

**AN ANALYSIS OF SEMANTIC DATA QUALITY DEFICIENCIES IN A NATIONAL  
DATA WAREHOUSE:  
A DATA MINING APPROACH**

by

**KIRSTIN BARTH**

submitted in accordance with the requirements for the degree of

**MASTER OF TECHNOLOGY**

in the subject of

**INFORMATION TECHNOLOGY**

at the

University of South Africa

Supervisor: Prof FO Bankole

Co-supervisor: Prof Christian W. Omlin

September 2018

## DECLARATION

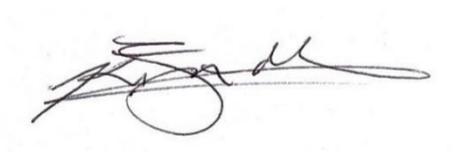
Name: Kirstin Tanja Barth

Student Number: 39929736

Degree: Master of Technology – Information Technology

An Analysis of Semantic Data Quality Deficiencies in a National Data Warehouse: A Data Mining Approach

I declare that the above dissertation is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.



Signature

28 March 2019

Date

**An Analysis of Semantic Data Quality Deficiencies in a National Data Warehouse:  
A Data Mining Approach**

by

**Kirstin Tanja Barth**

**SUMMARY**

This research determines whether data quality mining can be used to describe, monitor and evaluate the scope and impact of semantic data quality problems in the learner enrolment data on the National Learners' Records Database. Previous data quality mining work has focused on anomaly detection and has assumed that the data quality aspect being measured exists as a data value in the data set being mined. The method for this research is quantitative in that the data mining techniques and model that are best suited for semantic data quality deficiencies are identified and then applied to the data. The research determines that unsupervised data mining techniques that allow for weighted analysis of the data would be most suitable for the data mining of semantic data deficiencies. Further, the academic Knowledge Discovery in Databases model needs to be amended when applied to data mining semantic data quality deficiencies.

**KEYWORDS:**

Data warehouse; Data mining; Data quality mining; Exploratory data mining; Cluster analysis; Association rule; Knowledge Discovery in Databases; National Learners' Records Database; Learner enrolment data; Semantic data quality deficiencies

1	Chapter 1 .....	17
1.1	Introduction.....	17
1.2	Background .....	18
1.3	Motivation and problem statement .....	19
1.4	Research questions.....	21
1.5	Research aim and objectives .....	21
1.6	Justification and importance of the research.....	22
1.7	Outline of the research .....	22
2	Chapter 2: Literature review .....	25
2.1	Introduction.....	25
2.2	Definition of key concepts and ideas in the research.....	25
2.3	Review of literature.....	32
2.4	Chapter summary .....	37
3	Chapter 3: Methodology .....	39
3.1	Introduction.....	39
3.2	Research epistemology and ontology .....	39
3.3	Theoretical aspect of the research.....	40
3.4	Research methodology.....	43
3.5	Research method .....	45
3.6	Research context .....	48
3.6.1	Introduction.....	48
3.6.2	Defining the semantic business rules.....	49
3.6.3	Analysis of the data structures .....	50
3.6.4	Considerations that impact the analysis of the data.....	65
3.6.5	Selection of additional data for the analysis .....	70
3.6.6	Conclusion .....	71
3.7	Data collection .....	72
3.8	Data analysis procedure .....	72
3.8.1	Obtaining data from the NLRD .....	72
3.8.2	Raw data obtained from the NLRD .....	74
3.8.3	Overarching data derivation considerations .....	77
3.8.4	Learnership enrolment data selection, pre-processing and derivation.....	86

3.8.5	Qualification enrolment data selection, pre-processing and derivation.....	86
3.8.6	Unit Standard enrolment data selection, pre-processing and derivation .....	87
3.8.7	Data mining methods .....	87
3.9	Chapter summary .....	88
4	Chapter 4: Data analysis and research findings .....	89
4.1	Introduction.....	89
4.2	ETQE accreditation.....	90
4.2.1	Learnership enrolments.....	90
4.2.2	Qualification enrolments.....	94
4.2.3	Unit Standard enrolments .....	99
4.2.4	Conclusion .....	105
4.3	ETQE accreditation to quality assure the qualification or unit standard .....	106
4.3.1	Qualification enrolments.....	106
4.3.2	Unit Standard enrolments .....	114
4.3.3	Conclusion .....	124
4.4	Provider accreditation .....	125
4.4.1	Learnership enrolments.....	125
4.4.2	Qualification enrolments.....	130
4.4.3	Unit Standard enrolments .....	135
4.4.4	Conclusion .....	140
4.5	Provider accreditation to offer the qualification or unit standard.....	141
4.5.1	Qualification enrolments.....	141
4.5.2	Unit Standard enrolments .....	146
4.5.3	Conclusion .....	151
4.6	Assessor registration .....	152
4.6.1	Learnership enrolments.....	152
4.6.2	Qualification enrolments.....	159
4.6.3	Unit Standard enrolments .....	165
4.6.4	Conclusion .....	171
4.7	Assessor registration to assess the qualification or unit standard.....	171
4.7.1	Qualification enrolments.....	172
4.7.2	Unit Standard enrolments .....	178
4.7.3	Conclusion .....	184
4.8	Correlation between learnerships and their associated qualifications .....	185

4.8.1	No Qual Enrolment .....	188
4.8.2	Lshp Enrolled, Qual Achieved (Derived) .....	191
4.8.3	Lshp Enrolled, Qual Achieved.....	194
4.8.4	Lshp Completed, Qual Enrolled .....	195
4.8.5	Lshp Completed, Qual Enrolled (Derived) .....	196
4.8.6	Lshp Completed Before Qual (Derived).....	198
4.8.7	Lshp Completed Before Qual .....	199
4.8.8	Lshp Completed After Qual (Derived) .....	200
4.8.9	Lshp Completed After Qual.....	202
4.8.10	Summary of semantic infringements by ETQE.....	203
4.8.11	Conclusion .....	204
4.9	Qualification/Unit Standard registration.....	208
4.9.1	Qualification enrolments.....	209
4.9.2	Unit Standard enrolments .....	215
4.9.3	Conclusion .....	220
4.10	Unit Standard based qualification achievements .....	221
4.10.1	Insufficient Unit Standard Credits Achieved.....	225
4.10.2	No Unit Standard Credits Achieved .....	229
4.10.3	Incorrect Mix of Unit Standard Credits Achieved.....	232
4.10.4	Summary of semantic infringements by ETQE.....	236
4.10.5	Conclusion .....	237
4.11	Data quality affinity .....	238
4.11.1	Learnership enrolments.....	238
4.11.2	Qualification enrolments.....	240
4.11.3	Unit Standard enrolments .....	246
4.11.4	Conclusion .....	255
4.12	Chapter summary .....	255
5	Chapter 5 – Conclusions and recommendations .....	259
5.1	Response to research objectives and questions.....	259
5.2	Limitations .....	266
5.3	Recommendations for future research .....	266
6	Acronyms .....	267
7	Bibliography.....	268
	Appendix A .....	274

Appendix B.....	290
Appendix C.....	292
Appendix D .....	320
Appendix E.....	322
Appendix F.....	365
Appendix G .....	368
Appendix H .....	407
Appendix I.....	410
Appendix J.....	417
Appendix K .....	463
Appendix L.....	516
Appendix M.....	549
Appendix N .....	595
Appendix O .....	604
Appendix P.....	626

## Table of Figures

Figure 1.7.1 Diagrammatic representation of the outline of the research .....	23
Figure 3.2.1 Two dimensional matrix of the “Framework for Data and Information Quality Research” .....	41
Figure 3.3.1 Diagrammatic representation of application of criteria in the research .....	43
Figure 3.5.1 Diagrammatic representation of the KDD process .....	46
Figure 3.6.1 KDD phase - Selection.....	48
Figure 3.6.3.1 Conceptual diagram of the learner enrolment record.....	51
Figure 3.6.3.1.a.1 Conceptual diagram of the NQF concepts that participate in business rule 1.a .....	52
Figure 3.6.3.1.a.2 Conceptual diagram of the tables and fields that inform business rule 1.a.	52
Figure 3.6.3.1.b.1 Conceptual diagram of the NQF concepts that participate in business rule 1.b .....	53
Figure 3.6.3.1.b.2 Conceptual diagram of the tables and fields that inform business rule 1.b.	54
Figure 3.6.3.2.a.1 Conceptual diagram of the NQF concepts that participate in business rule 2.a .....	54
Figure 3.6.3.2.a.2 Conceptual diagram of the tables and fields that inform business rule 2.a.	55
Figure 3.6.3.2.b.1 Conceptual diagram of the NQF concepts that participate in business rule 2.b .....	55
Figure 3.6.3.2.b.2 Conceptual diagram of the tables and fields that inform business rule 2.b.	56
Figure 3.6.3.3.a.1 Conceptual diagram of the NQF concepts that participate in business rule 3.a .....	56
Figure 3.6.3.3.a.2 Conceptual diagram of the tables and fields that inform business rule 3.a.	57
Figure 3.6.3.3.b.1 Conceptual diagram of the NQF concepts that participate in business rule 3.b .....	57
Figure 3.6.3.3.b.2 Conceptual diagram of the tables and fields that inform business rule 3.b.	58
Figure 3.6.3.4.1 Conceptual diagram of the NQF concepts that participate in business rule 4	59
Figure 3.6.3.4.2 Conceptual diagram of the tables and fields that inform business rule 4.....	59
Figure 3.6.3.5.1 Conceptual diagram of the NQF concepts that participate in business rule 5	60
Figure 3.6.3.5.2 Conceptual diagram of the tables and fields that inform business rule 5.....	61
Figure 3.6.3.6.a.1 Conceptual diagram of the NQF concepts that participate in business rule 6.a .....	62
Figure 3.6.3.6.a.2 Conceptual diagram of the tables and fields that inform business rule 6.a.	63

Figure 3.6.3.6.b.1 Conceptual diagram of the NQF concepts that participate in business rule 6.b .....	64
Figure 3.6.3.6.b.2 Conceptual diagram of the tables and fields that inform business rule 6.b.65	
Figure 3.6.4.2.1 Figure depicting active accreditations/registrations with gaps .....	68
Figure 3.6.4.2.2 Figure depicting active accreditation/registration with gaps removed .....	68
Figure 3.8.1 KDD phases – Pre-processing and Transformation .....	72
Figure 4.1 KDD phases – Data Mining .....	89
Figure 4.2.1.1 % records according to the semantic business rule that requires that the ..... ETQE must be accredited for the duration of the learner’s active enrolment on the learnership .....	93
Figure 4.2.2.1 % records according to the semantic business rule that requires that the ..... ETQE must be accredited for the duration of the learner’s active enrolment on the qualification .....	97
Figure 4.2.3.1 % records according to the semantic business rule that requires that the ..... ETQE must be accredited for the duration of the learner’s active enrolment on the unit standard .....	102
Figure 4.3.1.1 % records according to the semantic business rule that requires that the ..... ETQE must be accredited to quality assure the qualification for the duration of the learner’s active enrolment on the qualification .....	109
Figure 4.3.2.1 % records according to the semantic business rule that requires that the ..... ETQE must be accredited to quality assure the unit standard for the duration of the learner’s active enrolment on the unit standard .....	117
Figure 4.4.1.1 % records according to the semantic business rule that requires that the provider must be accredited for the duration of the learner’s active enrolment on the learnership .....	128
Figure 4.4.1.2 % records that infringe the semantic business rule that requires that the provider must be accredited for the duration of the learner’s active enrolment on the learnership by category .....	129
Figure 4.4.2.1 % records according to the semantic business rule that requires that the provider must be accredited for the duration of the learner’s active enrolment on the qualification .....	133
Figure 4.4.2.2 % records that infringe the semantic business rule that requires that the provider must be accredited for the duration of the learner’s active enrolment on the qualification by category .....	134

Figure 4.4.3.1 % records according to the semantic business rule that requires that the provider must be accredited for the duration of the learner’s active enrolment on the unit standard.....	138
Figure 4.4.3.2 % records that infringe the semantic business rule that requires that the provider must be accredited for the duration of the learner’s active enrolment on the unit standard by category.....	139
Figure 4.5.1.1 % records according to the semantic business rule that requires that the provider must be accredited to offer the qualification for the duration of the learner’s active enrolment on the qualification.....	145
Figure 4.5.1.2 % records that infringe the semantic business rule that requires that the provider must be accredited for the duration of the learner’s active enrolment on the qualification by category .....	145
Figure 4.5.2.1 % records according to the semantic business rule that requires that the provider must be accredited to offer the unit standard for the duration of the learner’s active enrolment on the unit standard .....	149
Figure 4.5.2.2 % records that infringe the semantic business rule that requires that the provider must be accredited for the duration of the learner’s active enrolment on the unit standard by category.....	150
Figure 4.6.1.1 % records according to the semantic business rule that requires that .....	155
the assessor must be registered at the time of the completion of the learnership.....	155
Figure 4.6.2.1 % records according to the semantic business rule that requires that .....	161
the assessor must be registered at the time of the achievement of the qualification .....	161
Figure 4.6.3.1 % records according to the semantic business rule that requires that .....	167
Figure 4.7.1.1 % records according to the semantic business rule that requires that .....	174
the assessor must be registered to assess the qualification at the time of the achievement of the qualification.....	174
Figure 4.7.2.1 % records according to the semantic business rule that requires that .....	180
the assessor must be registered to assess the unit standard at the time of the achievement of the unit standard .....	180
Figure 4.8.1 % records according to the semantic business rule that requires that the intrinsic relationship between the completion of a learnership and achievement of its related qualification has been upheld .....	187

Figure 4.8.2 % records that infringe the semantic business rule that requires that the intrinsic relationship between the completion of a learnership and achievement of its related qualification has been upheld .....	188
Figure 4.8.1.1 % records by ETQE where an associated qualification enrolment record does not exist for the learnership enrolment record.....	189
Figure 4.8.2.1 % records by ETQE where the learnership enrolment record has a completion status of enrolled whilst it's associated qualification enrolment record has an enrolment status of achieved.....	192
Figure 4.8.3.1 % records by ETQE where the learnership enrolment record has a completion status of enrolled whilst its associated qualification enrolment record has an enrolment status of achieved.....	195
Figure 4.8.4.1 % records by ETQE where the learnership enrolment record has a completion status of completed whilst its associated qualification enrolment record has an enrolment status of not achieved .....	196
Figure 4.8.5.1 % records by ETQE where the learnership enrolment record has a completion status of completed whilst its derived associated qualification enrolment record has an enrolment status of not achieved .....	197
Figure 4.8.6.1 % records by ETQE where the learnership enrolment record was completed more than a year prior to the achievement of the associated qualification enrolment record	199
Figure 4.8.7.1 % records by ETQE where the learnership enrolment record was completed more than a year prior to the achievement of the derived associated qualification enrolment record.....	200
Figure 4.8.8.1 % records by ETQE where the learnership enrolment record was completed more than a year after the achievement of the derived associated qualification enrolment record.....	201
Figure 4.8.9.1 % records by ETQE where the learnership enrolment record was completed more than a year after the achievement of the associated qualification enrolment record.....	202
Figure 4.9.2.1 % records according to the semantic business rule that requires that the qualification must be registered for the duration of the learner's active enrolment on the qualification.....	211
Figure 4.9.2.1 % records according to the semantic business rule that requires that the unit standard must be registered for the duration of the learner's active enrolment on the unit standard.....	217

Figure 4.10.1 % records according to the semantic business rule that requires that in the case where the learner has achieved the qualification, and the qualification is a unit standards based qualification, the learner has achieved the correct number and mix of credits for the qualification.....	224
Figure 4.10.2 % records that infringe the semantic business rule that requires that in the case where the learner has achieved the qualification, and the qualification is a unit standards based qualification, the learner has achieved the correct number and mix of credits for the qualification.....	225
Figure 4.10.1.1 % records by ETQE where the learner has not achieved the correct number of credits for the qualification.....	226
Figure 4.10.2.1 % records by ETQE where the learner has achieved the qualification, the qualification is a unit standards based qualification and the learner has not achieved any credits for the qualification.....	230
Figure 4.10.3.1 % records by ETQE where even though the learner has achieved the correct number of credits for the qualification, the number of credits derived from core, fundamental or elective unit standards is incorrect .....	233
Figure 5.1 KDD phases – Interpretation.....	259
Figure A.1 Conceptual diagram of seven concepts of the National Qualifications Framework .....	274
Figure A.2 Conceptual diagram of qualifications and unit standards .....	274
Figure A.3 Conceptual diagram of learnerships and their relationship to qualifications .....	277
Figure A.4.1 Conceptual diagram of ETQAs and their relationship to qualifications and unit standards .....	278
Figure A.4.2 Conceptual diagram of ETQAs and their relationship to providers and assessors .....	279
Figure A.5 Conceptual diagram of learner enrolments and their relationship to ETQAs, providers, assessors, learnerships, qualifications and unit standards .....	280
Figure A.6 Conceptual diagram of learner enrolments and their relationship to ETQEs, providers, assessors, learnerships, qualifications and unit standards .....	285
Figure C.3.4.1 Illustrative diagram of ETQE_IND development .....	303
Figure C.3.5.1 Illustrative diagram of PROV_IND development.....	308
Figure C.3.6.1 Illustrative diagram of ASOR_IND development.....	311
Figure C.3.7.1 Illustrative diagram of QUAL_IND development .....	314
Figure E.3.4.1 Illustrative diagram of ETQE_IND development .....	333

Figure E.3.5.1 Illustrative diagram of ETQE_ACCRED_IND development .....	338
Figure E.3.6.1 Illustrative diagram of PROV_IND development .....	343
Figure E.3.7.1 Illustrative diagram of PROV_ACCRED_IND development.....	346
Figure E.3.8.1 Illustrative diagram of ASOR_IND development .....	350
Figure E.3.9.1 Illustrative diagram of ASOR_REGSTR_IND development.....	352
Figure E.3.10.1 Illustrative diagram of QUAL_REGSTR_IND development .....	355
Figure G.3.4.1 Illustrative diagram of ETQE_IND development .....	379
Figure G.3.5.1 Illustrative diagram of ETQE_ACCRED_IND development.....	384
Figure G.3.6.1 Illustrative diagram of PROV_IND development.....	389
Figure G.3.7.1 Illustrative diagram of PROV_ACCRED_IND development .....	392
Figure G.3.8.1 Illustrative diagram of ASOR_IND development.....	396
Figure G.3.9.1 Illustrative diagram of ASOR_REGSTR_IND development .....	399
Figure G.3.10.1 Illustrative diagram of USTD_REGSTR_IND development.....	401
Figure I.2.1 Screenshot of exploratory data mining results generated the data mining tool ..	411
Figure I.2.2 Screenshot of automatically generated graph showing the distribution of value for a data field .....	411

## Table of Tables

Table 2.3.1 Summary of the type of implementation of association rule mining research .....	33
Table 2.3.2 Summary of the further development of association rule mining research.....	34
Table 3.4.1 Comparing characteristics of qualitative and quantitative research .....	44
Table 4.2.1.1 ETQE accreditation categories for learnership enrolments .....	91
Table 4.2.2.1 ETQE accreditation categories for qualification enrolments .....	95
Table 4.2.3.1 ETQE accreditation categories for unit standard enrolments.....	100
Table 4.3.1.1 ETQE accreditation to quality assure the qualification categories.....	107
Table 4.3.2.1 ETQE accreditation to quality assure the unit standard categories .....	114
Table 4.4.1.1 Provider accreditation categories for learnership enrolments .....	125
Table 4.4.2.1 Provider accreditation categories for qualification enrolments.....	131
Table 4.4.3.1 Provider accreditation categories for unit standard enrolments .....	136
Table 4.5.1.1 Provider accreditation to offer the qualification categories .....	142
Table 4.5.2.1 Provider accreditation to offer the unit standard categories.....	147
Table 4.6.1.1 Assessor registration categories for learnership enrolments .....	154
Table 4.6.1.2 ‘No Registration’ records by submitting ETQE identifier and learnership identifier, count of assessors, % learnership enrolment records in the category and records in this category as a % of the records submitted by the ETQE .....	158
Table 4.6.2.1 Assessor registration categories for qualification enrolments.....	159
Table 4.6.2.2 ‘No Registration’ records by submitting ETQE identifier, count of qualification identifier, count of assessors, % qualification enrolment records in the category and records in this category as a % of the records submitted by the ETQE .....	163
Table 4.6.3.1 Assessor registration categories for unit standard enrolments .....	165
Table 4.6.3.2 ‘No Registration’ records by submitting ETQE identifier, count of unit standard identifier, count of assessors, % unit standard enrolment records in the category and records in this category as a % of the records submitted by the ETQE .....	169
Table 4.7.1.1 Assessor registration to assess the qualification categories .....	172
Table 4.7.1.2 ‘No Registration’ records by submitting ETQE identifier, count of qualification identifier, count of assessors, % qualification enrolment records in the category and records in this category as a % of the records submitted by the ETQE .....	176
Table 4.7.2.1 Assessor registration to assess the unit standard categories.....	178

Table 4.7.2.2 ‘No Registration’ records by submitting ETQE identifier, count of unit standard identifier, count of assessors, % unit standard enrolment records in the category and records in this category as a % of the records submitted by the ETQE .....	182
Table 4.8.1 A corresponding qualification achievement record has been submitted .....	185
for completed learnership categories .....	185
Table 4.8.10.1 % of records submitted by an ETQE that do not have an associated qualification enrolment record, % of records submitted by an ETQE that have a category that describes a semantic business rule issue, and the sum percentage of both .....	204
Table 4.9.2.1 Qualification was registered for the duration of the learner’s active enrolment on the qualification categories .....	209
Table 4.9.2.1 Unit standard was registered for the duration of the learner’s active enrolment on the unit standard categories .....	215
Table 4.10.1 Learner has achieved the correct number and mix of credits for the qualification categories .....	222
Table 4.10.4.1 % of records submitted by an ETQE that in the case where the learner has achieved the qualification, and the qualification is a unit standards based qualification, the learner has achieved the correct number and mix of credits for the qualification.....	237
Table I.2.3 Example of statistics generated for records that have ETQE_IND_DESC ‘Start Before, End During’ .....	412
Table J.1.2.1 ‘No Accreditation’ records by submitting ETQE identifier, count of Not Primary ETQE providers, count of Primary ETQE of provider and % learnership enrolment records in the category.....	420
Table J.1.9.1 % of records submitted by an ETQE that have a category that describes a semantic business rule issue .....	432
Table J.2.3.1 ‘No Accreditation’ records by submitting ETQE identifier, count of Not Primary ETQE providers, count of Primary ETQE of provider and % qualification enrolment records in the category .....	436
Table J.2.9.1 % of records submitted by an ETQE that have a category that describes a semantic business rule issue .....	447
Table J.3.4.1 ‘No Accreditation’ records by submitting ETQE identifier, count of Not Primary ETQE providers, count of Primary ETQE of provider and % unit standard enrolment records in the category .....	452
Table J.3.9.1 % of records submitted by an ETQE that have a category that describes a semantic business rule issue .....	462

Table L.1.1.1 ‘No Accreditation’ records by submitting ETQE identifier, count of Not Primary ETQE providers, count of Primary ETQE of provider and % qualification enrolment records in the category.....	518
Table L.1.9.1 % of records submitted by an ETQE that have a category that describes a semantic business rule issue .....	531
Table L.2.1.1 ‘No Accreditation’ records by submitting ETQE identifier, count of Not Primary ETQE providers, count of Primary ETQE of provider and % unit standard enrolment records in the category.....	534
Table L.2.9.1 % of records submitted by an ETQE that have a category that describes a semantic business rule issue .....	547

# **1 Chapter 1**

## **1.1 Introduction**

The South African Qualifications Authority (SAQA) implemented the National Learners' Records Database (NLRD) to support the implementation, monitoring and evaluation, of the National Qualifications Framework (NQF) in the Republic of South Africa. The NLRD comprises in part a data warehouse which is the most comprehensive centralized data source of learning enrolment across all of the education sectors in South Africa.

The NLRD data warehouse obtains data from a variety of data sources and as a result is prone to data quality issues. SAQA has implemented technical quality control measures on data received by the NLRD to ensure that the form of the data submitted to the NLRD is correct. However the semantic quality of the data, specifically in regard to the explicit and implicit rules of the NQF, cannot not be ensured. The NLRD data warehouse currently has no mechanisms with which to describe, monitor and evaluate the scope and impact of data that contains semantic data quality deficiencies

Data mining is the process of examining large established data sets in order to produce new information. The implementation of data mining techniques on data warehouses in order to discover meaningful trends and patterns in large amounts of data stored in data warehouses is an intuitive application of data mining technology. Data quality mining is a discrete data mining approach that entails the implementation of data mining techniques for the purpose of measuring data quality and the improvement of data.

This research determines whether data quality mining can be used to describe, monitor and evaluate the scope and impact of semantic data quality problems in the learner enrolment data on the NLRD. Further, the research determines which data mining techniques and model are best suited for the identification, measurement and description of semantic data quality deficiencies and how these techniques and model need to be amended in order to provide a mechanism with which to determine the semantic quality aspects of learner enrolment records in the NLRD.

## 1.2 Background

SAQA is a juristic person, regulated in terms of the National Qualifications Framework Act No. 67 of 2008 (Ministry in the Office of the President, National Qualifications Framework, Act 67 of 2008, p. 6). SAQA's objectives are to advance the objectives of the National Qualifications Framework (NQF), oversee the further development and implementation of the NQF and co-ordinate the sub-frameworks (Ministry in the Office of the President, National Qualifications Framework, Act 67 of 2008, p. 6). The NQF is "... a comprehensive system..." approved by the Minister of Higher Education and Training for "...the classification, registration, publication and articulation of quality assured qualifications" (Ministry in the Office of the President, National Qualifications Framework, Act 67 of 2008, p. 3).

In 1999, SAQA implemented the NLRD to support the implementation, monitoring and evaluation, of the NQF. The NLRD provides SAQA with functionality that provides it with:

1. an operational information system that allows it to combine education and training into a single framework, the NQF, and
2. a data warehouse that allows for the monitoring and evaluation of the implementation of the NQF.

The operational information system and data warehousing components of the NLRD are integrated into a single information system.

The data warehouse aspect of the NLRD specifically collects data in regard to legacy and current learner enrolment (although commonly referred to as learner achievements, the NLRD in fact stores data related to learner enrolments regardless of their achievement status) records from both the public and private education sectors. The learner enrolment aspect of the NLRD data warehouse is not limited to only learner enrolments; rather it includes all data aspects related to the learner enrolment record including provider, provider accreditation, assessor, assessor registrations and professional designation data.

The NLRD data warehouse is the most comprehensive centralized data source of learning enrolment across all of the education sectors in South Africa. The NLRD data warehouse provides information that is of fundamental importance to the continuous monitoring, planning and policy development for the education sector in South Africa. The value of the information extracted from the NLRD data warehouse is however directly dependent on the quality of the data stored therein.

In addition, the NLRD data warehouse's Extract-Transform-Load process has implemented technical quality control measures that ensure that data submitted to and loaded into this data warehouse conforms to a predefined format and minimum data quality standard. In other words, the technical quality control measures of the NLRD ensure that the form of the data submitted to the NLRD is correct. Due to the operational scope of these technical quality control measures, the semantic quality of the data, specifically in regard to the explicit and implicit rules of the NQF, cannot be ensured. Semantic quality of the data in this instance denotes the interpretation of the data within the context of the business rules of the framework of the NQF. The NLRD data warehouse currently has no mechanisms with which to describe, monitor and evaluate the scope and impact of data that contains semantic data quality deficiencies.

A preliminary literature review suggested that data quality mining could be utilized to describe, monitor and evaluate the scope and impact of data quality deficiencies in a dataset. Further research however needed to be conducted to determine whether a modified data quality mining process could provide a suitable method and model for the NLRD data warehouse in regard to semantic data quality deficiencies.

### **1.3 Motivation and problem statement**

Since its establishment in 1999, more than 85 million learner enrolment records have been loaded into the NLRD data warehouse. Data that is imported into the NLRD data warehouse conforms to strict technical requirements that ensure the technical quality of the data that the data warehouse stores. The scope of the process and procedures that ensure the quality of the data in the NLRD data warehouse however cannot, and in some instances should not, ensure that all data meets expected instance level quality constraints described both explicitly and implicitly in the NQF. The NLRD data warehouse currently has no mechanisms with which to describe, monitor and evaluate the scope and impact of data records that do not conform to these types of semantic data quality rules. A sub-discipline of data mining, namely data quality mining, may be useful in providing such mechanisms for the NLRD data warehouse.

A preliminary literature review has however shown that research related to data quality mining has largely focused on the cleaning of data, with only some research having been conducted in regard to how to measure and explain data quality deficiencies.

This includes research that focused on the cleaning of data using:

- algorithms (Das & Saha, 2009, p. 109), (Farzi & Dastjerdi, 2010, p. 115),
- classifiers (Grüning, 2007, p. 2)
- clustering and fuzzy techniques (Khosravani, 2012, p. 8) and
- supervised learning and modelling (Sheng, Provost, & Ipeirotis, 2008, p. 614).

Whereas approaches with which to measure and explain data quality deficiencies included:

- a statistical approach to data quality assessment (Dasu & Johnson, 2003, p. 164)
- the use of multiple target rules to identify inconsistent values (Natarajan, Li, & Koronios, Data Mining Techniques for Data Cleaning, 2009, p. 296)
- the employment of association rules to measure data quality (Vizhi & Bhuvaneshwari, 2012, p. 33)

Research that has focused on the measurement and explanation of data quality deficiencies assumes that data quality deficiencies are only a measure of whether or not the data is a true representation of the object that it represents in the real world.

The nature and complexity of the instance level data quality issues that exist in the NLRD data warehouse contain an additional data quality dimension in that data quality deficiencies are also a measure of whether or not the data conforms to both the explicit and implicit semantic business rules of the NQF.

The NLRD data warehouse is in need of a suitable method and model with which to describe, monitor and evaluate the scope and impact of data records that do not conform to the semantic data quality rules. The NLRD data warehouse currently is the most comprehensive centralized data source that describes learning enrolment across all of the education sectors in South Africa. The NLRD data warehouse has the potential to provide information that is of fundamental importance to continuous monitoring, planning and policy development for the education sector in South Africa. However, the value of the information extracted from the NLRD data warehouse is directly dependent on the quality of the data stored in the data warehouse. Data quality mining can potentially provide such a method and model, but the way in which it can be implemented for this purpose has not yet been explored.

## **1.4 Research questions**

The main research question for the research is:

- How can data quality mining be used to describe, monitor and evaluate the scope and impact of semantic data quality problems in the learner enrolment data on the National Learners' Records Database?

In order to properly answer the main research question the following sub-questions need to be answered:

- Which data mining techniques are best suited to identify, measure and describe semantic data quality deficiencies?
- How will the existing data mining methods and models need to be amended in order to provide a “pure” mechanism with which to determine the semantic quality aspects of a database?

## **1.5 Research aim and objectives**

This research aims to determine which data mining techniques are best suited for the identification, measurement and description of semantic data quality deficiencies and which aspects of existing data quality mining methods and models can be utilized in order to determine semantic data quality deficiencies. Thereafter a standard set of data mining techniques and an adapted data mining method and model will be developed for the NLRD data warehouse with the aim to assess data quality deficiencies in learner enrolment records in the NLRD data warehouse in an ongoing manner.

The objectives of this research are:

1. To determine which data mining techniques are best suited for the identification, measurement and description of semantic data quality deficiencies.
2. To determine which aspects of existing data quality mining methods and models can be utilized in order to define semantic data quality deficiencies.
3. To develop a standard set of data mining techniques that can identify, measure and describe semantic data quality deficiencies related to learner enrolment records in the NLRD data warehouse.
4. To develop an adapted data quality mining method and model that can be utilized by the NLRD data warehouse to continuously assess data quality deficiencies in learner enrolment records in the NLRD data warehouse.

## **1.6 Justification and importance of the research**

This research aims to determine the manner in which data quality mining can be utilized to describe, monitor and evaluate the scope and impact of semantic data quality problems in a data set. The research will further determine the data mining techniques that are best suited to identify, measure and describe semantic data quality deficiencies. This research will also determine a method and model which can be specifically applied to the mining of semantic data quality problems in data.

On the 16<sup>th</sup> of April 2012 SAQA signed the Groningen Declaration, thereby committing South Africa to the goal of the “...*free movement of students and skilled workers on a global scale*”. One of the objectives of this declaration is “...*promoting acceptance, for purposes of recognition, of digital student data in lieu of paper documents*”. As the definitive source of such digital data from South Africa, the legitimacy and accuracy of the data contained in the NLRD is of vital importance.

Further, all of the SADC Member States are in the process of implementing National Qualification Frameworks (Krönner, 2005, p. 10). Inherently this means that potentially twelve (12) SADC State Members are developing data warehouses that will allow the SADC State Member to evaluate the implementation of their NQF. An adapted data mining method and model that will allow such SADC State Members to monitor and evaluate the quality of the data in their own data warehouses could prove valuable.

Finally, in addition to contributing to the continuous monitoring, planning and policy development for the education sector in South Africa, the NLRD data warehouse has much to offer policymakers and practitioners whose long standing goal is to make data-based decisions utilizing high-quality data. An adapted method and model with which to describe, monitor and evaluate the scope and impact of the semantic quality of data records in the NLRD will ultimately provide SAQA with the mechanism with which to assure the data that is provided to these stakeholders.

## **1.7 Outline of the research**

This thesis is divided into five chapters as illustrated in Figure 1.7.1. The first chapter provides background to the topic of the research and a description of the research problem and

question. Finally, the chapter justifies the need for the research and defines the information that it will provide.

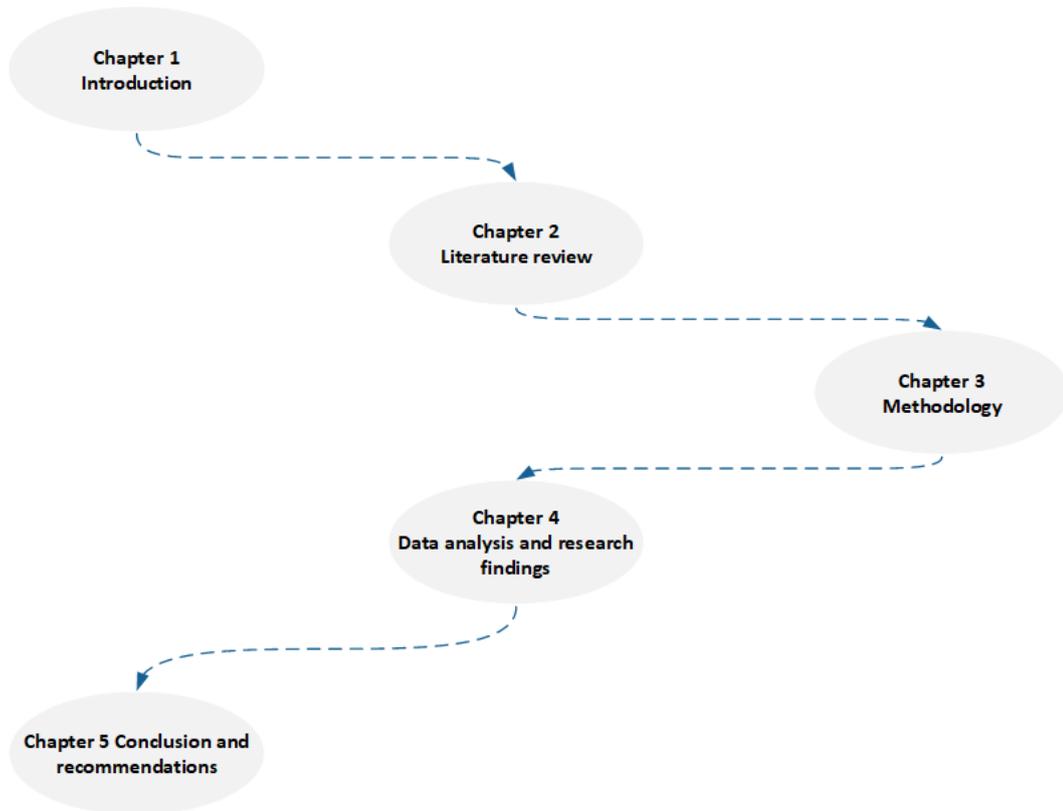


Figure 1.7.1 Diagrammatic representation of the outline of the research

Chapter 2 reviews literature aligned to the theoretical argument being made in the research topic. The chapter further identifies gaps in the literature and how they impact the research being conducted. In closing, this chapter identifies applicable data mining techniques as well as the related research process.

Chapter 3 highlights that the research has two main focus areas and discusses the research using the four basic elements of the research process (Crotty, 1998, p. 3). This is followed by a section that provides background to the topic, thereby giving an understanding of the data structures of the NLRD and the applicable semantic business rules that form the core of this research. Further, the chapter notes that there was no data collection because the data mined is pre-existing. The chapter goes on to detail how the data was obtained and which data was obtained. This is followed by a discussion as to how the physical data received from the NLRD is processed and prepared for data mining. Finally, the chapter describes the data mining methods, as identified in Chapter 2, that are applied to the data.

Chapter 4 presents the results of the data mining conducted on the data received from the NLRD. The results are presented by semantic rule as applied to learnership, qualification and/or unit standard enrolment records. Each analysis comprises of the results of the exploratory data mining.

Where the number of records that do not comply with a semantic business rule exceeds 5%, and the results lend themselves to further data mining efforts, the data subset is further analysed utilizing cluster data mining techniques. In these instances both a description of the most pertinent aspects of the resultant clusters and the technical description of the cluster are provided. Further, a summary of semantic infringements by ETQE is also given in order to provide clarity in regard to the results when compared to an overall view of the percentage of infringements by ETQE. Each analysis is finalized with a conclusion that highlights the specific recommendations for SAQA in regard to the findings that were made.

The data is also analysed using association data mining techniques where associations are sought amongst records that have one or more semantic business rule infringement. These analyses are conducted by learnership, qualification and/or unit standard enrolment records. As with the exploratory and cluster data mining technique analyses, this analysis is also finalized with a conclusion that highlights the findings of the analysis and makes specific recommendations for SAQA in regard to the findings that were made.

The chapter draws a final conclusion that highlights findings across all the semantic business rules and learnership, qualification and unit standard enrolment records.

Chapter 5 reviews the results of the research in light of the research question, highlighting whether there are any limitations in regard to achieving the objectives of the research as described in Chapter 1 and provides recommendations for future research.

## 2 Chapter 2: Literature review

### 2.1 Introduction

Data mining is an interdisciplinary subfield of computer science with diverse applications across different domains resulting in a variety of approaches and perspectives (Han & Kamber, 2001, p. 13). The literature review focused on the identification of data mining techniques that support the description, monitoring and evaluation of the scope and impact of data records in the NLRD that do not meet the explicit and implicit rules of the NQF as described in Section 3.6.2. The main objective of the review is to determine which data mining techniques have been used in other studies that relate to the identification, measurement and description of data quality deficiencies in a data set. Further, the literature review seeks to determine which of the identified data mining techniques specifically lend themselves to the identification, measurement and description of semantic data quality deficiencies.

Section 2.2 provides a brief introduction to technical concepts such as data warehouses, data mining and data quality mining and how and why they are linked to this research. Further, concepts such as descriptive and predictive data mining concepts and their supporting techniques are also introduced in Section 2.2 for the purposes of highlighting which apply to this research. Section 2.3 follows with a review of current research related to data quality mining in which the specific data mining techniques utilized are considered and their applicability to this research is determined.

Consequently, additional literature is also reviewed in order to find clarity on issues such as:

- how to approach the data mining of data that has never been mined before,
- how the complexities of the data as described in Sections 3.6.3 may influence the overall research process, and
- how possible strategies that could be implemented to address the additional considerations that will impact the analysis of the data as described in Section 3.6.4.

### 2.2 Definition of key concepts and ideas in the research

A data warehouse is defined as “*a subject-oriented, integrated, non-volatile, and time-variant collection of data in support of management’s decisions*” (Inmon, 2005, p. 29). In this context;

- **subject-oriented** means that the data warehouse is organized around high-level entities of the enterprise (Hoffer, Prescott, & McFadden, 2008, p. 422),
- **integrated**, one of the most important aspects of a data warehouse, defines that the data within a data warehouse which is obtained from disparate sources has a single physical corporate image (Inmon, 2005, p. 30),
- **non-volatile** means that the data that is stored in a data warehouse is not updated (Inmon, 2005, p. 32) by the end user of the data warehouse (Hoffer, Prescott, & McFadden, 2008, p. 422), and
- **time-variant** indicates that each unit of data inside the data warehouse is accurate as of some moment in time (Inmon, 2005, p. 32).

The process known as ETL is core to the implementation of any data warehouse. ETL is the end-to-end process of taking data from disparate data sources and loading these into the data warehouse (Agrawal, Chafle, Goyal, Mittal, & Mukherjea, 2008, p. 1278). The ETL process periodically *extracts* data from the disparate data sources, *transforms* the data into a common format and *loads* the data into the data warehouse (Eckerson & White, 2003, p. 5).

The fact that data warehouses obtain data from a variety of data sources make them prone to data quality issues (Rahm & Hai Do, 2000, p. 1). The probability of receiving low quality data in a data warehouse is inflated when integrating large databases into a single data warehouse (Januzaj & Januzaj, 2009, p. 17). Data quality issues in data warehouses fall broadly into the classification of multi-source data quality problems and can range from schema-level (naming conflicts, structural conflicts) to instance-level quality problems (overlapping, contradicting and inconsistent data) (Rahm & Hai Do, 2000, p. 4).

Nisbet, Elder & Miner (Nisbet, Elder, & G., 2009, p. 17) define data mining as “... *the use of machine learning algorithms to find faint patterns of relationship between data elements in large, noisy, and messy data sets, which can lead to actions to increase benefit in some form (diagnosis, profit, detection, etc.)*”. The efficiency of data mining is however critically subject to the quality of the data being mined (Apiletti, Bruno, Ficarra, & Baralis, 2006, p. 2).

The implementation of data mining techniques in order to discover meaningful trends and patterns in large amounts of data stored in data warehouses is an intuitive application of data

mining technology. There is however, based on the statements above, a tension between data warehousing and data mining, where one technology is prone to data quality issues and the other requires quality data in order to generate meaningful outputs. The evolution of this tension to the development of a new data mining approach that focuses on data quality is seemingly inevitable.

The new data mining approach, namely data quality mining, was first articulated in 2001 and is defined as the “... *deliberate application of data mining techniques for the purpose of data quality measurement and improvement.*” (Hipp, Guntzer, & Grimmer, 2001, p. 2). The goals of data quality mining are defined as “... *to detect, quantify, explain, and correct data quality deficiencies in very large databases*” (Hipp, Guntzer, & Grimmer, 2001, p. 2).

The four main focus areas for data quality mining are defined as (Hipp, Guntzer, & Grimmer, 2001, p. 3):

- Employment of data mining methods to measure and explain data quality deficiencies
- Employment of data mining methods to correct deficient data
- Extension of Knowledge Discovery in Databases (KDD) process models to reflect the potentials of data quality mining
- Development of specialized process models for “pure data quality mining”

The main focus of this research is related to the “*employment of data mining methods to measure and explain data quality deficiencies*”, in other words this research aims to measure and describe semantic data quality deficiencies in the NLRD.

With the purpose of achieving this focus, the research will need to determine which data mining techniques are best suited for the identification, measurement and description of data quality deficiencies in the NLRD. Further, the research will need to determine which aspects of existing data quality mining methods and models can be utilized and adapted in order to determine such data quality deficiencies in the NLRD.

Data mining concepts can be broadly split into one of two categories, which relate to the overall goal of the data mining, namely (Fayyad, Piatetsky-Shapiro, & Smyth, From Data Mining to Knowledge Discovery in Databases, 1996, p. 85):

- predictive data mining where the primary goal is to forecast values determined from known results, and
- descriptive data mining where the primary goal is to find patterns and to present them in a user understandable format.

The primary goal of this research is to find patterns in data related to the semantic business rules described in Section 3.6.2 and present them in a user understandable format. As a result mostly descriptive data mining concepts were considered during the review of current research.

Predictive and descriptive data mining concepts are respectively supported by supervised and unsupervised machine learning techniques. Supervised machine learning techniques require that data classifiers are trained on a collection of representative data and the algorithm is provided with a specific target variable whereas unsupervised machine learning techniques do not require a target variable or prior learning in order to mine the data (Chaovalit & Zhou, 2005, p. 2). However there are instances in which supervised machine learning is used for descriptive data mining and unsupervised machine learning is used for predictive data mining. As a result, this review mostly focused on unsupervised machine learning techniques, although supervised machine learning techniques as applied to DQM were also considered.

Most of the current research into data mining makes the assumption that the nature of the errors in the data is discernible in the data set used for the data mining activity. In other words the assumption is that the frequency of a data error is small enough to be detected as an anomaly within the overall data set. The assumption that there are sufficient data records in the NLRD that comply with the semantic business rules as described in Section 3.6.2, in order to determine which data records do not comply with the semantic business rules using data mining techniques, cannot be assured and will need to be tested in the research project.

The above concern touches on a broader issue, namely that the data in the NLRD has never been interrogated in line with the semantic business rules described in Section 3.6.2. As a result the resultant data set is unfamiliar to both SAQA and the researcher. To this end the utilization of Exploratory Data Mining (EDM) techniques are deemed appropriate for this research. EDM produces simple and fast analyses and summaries of the data in order to reveal the characteristics of the data (Dasu & Johnson, 2003). EDM tasks can be broadly categorised into one of three different types namely; summarizing the data, finding hidden relationships

and making predictions (Myatt, 2006, p. 2). For the purposes of this research the data mining that will be conducted is descriptive in nature and not predictive and as a result only the first two tasks are applicable. Based on these tasks the following types of EDM methods could be utilized for this research (Myatt, 2006, p. 4):

- summary tables that present raw information summarized in a number of ways,
- graphs that present information graphically in order to allow for the visual identification of trends and relationships,
- descriptive statistics that summarize information about particular data columns such as average or extreme values,
- inferential statistics that allow claims to be made in regard to the data with confidence,
- correlation statistics which quantify relationships within the data, and
- searching and grouping for organizing data into smaller groups and to quantify any conclusions with more information.

The research reviewed highlights that the data mining techniques best suited to this research are association rule mining and clustering, both of which are unsupervised data mining techniques. Association rule mining searches for similarities or events that occur together within data records and tries to infer rules that express those relationships (Agrawal, Imieliński, & Swami, Mining association rules between sets of items in large databases, 1993, p. 208). Clustering involves separating groups of data into collections that include consistent patterns.

The predominant research method for this research is the KDD process which is defined as (Fayyad, Piatetsky-Shapiro, & Smyth, The KDD Process for Extracting Useful Knowledge from Volumes of Data, 1996, p. 30), “...*the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.*” Data mining is a step within the overall process model of KDD. There are generally three different types of KDD process models, namely academic, industrial and hybrid (Crios, Pedryzc, Swiniarski, & Kurgan, 2007). The academic model developed by Fayyad et al. (Fayyad, Piatetsky-Shapiro, & Smyth, The KDD Process for Extracting Useful Knowledge from Volumes of Data, 1996, p. 30) will be used for this research. The process contains 9 discrete steps that fall into 5 different phases. The 5 different phases are:

1. Selection

The selection phase focuses on selecting a sub-set of the data that is concise enough to be processed within a reasonable time period whilst also large enough to contain a representation of the specific data quality dimension being studied.

## 2. Pre-processing

The pre-processing phase addresses the cleaning of data in regard to missing data values and the removal of statistical noise (i.e. unexplained variations in the data).

## 3. Transformation

The transformation phase focuses on determining which data fields need to be utilized in order to provide meaningful patterns in regard to the data quality dimension being studied. The transformation phase will result in a constrained dataset with limited features as required for the data mining process.

## 4. Data-mining

The data mining phase will entail the selection of a relevant data mining task and one or more data mining techniques identified in the literature review with which to mine the data.

## 5. Interpretation and Evaluation

The post data mining step will entail the analysis of the results of the data mining in which interesting patterns that represent knowledge are identified. The pre-final step of the overall process will be the validation of the results found by applying the same data mining techniques to the complete dataset to ensure that the same patterns occur in either a new subset of data or the complete dataset. The final step in the overall process will entail documenting the new understanding of the data and possibly the development of recommendations that can be used as a basis for change.

The pre-processing and transformation phases of the KDD process seem to lend themselves to the new process of deriving data elements that describe compliance to the semantic business rules into the data set. Unfortunately by definition neither the pre-processing nor transformation phases can completely accommodate the new process. Further, the placement of the tasks of the new process in the pre-processing and transformation phases which occur after the selection phase may prove to be problematic. For example in practice the derived data elements will need to be specifically created in line with the specific data mining technique, for each data mining technique. This approach could lead to inconsistent handling in the manner in which the data elements are derived.

As a result the research may need to define the new process of deriving data elements that describe compliance to the semantic business rules into the data set as a new KDD process phase. Alternatively the research may need to define discrete steps in the pre-processing and transformation phases of the KDD process that sufficiently and consistently accommodate the requirements of the new process.

Having broadly determined the types of data mining techniques that could be used in this research and overall process of the data mining activities, the nature of the data to be mined as described in Section 3.6.3 and Section 3.6.4 must be considered.

The analysis presented in Section 3.6.3 shows that the data to be mined is stored in a relational database and as a consequence the data mining of relational data needs to be considered in the review of current research. Traditional data mining techniques cannot be directly applied to relational databases (Houshmand & Alishahi, 2011, p. 332). MRDM (also referred to as Relational Data Mining) is considered a lively and interesting research area (Dzeroski, 2003, p. 14), with recent research showing that the implementation of relational database mining techniques are more efficient and effective in regard to execution time and memory utilization (Padhy & Panigrahi, 2012, p. 31). Unfortunately only one commercially available data mining software application, Safarii, applies MRDM in its general form. Other data mining software applications contain algorithms that require the flattening of the data prior to processing.

The data consideration described in Section 3.6.4.2, which highlights the requirement to derive an active accreditation or registration shows that the complexity of the relations in the data structures that will inform this research far exceed the normal requirements of mining relational data. The preparation of this data, in the required format, will need to be addressed as a pre-processing task prior to the mining of the data and as a result very little advantage can be envisaged in maintaining the relational design for the data mining task. For this reason the research will not be conducted utilizing MRDM data mining techniques.

The consideration described in Section 3.6.4.4 in regard to the representation of temporal data was also considered during the review of current research. Standard date formats are rarely directly supported by data mining techniques and in instances where they are supported this

type of value is treated as a single continuous variable which rules out the discovery of interesting patterns (Gupta, 2011, p. 259).

The derived start and end dates for learner enrolment records can be expected to display some trend/seasonal/cyclic components. Further, the analysis of the data mining results when presented in a time series that is represented by trend/seasonal/cyclic components will increase the understanding of the data. The data that will be mined represents more than 15 years of learner enrolment data and in order to facilitate the analysis process these dates will need to be broken down into new data fields that represent categories of time. The actual definition of the categories of time cannot be predicted prior to the analysis of the data. The implementation of EDM tasks such as those described in Section 2.2 will provide an understanding of suitable trends in the data that will guide the implementation time categories for these types of fields.

The representation of data in relation to a point in time, also described in Section 3.6.4.4, will require the implementation of a time delta (Witten, Frank, & Hall, 2011), in other words a value that represents the difference between two dates. In the case of the data to be mined, time deltas will need to be introduced that represent the difference between the learner enrolment start date and for example the ETQE's active accreditation. The same considerations as given to the start and end dates of learner enrolment records will need to be given to these values. These values will need to be broken down into new data fields that represent categories of time deltas. Similarly the actual definition of the categories cannot be predicted and will be determine with the assistance of EDM tasks.

### **2.3 Review of literature**

A review of current literature related to DQM shows that most research efforts focus on the development of mathematical and programming methods to correct deficient data. These data deficiencies broadly fall into one of the following data quality classes (Berti-Equille, 2007, p. 106):

- duplicate detection/record matching,
- instance conflict resolution,
- missing/incomplete data, and
- data staleness.

The implementation of data mining techniques to correct data deficiencies generally has a step prior to the actual cleaning of the data which implements some form of data mining technique to identify the data records that have deficient data quality. A review of this type of research has some bearing to the current research project if the methods used can also explain and measure the deficient data quality.

The data mining technique that currently receives the most attention is association rule mining and its application to data quality mining. Association rule mining discovers relationships among variables in a dataset and produces if-then statements in regard to the values of the variables (García, Romero, Ventura, & Calders, 2007, p. 13). The strengths of such if-then statements (rules) are measured in terms of the support and confidence of the rule generated (García, Romero, Ventura, & Calders, 2007, p. 13). The research reviewed describes the implementation of association rule mining to:

Table 2.3.1 Summary of the type of implementation of association rule mining research

Identify data deficiencies on a data set	Natarajan & Koronios (Natarajan, Li, & Koronios, Data Mining Techniques for Data Cleaning, 2009, p. 796) propose that by deriving the association rules from a given dataset, and then using these rules to identify records in the data set that do not adhere to these rules, it is possible to identify data records of deficient quality. Further, the authors propose that a mechanism with which to rank and sort data records that have poor quality can be developed by allocating the sum of the confidence and support values of the association rule to data records that do not adhere to the association rule. The resultant data set can then be reviewed and corrected by a domain expert.
Identify non-enforcement of integrity constraints in data	Mehta & Rajalakshmi (Mehta & Rajalakshmi, 2014, p. 24) propose that by deriving association rules for data fields that contain categorical attributes and then combining rules that have the same antecedent, the derived rules can be utilized to identify data that has deficient quality in terms of the semantic integrity of the data. The description of the data and the nature of the overall implementation of this algorithm imply a non-relational

	data set that contains errors that would not be found in a data set derived from a properly designed relational database.
Determine multi relation association rules in order to identify data deficiencies	Houshmand & Alishahi (Houshmand & Alishahi, 2011, p. 332) propose the implementation of Multi-Relational Data Mining (MRDM) techniques in combination with association rule mining in order to create logical classifications in a data set. These logical classifications can then be utilized to identify data quality deficiencies for reporting purposes.
Identify data quality deficiencies in any number of attributes where a record may contain quality deficiencies in more than one attribute	Alpar & Winkelsträter (Alpar & Winkelsträter, 2014, p. 2261) propose the implementation of association rule mining on numerous attributes in the same data set. The confidence of each rule for each attribute is then allocated to a specific data record and used as a mechanism with which to identify data records that have data quality deficiencies.

Other research that was reviewed suggests the further development of association rule mining:

Table 2.3.2 Summary of the further development of association rule mining research

Optimal association rule mining to identify data deficiencies	Natarajan, Li, & Koronios (Natarajan, Li, & Koronios, Use Rule Based to Predict Dirty Values, 2012, p. 694) propose the implementation of optimal rule discovery in order to address speed issues related to the implementation of association rule mining with low confidence and support values. Optimal rule discovery prunes rules from the resultant association rule set that are considered weak. For example in optimal rule discovery, given two rules the first of which is more generalized than the second, with the second having a confidence level lower than the first, the second rule is pruned from the rule set. The proposed algorithm further implements multiple target rule association rule techniques to improve the prediction accuracy of the rules generated.
Fuzzy association rules in order to discover hidden	Alizamini, Pedram, Alishahi, & Badie (Alizamini, Pedram, Alishahi, & Badie, 2010, p. 469) propose the implementation of

rules in a data set	fuzzy association rules in order to address the limitation that standard association rule techniques have in detecting errors in quantitative data. The algorithm proposed mines the data set using fuzzy association rules and data records are allocated a rank and score of the highest confidence of a rule that the data record violates. The resultant rank and score are then used to determine which records have deficient quality.
---------------------	--

Finally, some research proposed the utilization of association rule mining used in conjunction with:

- feature selection with multi-objective genetic algorithms to detect, quantify and explain data quality (Das & Saha, 2009, p. 106),
- functional dependency to determine the quality of an input transaction prior to implementation on a database (Farzi & Dastjerdi, 2010, p. 116), and
- genetic algorithms to measure data quality on categorical data (Vizhi & Bhuvaneshwari, 2012, p. 40).

The focus on association rule mining, which is an unsupervised machine learning technique, is not unanticipated. Association rule mining is an important aspect of data mining and as a result is one of the best studied data mining tasks (García, Romero, Ventura, & Calders, 2007, p. 13). Association rule mining has some shortcomings in that the algorithm can generate uninteresting rules; that the resulting rules generated are too many, and has low performance (Moreno, Segrera, & López, 2005, p. 317). The research reviewed shows that these shortcomings are still being actively addressed by the data quality mining research community. Whilst considering the popularity of association rule mining and its shortfalls it is difficult to assess whether association rule mining is applicable for this research. The applicability of this data mining technique will as a result have to be tested during the research.

The review of current research in regard to DQM related research with a focus on data cleaning also highlights the utilization of the following additional data mining techniques:

- utilization of functional dependency mining and bagging support vector machines for the identification of data deficiencies and cleaning of data (Natarajan, Li, & Koronios, *Data Mining Techniques for Data Cleaning*, 2009, p. 796),
- utilization of support vector machines as a classification algorithm to identify and correct inconsistencies in a dataset (Grüning, 2007, p. 5),
- improvement of data labels using repeating classification (Sheng, Provost, & Ipeirotis, 2008, p. 614),
- evaluating the accuracy of data records using clustering and fuzzy techniques (Khosravani, 2012, p. 9), and
- development of a clustering algorithm for the identification of contradictions in data (Mehta, Sankarasubramaniam, & Rajalakshmi, 2012, p. 102).

The implementation of functional dependency mining is largely related to the improvement of the design of databases and the determination of data values that require data cleaning. As a result this specific data mining technique is not applicable for the purposes of this research. Support vector machine utilization as described in the research is also predominantly used to determine which data has deficient quality and does little to measure or explain the deficient data and as a result is also not suitable for the purposes of this research.

Classification and clustering techniques both find hidden patterns in a dataset. Clustering is an unsupervised data mining technique that groups related data records together in segments based on having similar values for data variables and is considered an exploratory data mining technique. Classification is a supervised data mining technique that also groups data records together in classes based on having similar values. The clustering technique differs from classification in that the user must define how the classes differ and is used to predict which class a new record would fall into. As already stated this research is descriptive and as a result one would expect that clustering data mining techniques would be applicable for the purposes of this research.

The review of current research related to the development of mathematical and programming methods to correct deficient data did not highlight any research that, besides identification of deficient data quality, would lend itself to the explanation and measurement of deficient data quality.

The review of current research found only one instance that clearly addresses the need to measure and explain data quality deficiencies. The research which was conducted on the 2002 Census of Agriculture utilized classification data mining techniques to identify the characteristics of specific errors in the data (McCarthy & Earp, 2009, p. 1). The research utilized a classification tree model which is constructed by segmenting a dataset using a series of simple rules. Each rule assigns observations to a segment of the data based on the value of one input variable. The rules are chosen to separate the sub-segments in the best way with respect to a chosen target variable. The rules are applied one after another, resulting in a hierarchy of segments within segments.

The hierarchy is called a tree, and each segment is called a node, a segment with all its successors is called a branch, with the final node being called a leaf (McCarthy & Earp, 2009, p. 2). A classification tree was grown for each type of error where the greatest frequency terminal tree nodes provided characteristics of the operations that lead to the error (McCarthy & Earp, 2009, p. 4). This specific research is relevant in that not only does it describe the implementation of a data mining technique that measures and explains a data quality deficiency, it also does this for a data quality deficiency that is semantic in nature. This research shows that even though classification techniques are supervised machine learning techniques they might be applicable to this research.

## **2.4 Chapter summary**

The review of current research assisted in the identification of data mining techniques that lend themselves to the identification, measurement and description of data quality deficiencies.

The association rule data mining technique has been identified as a possible data mining technique for this research. The applicability of this data mining technique will however need to be tested. Clustering has clearly been identified as a data mining technique that should be used for this research. Further, an example in which classification data mining techniques were used for the measurement and description of data quality deficiency means that the research could include classification data mining techniques as well.

The NLRD has never been interrogated from the vantage point of compliance to the semantic business rules described in Section 3.6.2 and therefore the implementation of the EDM technique will need to be utilized to ensure that both the researcher and SAQA can properly visualize the main characteristics of the data being analysed.

The review of current research also highlighted that the KDD process, which will be utilized as the research method for this study, may need to be adjusted to accommodate the large amount of data processing required in order to derive data elements that are needed to evaluate a record's compliance in accordance with the semantic business rules defined in Section 3.6.2. Further, current research in conjunction with the complexity of the required data processing negated the utilization of MRDM data mining techniques. Finally, the review of current research provided insight into the implementation of time categories and time delta categories to address the considerations around representing temporal data as described in Section 3.6.4.4.

## **3 Chapter 3: Methodology**

### **3.1 Introduction**

The first half of this chapter discusses the four basic elements of the research process (Crotty, 1998, p. 3) used to describe, monitor and evaluate the scope and impact of semantic data quality problems in the learner enrolment data on the NLRD using data mining.. These include the

- ontology and epistemology discussed in Section 3.2
- theoretical perspective discussed in Section 3.3
- methodology discussed in Section 3.4, and
- methods discussed in Section 3.5

Thereafter the chapter gives an overview of

- the context of the research,
- the raw data collected from the NLRD,
- overarching data derivation considerations that needed to be considered during the preparation of the data for the data mining,
- identification of the variables used in the research and the pre-processing and derivation conducted on the data, and
- the specific data mining techniques applied to the data.

### **3.2 Research epistemology and ontology**

The proposed research has two main focus areas, firstly, the development of an amended data quality mining framework suited to the assessment of semantic data quality deficiencies, and secondly, to develop an understanding of the semantic data quality deficiencies found in the learner enrolment records in the NLRD data warehouse.

The focus area pertaining to the development of an amended data quality mining framework is constructive in style in that the research aims to construct a new model “... *based on the existing knowledge used in novel ways, with possibly adding a few missing links.*” (Crnkovic, 2010, p. 360).

The focus area that addresses the development of an understanding of the semantic data quality deficiencies in the learner enrolment records in the NLRD data warehouse will be conducted by objective measurement methods (Easterby-Smith, Thorpe, & Lowe, 2002, p. 28) across a predefined set of variables, and as a result can be considered to be scientifically positivistic. This positivist philosophical assumption employed in this study ensured quantifiable measures of variables, drawing of inferences, understanding of relationships within an occurrence using structured instruments (Orlikowski & Baroudi, 1991, p. 9).

### 3.3 Theoretical aspect of the research

The research is related to data and information quality research and thus the “Framework for Data and Information Quality Research” proposed by Madnick et al. (Madnick, Wang, & Lee, 2009) is applicable. The framework is pragmatic in that it is based on two principles that assume that research methods and types continue to evolve in the data and information quality research (Madnick, Wang, & Lee, 2009, p. 5). As a result the framework has two dimensions - topics and methods - and assumes that any research related to data and information quality addresses certain topics using certain research methods (Madnick, Wang, & Lee, 2009, p. 5).

Topics	Methods
1. Data quality impact 1.1 Application area (e.g., CRM, KM, SCM, ERP) 1.2 Performance, cost/benefit, operations 1.3 IT management 1.4 Organizational change, processes 1.5 Strategy, policy 2. Database related technical solutions for data quality 2.1 Data integration, data warehouse 2.2 Enterprise architecture, conceptual modelling 2.3 Entity resolution, record linkage, corporate householding 2.4 Monitoring, cleansing 2.5 Lineage, provenance, source tagging 2.6 Uncertainty (e.g., imprecise, fuzzy data) 3. Data quality in the context of computer science and IT 3.1 Measurement, assessment 3.2 Information systems 3.3 Networks 3.4 Privacy 3.5 Protocols, standards	1. Action research 2. Artificial Intelligence 3. Case study 4. Data mining 5. Design science 6. Econometrics 7. Empirical 8. Experimental 9. Mathematical modelling 10. Qualitative 11. Quantitative 12. Statistical analysis 13. System design, implementation 14. Theory and formal proofs

Topics	Methods
3.6 Security 4. Data quality in curation 4.1 Curation - Standards and policies 4.2 Curation - Technical solutions	

Figure 3.2.1 Two dimensional matrix of the “Framework for Data and Information Quality Research”

(Madnick, Wang, & Lee, 2009, p. 6)

According to this framework the proposed research falls into the major research topic “Data Quality in the Context of Computer Science and Information Technology” which is defined as research that “... *develops technologies and methods to manage, ensure and enhance data quality*” (Madnick, Wang, & Lee, 2009, p. 11). Further, the sub-category of the research topic is “Measurement, Assessment” which is defined as the development of techniques for the systematic measurement of data quality within organizations or within the context of a particular application (Madnick, Wang, & Lee, 2009, p. 11).

The framework method for the proposed research is data mining, which the framework defines as a high level category research method and is recognized as being applicable for addressing data quality issues (Madnick, Wang, & Lee, 2009, p. 15).

The framework provided by Madnick et al. (Madnick, Wang, & Lee, 2009) falls short in providing a theoretical foundation for the data mining process. Although data mining is an applied area, a theoretical framework is required in order to maintain the focus of the study.

The establishment of a theoretical framework for the application of data mining is still ongoing. Current research being conducted by De Bie is notable, however the researcher notes that the work done constitutes only a starting point for the establishment of such a theoretical framework (De Bie, An Information Theoretic Framework for Data Mining, 2011) (De Bie & Spyropoulou, A Theoretical Framework for Exploratory Data Mining: Recent Insights and Challenges Ahead, 2013, p. 615) . A theoretical framework that addresses both KDD and data mining suggests that the following technical criteria should be considered when applying data mining (Fayyad, Piatetsky-Shapiro, & Smyth, Knowledge Discovery and Data Mining: Towards a Unifying Framework, 1996). The technical criteria provided by Fayyad et al.

(Fayyad, Piatetsky-Shapiro, & Smyth, Knowledge Discovery and Data Mining: Towards a Unifying Framework, 1996) provide sufficient guidelines to maintain the focus of the study and would need to be applied as follows:

1. When complex patterns are sought or when there are a large number of attributes in the data being investigated a large volume of data must be available to be mined.

The data mined for the study would need to include all records that are applicable to the study. Further, the attributes mined would need to be limited in scope to include only those attributes that are relevant to the study.

2. The data attributes that are used must be relevant to the discovery task.

In order to determine which data attributes are relevant to the discovery task, an analysis must be completed of the data structures of the NLRD in order to correctly identify which data attributes should be included in the study.

3. The data must contain few data errors.

Where data errors are encountered in the data, logic would need to be implemented to ensure ease of identification and exclusion of these records from the study.

4. Time orientated data must be handled in a manner that allows the application of the data mining task to be retrained on a newer version of data.

Temporal data will need to be encoded in such a manner as to ensure ease of interpretation of the results and future mining of new data.

5. An understanding of the domain should guide which attributes are important, what relationships in the data are likely, which patterns are already known, and which patterns have utility for the user.

A review of all relevant policies, acts and legislation must be conducted in order to determine the relevant attributes for the study and the nature of expected relationships in the data being minded.

Figure 3.3.1 illustrates the manner in which these criteria are applied in this research.

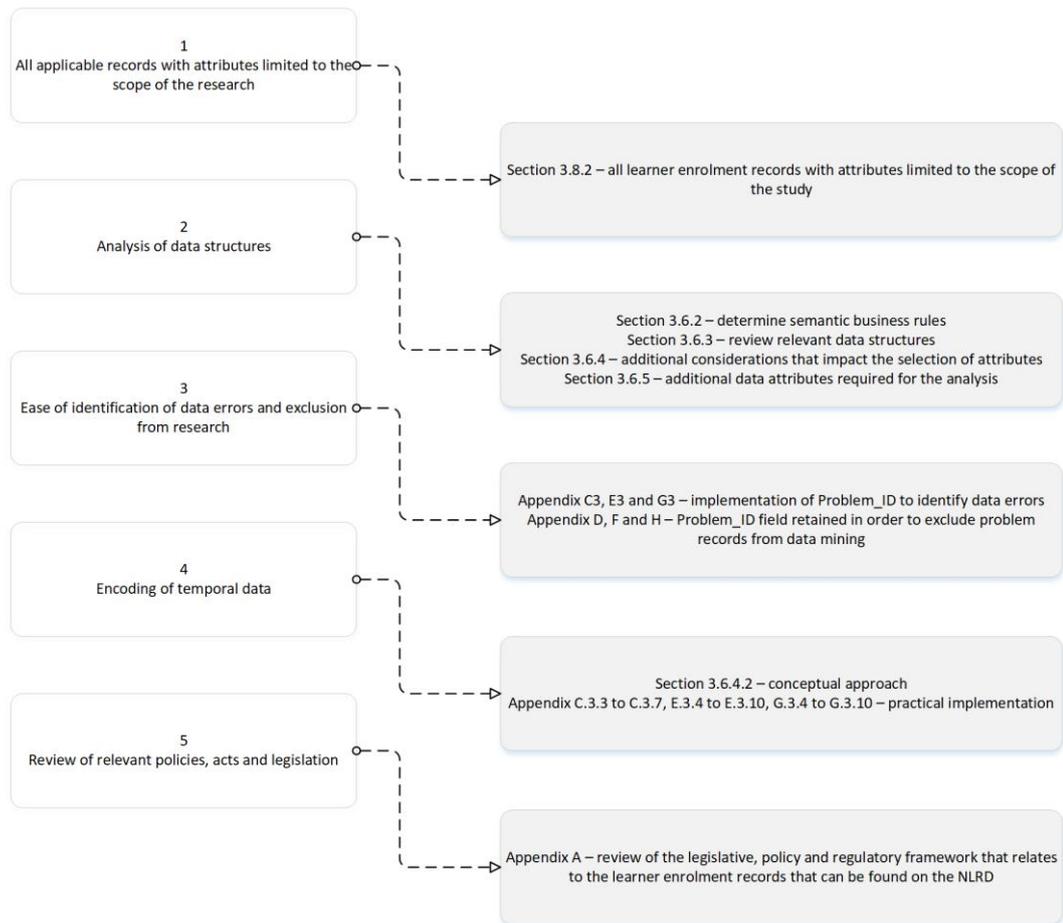


Figure 3.3.1 Diagrammatic representation of application of criteria in the research

### 3.4 Research methodology

The process of research has two main classifications, namely qualitative and quantitative. Creswell defines quantitative research as research that has traditionally provided a measurement orientation in which data can be gathered from many individuals and trends assessed across large geographic regions (Creswell, 2011). Further, Creswell defines qualitative research as research that yields detailed information reported in the voice of the participants and contextualized in the settings in which they provide experiences and the meanings of their experiences (Creswell, 2011). Table 3.4.1 below presents a comparison of the attributes of qualitative and quantitative research methods (Creswell, 2011):

Table 3.4.1 Comparing characteristics of qualitative and quantitative research  
(Creswell, 2011)

<b>Attribute</b>	<b>Quantitative</b>	<b>Qualitative</b>
Research problem	Describe and justify	Explore and understand
Literature review	Major role	Minor role
Purpose statements, research questions, and hypotheses	Specific, narrow, measurable and observable	General and broad
Data	Large, structured, small number of variables	Small, unstructured, large number of variables
Analysis	Statistical and compared to predictions or previous results	Themes and interpretation of larger meaning of results
Report	Fixed structures and evaluation criteria	Flexible, emerging structures and evaluation criteria
Results	Objective, unbiased approach	Researcher's subjective and reflexivity bias

The above clearly illustrates the expectation that exploratory research could be qualitative and quantitative. However, the research differs from the attributes of only qualitative research in that the data for the research will be large (in excess of 85 million records), structured and with a small number of variables, and the results of the research will be predominantly quantitative in nature. As a result, the proposed research will be quantitative.

Attempting to discover trends or patterns from a dataset as large as the learner enrolment records stored in the NLRD data warehouse cannot be sought using traditional manual research approaches. Gaining meaningful understanding of this data will require the utilization of data mining techniques. Data mining however is not in itself a research design; rather it is the application of specific algorithms for the extraction of patterns in data (Fayyad, Piatetsky-Shapiro, & Smyth, From Data Mining to Knowledge Discovery in Databases, 1996, p. 39).

There are two broad cultures in the data mining field, the first - a statistical culture - emphasizes the role of predictive modelling and the second - an artificial intelligence -

emphasizes the role of knowledge discovery (Sumathi & Sivanandam, 2006, p. 238). This study does not focus on the prediction of future trends in data; rather, it strives to extract useful information from a large dataset. As a result, this study required the implementation of data mining techniques that focus on knowledge discovery and the study falls within the field of KDD. The KDD is defined as follows (Fayyad, Piatetsky-Shapiro, & Smyth, The KDD Process for Extracting Useful Knowledge from Volumes of Data, 1996, p. 30):

“Knowledge Discovery in Databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”

Data mining is a step within the overall process model of KDD. There are generally three different types of KDD process models, namely academic, industrial and hybrid (Crios, Pedryzc, Swiniarski, & Kurgan, 2007). The academic model developed by Fayyad et al. (Fayyad, Piatetsky-Shapiro, & Smyth, The KDD Process for Extracting Useful Knowledge from Volumes of Data, 1996, p. 30) has been used for this study and contains the following 9 discrete steps that are both interactive and are used in an iterative manner:

1. Understand the application domain
2. Create a target dataset
3. Clean and pre-process the dataset
4. Reduce and project the dataset
5. Choose the data mining function
6. Choose the data mining algorithm
7. Mine the dataset
8. Interpret the data mining results
9. Utilize the discovered knowledge

These discrete steps, combined with the interactive and iterative nature of the process model, are exploratory in nature and are thus suitable for an exploratory study such as this one.

### **3.5 Research method**

The research was initiated with a literature review, which focused on determining which data mining techniques have been used in other studies to identify, measure and describe data quality deficiencies. Further, the review determined which of the identified data mining

techniques specifically lend themselves to the identification, measurement and description of semantic data quality deficiencies.

Having identified the most applicable data mining techniques, the academic KDD process was implemented as the methodology for the mining of the data. The process contains 9 discrete steps (Fayyad, Piatetsky-Shapiro, & Smyth, *The KDD Process for Extracting Useful Knowledge from Volumes of Data*, 1996, p. 30):

1. Understand the application domain
2. Create a target dataset
3. Clean and pre-process the dataset
4. Reduce and project the dataset
5. Choose the data mining function
6. Choose the data mining algorithm
7. Mine the dataset
8. Interpret the data mining results
9. Utilize the discovered knowledge

These 9 steps fall into 5 different phases (see Figure 3.5.1):

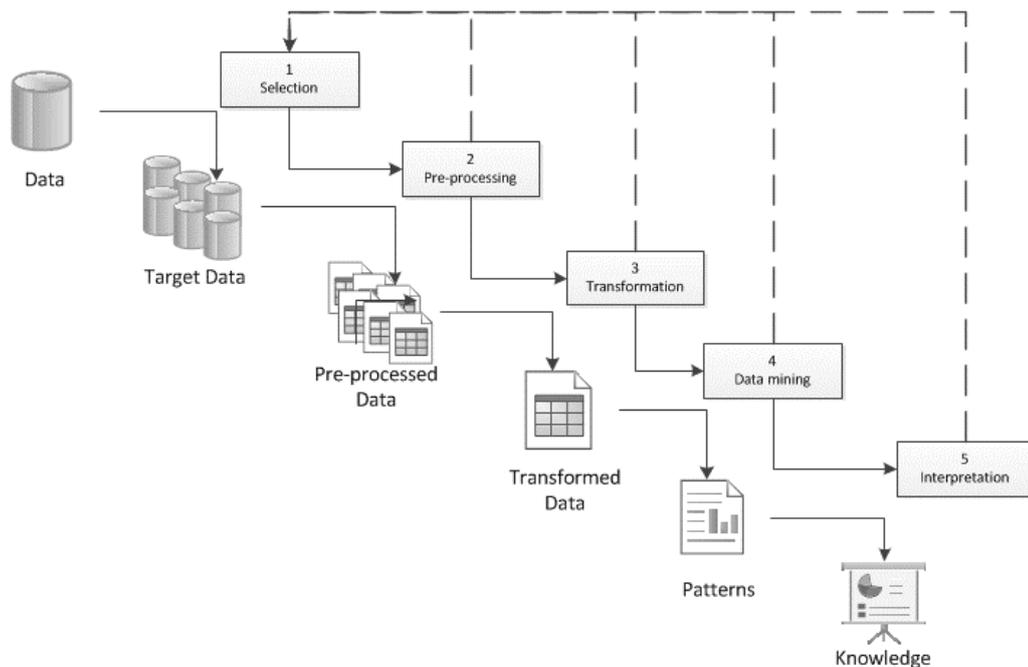


Figure 3.5.1 Diagrammatic representation of the KDD process

1. Selection

The selection phase focused on selecting a sub-set of the data that is concise enough to be processed within a reasonable time period whilst also large enough to contain a representation of the specific data quality dimension being studied.

2. Pre-processing

During the pre-processing phase the study addresses the cleaning of data in regard to missing data values and the removal of statistical noise (i.e. unexplained variations in the data).

3. Transformation

The transformation phase focuses on determining which data fields in the learner enrolment tables need to be utilized in order to provide meaningful patterns in regard to the data quality dimension being studied. Further, the transformation phase also determines data fields related to the data quality dimension that are found in data tables other than the learner enrolment tables of the NLRD data warehouse, such as provider accreditation details, assessor accreditation details etc. The transformation phase resulted in a constrained dataset with limited features from both the learner enrolment tables and other tables as required for the data mining process.

4. Data-mining

The data mining phase entailed the selection of a relevant data mining task and one or more data mining techniques identified in the literature review with which to mine the data.

5. Interpretation and Evaluation

The post data mining step entailed the analysis of the results of the data mining in which interesting patterns that represent knowledge are identified. The pre-final step of the overall process was the validation of the results found by applying the same data mining techniques to the complete NLRD data warehouse dataset to ensure that the same patterns occur in either a new subset of data or the complete dataset.

The final step in the overall process entailed documenting the new understanding of the data in the NLRD data warehouse and the development of recommendations that can be used as a basis for change.

The KDD process is both interactive and iterative and required numerous decisions to be made during the course of the overall process. All decisions and results were deliberated with domain experts. In this study the primary domain expert was the Director of the NLRD at SAQA.

### 3.6 Research context

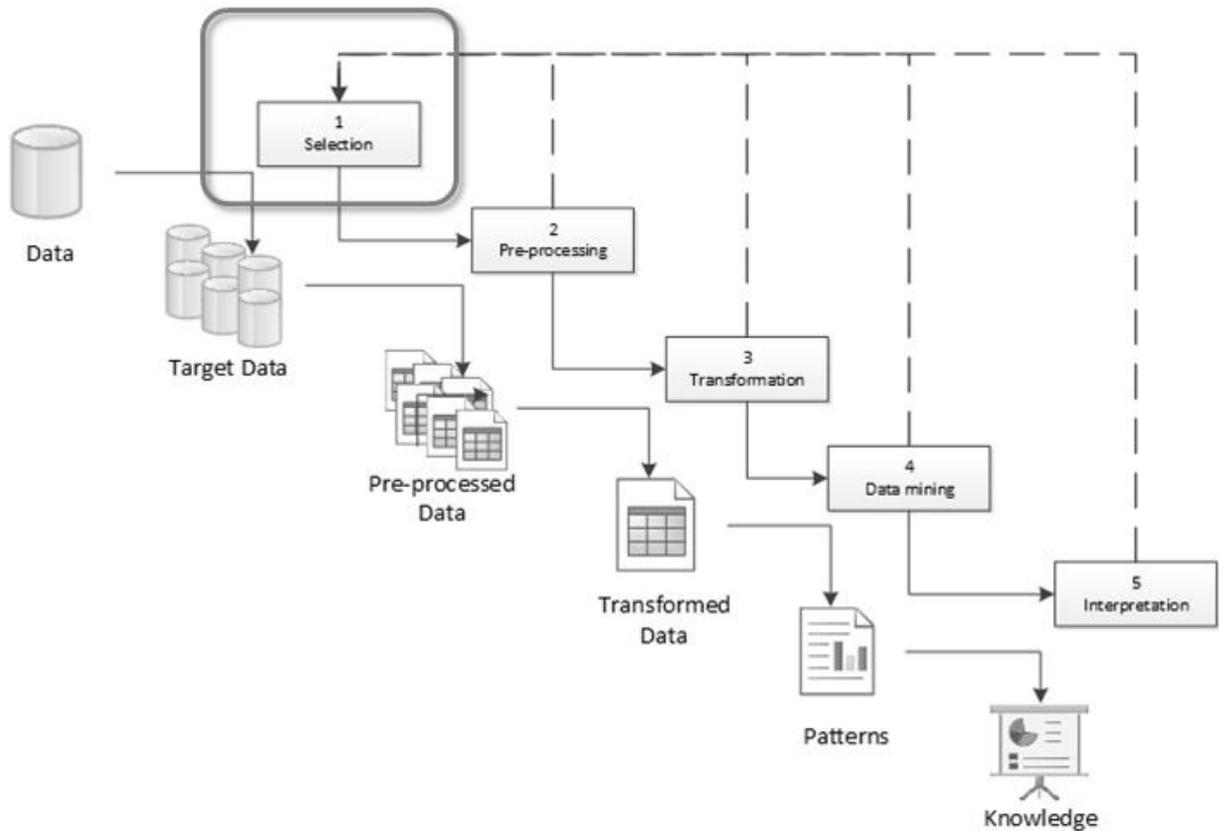


Figure 3.6.1 KDD phase - Selection

#### 3.6.1 Introduction

This section commences with a review to the legislative, policy and regulatory framework that relates to learner enrolment records stored on the NLRD. The review highlights ten (10) discrete semantic data quality aspects which are identified as the core semantic business rules for this research.

Based on these semantic business rules a review of the NLRD data tables that are related to the semantic business rules was conducted. The results of the review are presented as a generalized description of the physical data structures in the NLRD. The data tables and

data fields that relate to the determination of the compliance of a learner enrolment record in regard to each specific semantic business rule are described.

The description of the physical data structures in the NLRD that inform the semantic business rules raises specific aspects in regard to the manner in which the data is stored in the NLRD. As a result this section is followed by a section that specifically addresses some considerations that impact on the manner in which the data needs to be prepared in order to derive useful information from the analysis of the data.

Finally, a review is conducted of other data fields that are present in the NLRD learner enrolment record and parent tables of the learner enrolment record. Based on this review, additional data fields that could prove useful during the analysis of the data are identified for this research.

This review in conjunction with the review of the data structures that inform the semantic business are instrumental in the development of an understanding of the data tables and data fields in the NLRD that will form the basis of this research.

### ***3.6.2 Defining the semantic business rules***

In order to define the semantic business rules that are applicable to this study a review of the legislative, policy and regulatory framework that relates to the learner enrolment records that can be found on the NLRD was conducted (see Appendix A). The review introduced six discrete concepts that form part of the South African National Qualification Framework (NQF) namely; Education and Training Quality Assurance Bodies, providers, assessor, learnerships, qualifications and unit standards.

The review further highlights how these six concepts relate to learner enrolment records and form the basis of ten (10) discrete semantic business rules that form the core of this study (see Appendix A):

1. that the ETQE that submitted the record
  - a. was accredited for the duration of the learner's active enrolment on the learnership/qualification/unit standard

- b. was accredited to quality assure the qualification/unit standard for the duration of the learner's active enrolment on the qualification/unit standard
- 2. that the provider
  - a. was accredited for the duration of the learner's active enrolment on the learnership/qualification/unit standard
  - b. was accredited to offer the qualification/unit standard for the duration of the learner's active enrolment on the learnership/qualification/unit standard
- 3. that if the learner has completed the learnership or achieved the qualification/unit standard and the details of the assessor are supplied, that the assessor
  - a. was registered at the time of the completion of the learnership or achievement of the qualification/unit standard
  - b. was registered to assess the qualification/unit standard at the time of the completion of the learnership or achievement of the qualification/unit standard
- 4. that the qualification/unit standard was registered for the duration of the learner's active enrolment on the qualification/unit standard
- 5. that if the learner has completed the learnership, then due to the intrinsic nature of a learnership and qualification the learner would have achieved the qualification on or before the completion of the learnership
- 6. that if the learner has achieved the qualification, and the qualification is a unit standards based qualification
  - a. the learner would have achieved the minimum required number of credits for the qualification, on or before the achievement of the qualification, based on the achievement of unit standards related to the qualification
  - b. the learner would have achieved the correct range of credits for the qualification, on or before the achievement of the qualification, based on the achievement of unit standards that have been defined as core, fundamental and elective unit standards for the qualification

### ***3.6.3 Analysis of the data structures***

Having defined the semantic business rules that form the basis of this research this section entails a review of the data structures in the NLRD that store the data that are related to the semantic business rules. The data structures for each semantic business rule and/or sub-rule is briefly described and illustrated.

Thus far the following conceptual figure (see Figure 3.6.3.1) has been utilized to acclimatize the reader to the concepts of the NQF that relate to this research. The same type of figure will be utilized in this section to orientate the reader in regard to the specific NQF concepts that are participants to a specific rule.

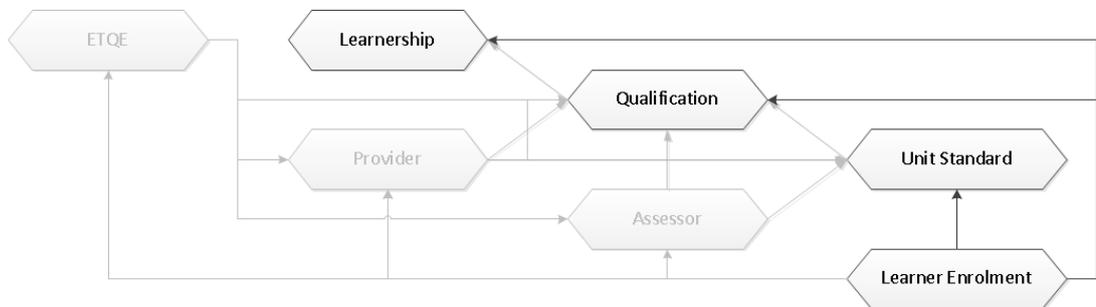


Figure 3.6.3.1 Conceptual diagram of the learner enrolment record

The reader should note the following in regard to the manner in which the storage of the data in the NLRD is described in this section:

- In order to keep this section succinct, the diagrams that follow provide only conceptual illustrations of the manner in which the data is saved in the NLRD.
- Although the diagrams make reference to “History” tables in certain instances, these tables are physically manifested as audit tables on the NLRD from which a record of history for a specific data record can be derived.
- The diagrams depict that start and end dates are NOT NULL fields. This constraint holds true only for active records in the respective NLRD tables.

1. The ETQE that submitted the record

- a. was accredited for the duration of the learner’s active enrolment on the learnership/qualification/unit standard

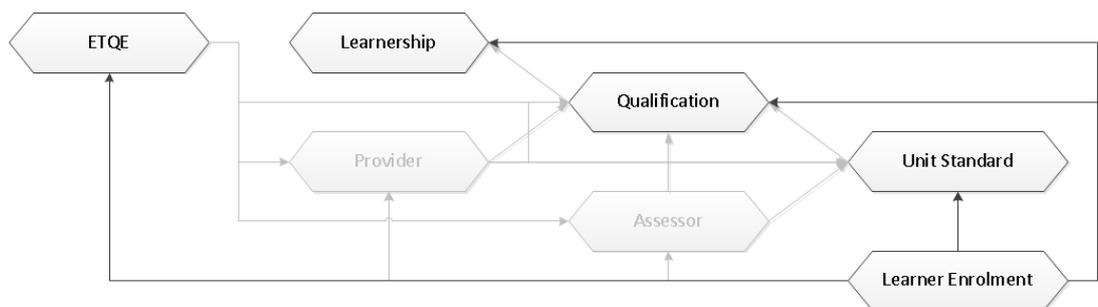


Figure 3.6.3.1.a.1 Conceptual diagram of the NQF concepts that participate in business rule 1.a

The linkage between a learner enrolment record and an ETQE is determined based on the ETQE ID of the ETQE that submitted the record to the NLRD which is stored on the learner enrolment record. Data in regard to an ETQE is maintained on the NLRD by SAQA.

The accreditation status of the ETQE for the duration of the learner’s active enrolment is derived from the tables ETQE and ETQE History using a combination of status and start and end date.

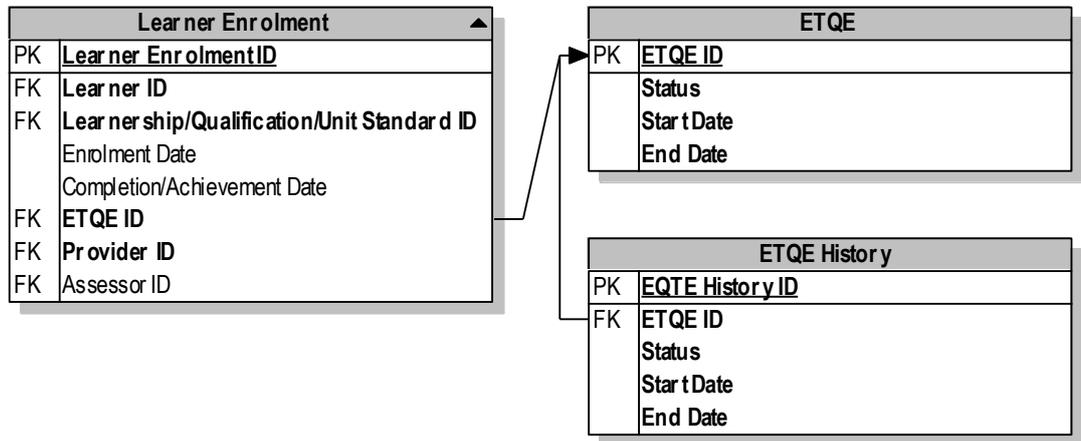


Figure 3.6.3.1.a.2 Conceptual diagram of the tables and fields that inform business rule 1.a

- b. was accredited to quality assure the qualification/unit standard for the duration of the learner’s active enrolment on the qualification/unit standard

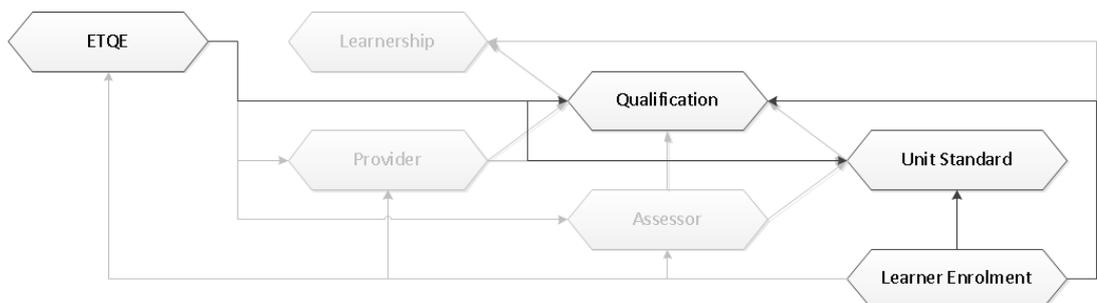


Figure 3.6.3.1.b.1 Conceptual diagram of the NQF concepts that participate in business rule 1.b

As stated previously the linkage between a learner enrolment record and an ETQE is determined based on the ETQE ID of the ETQE that submitted the record to the NLRD which is stored on the learner enrolment record. The accreditation of an ETQE to quality assure a specific qualification/unit standard is however not recorded against the learner enrolment record. Data in regard to the accreditation of an ETQE to quality assure a qualification/unit standard is maintained on the NLRD by SAQA.

The accreditation status of the ETQE to quality assure the qualification/unit standard for the duration of the learner’s active enrolment is derived from the tables ETQE Accreditation and ETQE Accreditation History using a combination of status and start and end date. Deriving this information is done independently from the ETQE and ETQE History tables.

There may be instances where records of ETQE accreditation to quality assure a qualification/unit standard fall outside of the scope of the ETQE’s overall accreditation. This would be considered a data capturing error and will be reported to SAQA for further action or interpretation. The analysis of this type of issue falls outside of the scope of this research.

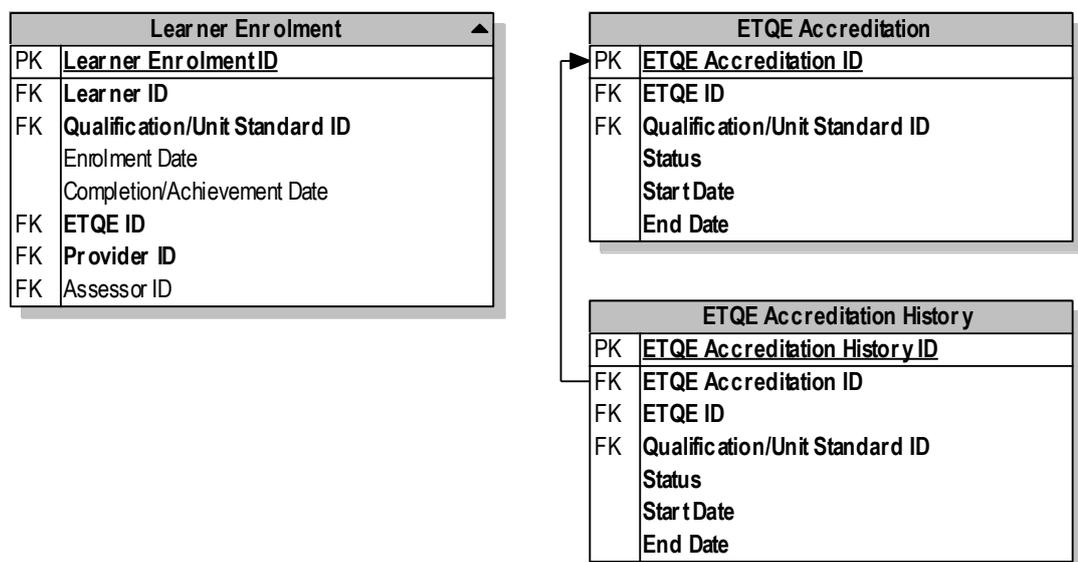


Figure 3.6.3.1.b.2 Conceptual diagram of the tables and fields that inform business rule

1.b

2. The provider

- a. was accredited for the duration of the learner’s active enrolment on the learnership/qualification/unit standard

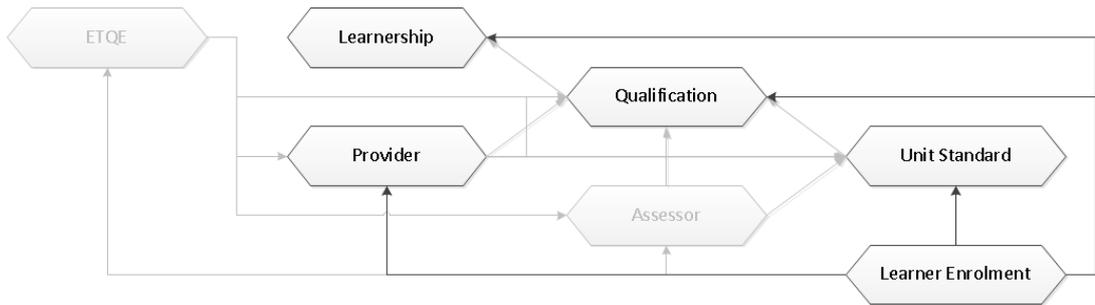


Figure 3.6.3.2.a.1 Conceptual diagram of the NQF concepts that participate in business rule 2.a

The linkage between a learner enrolment record and a provider is determined based on the Provider ID value which is submitted to the NLRD by the ETQE as part of the learner enrolment record. Data in regard to a provider is maintained and submitted to the NLRD by the ETQE.

The accreditation status of the provider for the duration of the learner’s active enrolment is derived from the tables Provider and Provider History using a combination of status and start and end date.

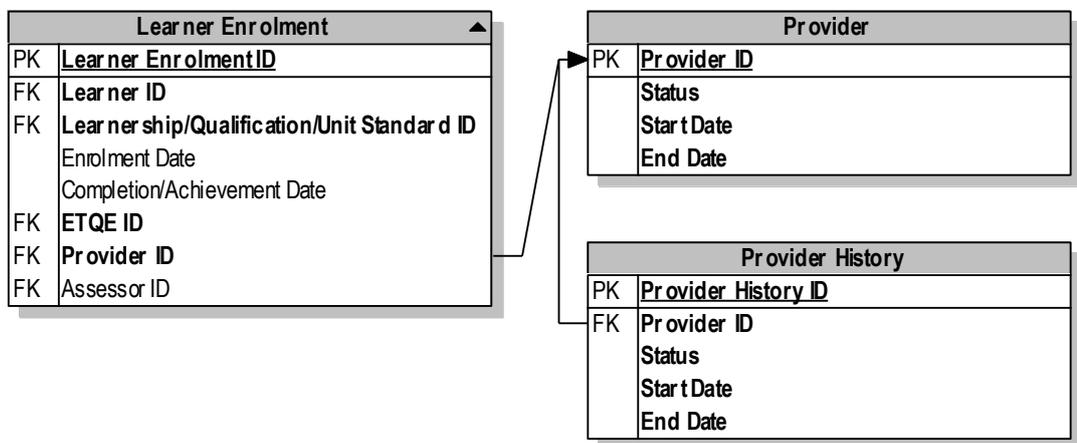


Figure 3.6.3.2.a.2 Conceptual diagram of the tables and fields that inform business rule 2.a

- b. was accredited to offer the qualification/unit standard for the duration of the learner’s active enrolment on the qualification/unit standard

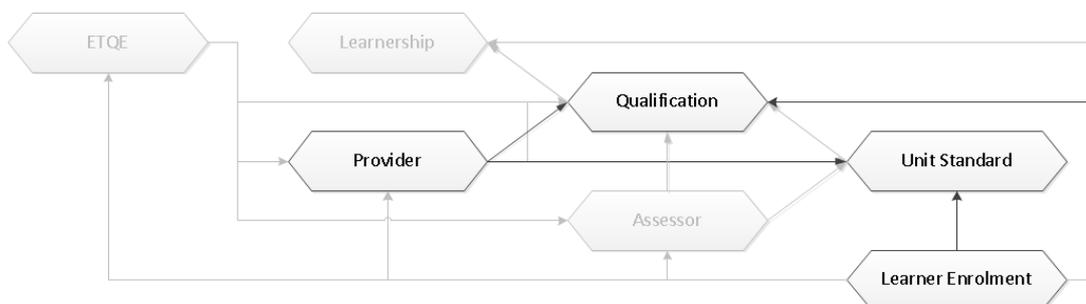


Figure 3.6.3.2.b.1 Conceptual diagram of the NQF concepts that participate in business rule 2.b

As stated previously the linkage between a learner enrolment record and a provider is determined based on the Provider ID value which is submitted to the NLRD by the ETQE as part of the learner enrolment record. The accreditation of a provider to offer a specific qualification/unit standard is however not recorded against the learner enrolment record. Data in regard to the accreditation of providers to offer a qualification/unit standard is maintained and submitted to the NLRD by the ETQE.

The accreditation status of the provider to offer the qualification/unit standard for the duration of the learner’s active enrolment is derived from the tables Provider Accreditation and Provider Accreditation History using a combination of status and start and end date. Deriving this information is done independently from the Provider and Provider History tables.

There may be instances where provider records for accreditation to offer a qualification/unit standard fall outside of the scope of the provider’s overall accreditation. This would be considered a data capturing error and will be reported to SAQA for further action or interpretation. The analysis of this type of issue falls outside of the scope of this research.

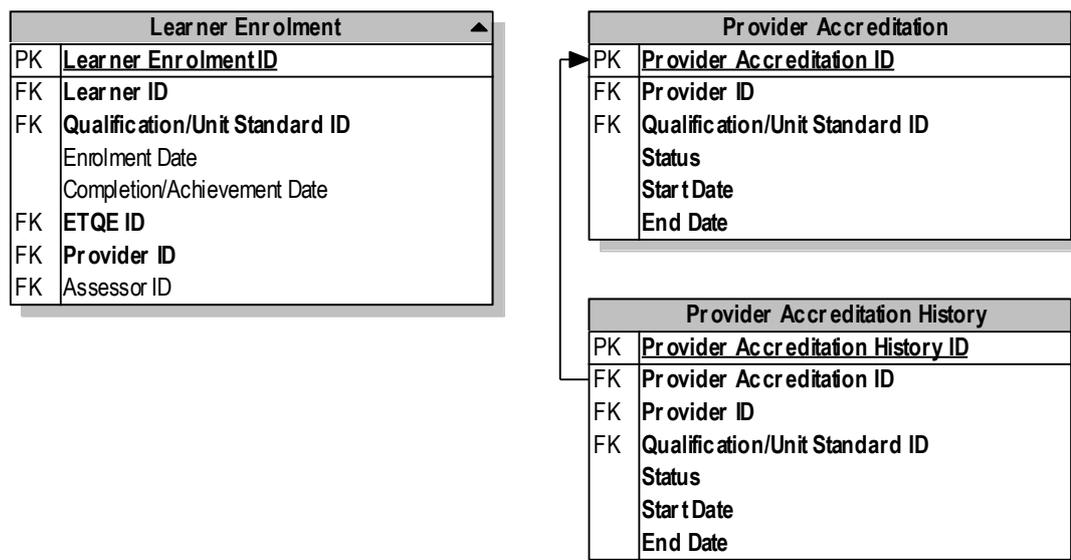


Figure 3.6.3.2.b.2 Conceptual diagram of the tables and fields that inform business rule 2.b.

3. If the learner has completed the learnership or achieved the qualification/unit standard and the details of the assessor are supplied, that the assessor
  - a. was registered at the time of the completion of the learnership or achievement of the qualification/unit standard

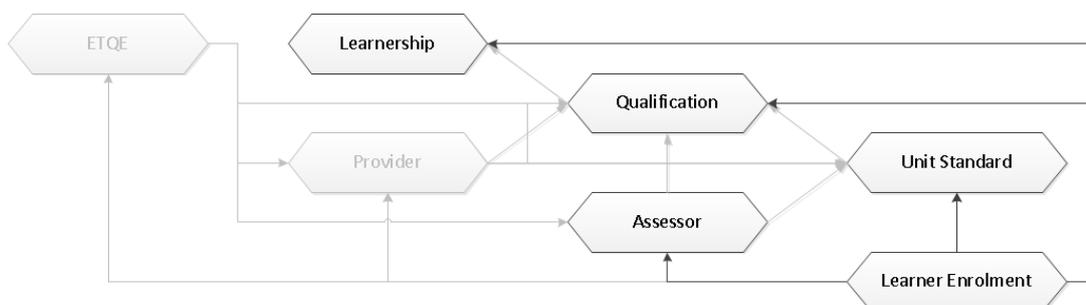


Figure 3.6.3.3.a.1 Conceptual diagram of the NQF concepts that participate in business rule 3.a

The linkage between a learner enrolment record with a completed/achieved status and an assessor is determined based on the Assessor ID value which is submitted to the NLRD by the ETQE as part of the learner enrolment record. Data in regard to an assessor is maintained and submitted to the NLRD by the ETQE.

The registration status of the assessor at the time of the completion of the learnership or achievement of the qualification/unit standard is derived from the tables Assessor and Assessor History using a combination of status and start and end date.

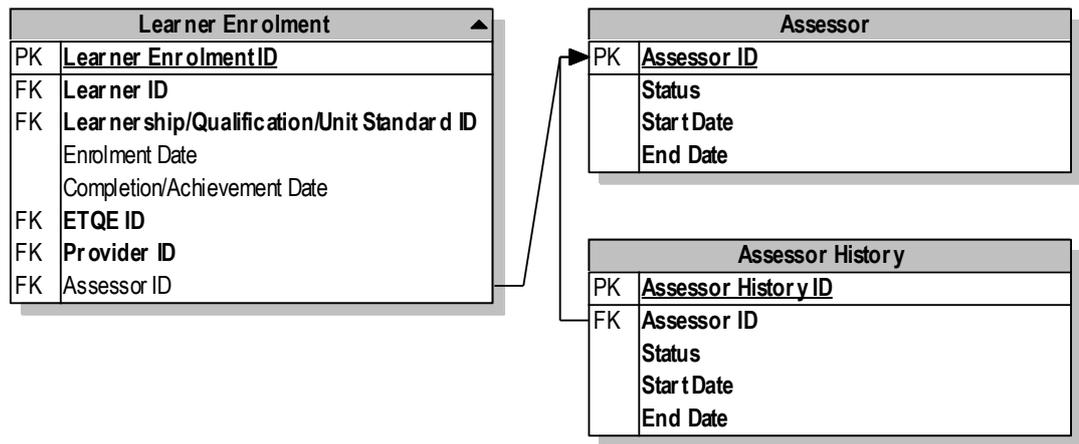


Figure 3.6.3.3.a.2 Conceptual diagram of the tables and fields that inform business rule 3.a

b. was registered to assess the qualification/unit standard at the time of the achievement of the qualification/unit standard

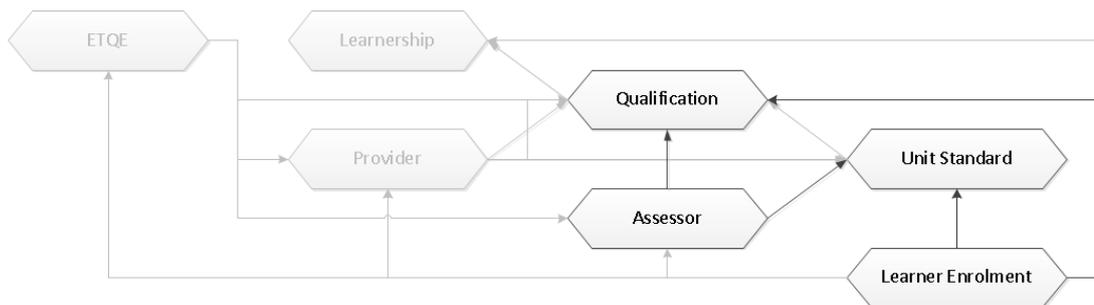


Figure 3.6.3.3.b.1 Conceptual diagram of the NQF concepts that participate in business rule 3.b

As stated previously, the linkage between a learner enrolment record that has been completed/achieved and an assessor is determined based on the Assessor ID value which is submitted to the NLRD by the ETQE as part of the learner enrolment record. The registration of an assessor to assess a specific

qualification/unit standard is however not recorded against the learner enrolment record. Data in regard to the registration of assessors to assess a qualification/unit standard is maintained and submitted to the NLRD by the ETQE.

The registration status of the assessor to assess the qualification/unit standard at the time of the achievement of the qualification/unit standard is derived from the tables Assessor Registration and Assessor Registration History using a combination of status and start and end date. Deriving this information is done independently from the Assessor and Assessor History tables.

There may be instances where assessor records for registration to assess a qualification/unit standard fall outside of the scope of the assessor’s overall registration. This would be considered a data capturing error and will be reported to SAQA for further action or interpretation. The analysis of this type of issue falls outside of the scope of this research.

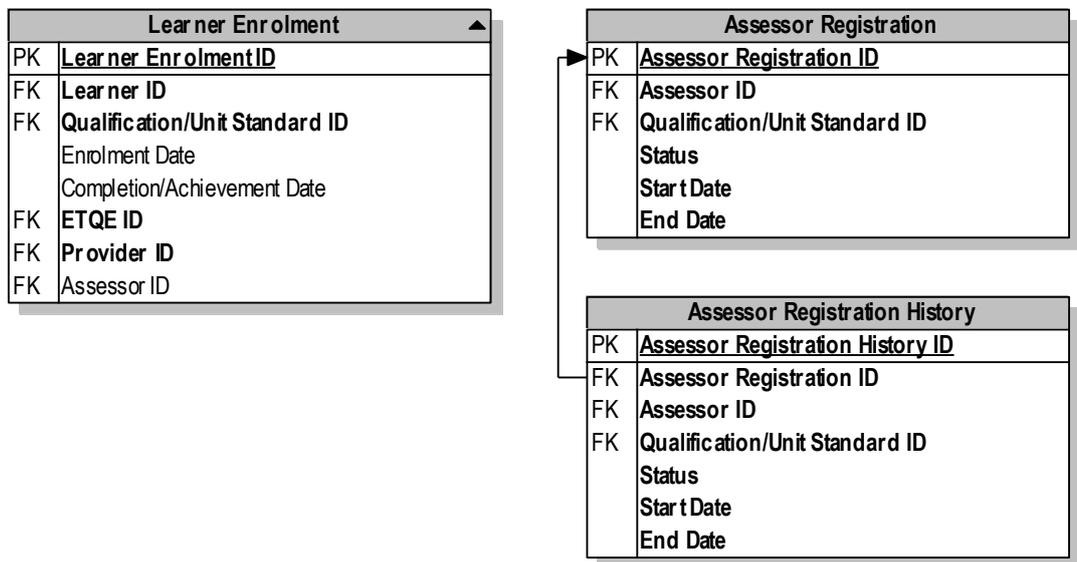


Figure 3.6.3.3.b.2 Conceptual diagram of the tables and fields that inform business rule

3.b

4. The learnership/qualification/unit standard was registered for the duration of the learner’s active enrolment on the learnership/qualification/unit standard

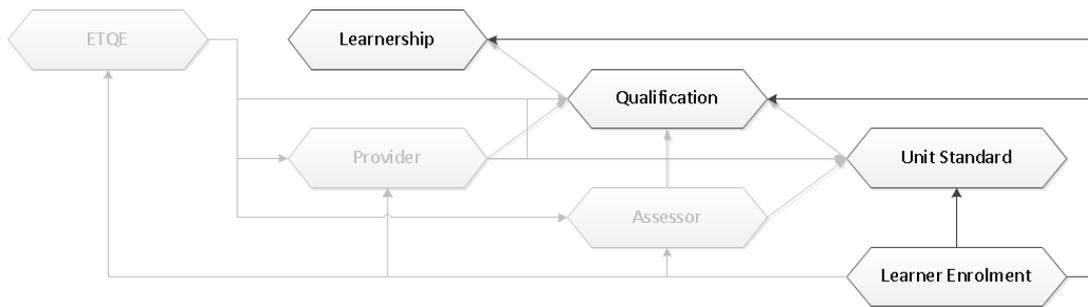


Figure 3.6.3.4.1 Conceptual diagram of the NQF concepts that participate in business rule

4

The linkage between a learner enrolment record and a learnership/qualification/unit standard is determined based on the Learnership/Qualification/Unit Standard ID value which is submitted to the NLRD by the ETQE as part of the learner enrolment record. Data in regard to a learnership/qualification/unit standard is maintained on the NLRD by SAQA.

The registration status of the learnership/qualification/unit standard for the duration of the learner's active enrolment is derived from the tables Learnership/Qualification/Unit Standard and Learnership/Qualification/Unit Standard History using a combination of status and start and end date. The manner in which this will be derived will be sensitive to the fact that if a learner engaged on a unit standard in order to acquire the required number of credits for a specific qualification, then the registration status and term of the qualification is applicable.

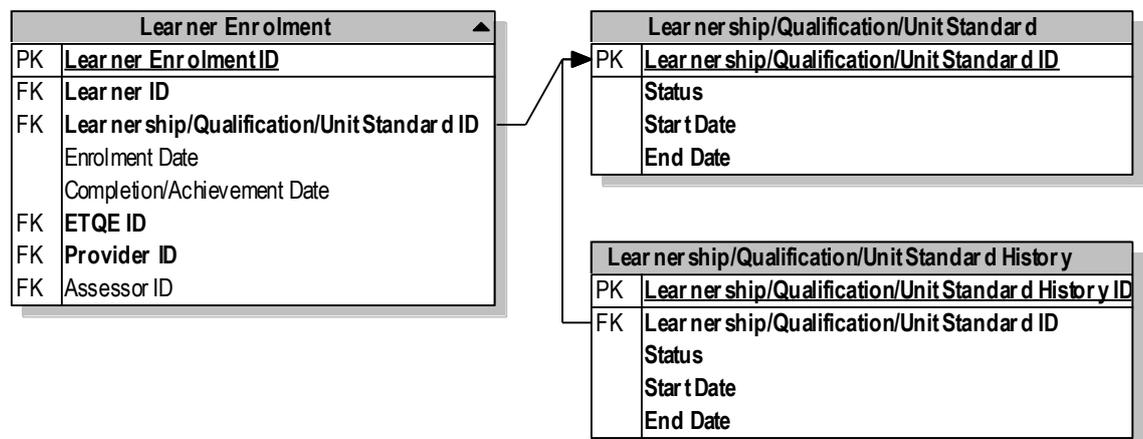


Figure 3.6.3.4.2 Conceptual diagram of the tables and fields that inform business rule 4

5. If the learner has completed the learnership, that due to the intrinsic nature of a learnership and qualification the learner would have achieved the qualification on or before the completion of the learnership

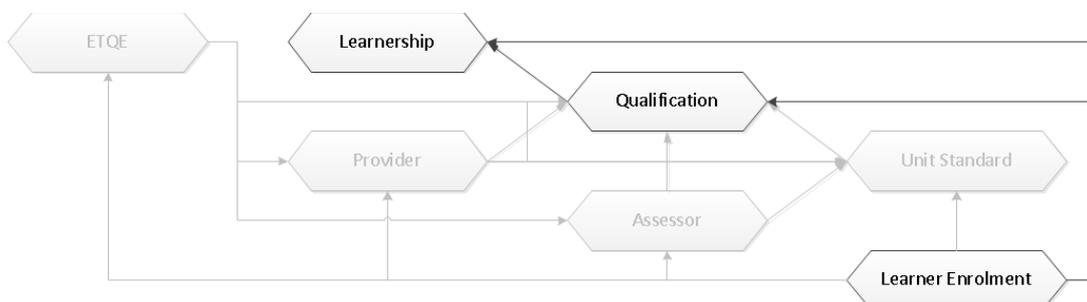


Figure 3.6.3.5.1 Conceptual diagram of the NQF concepts that participate in business rule 5

The linkage between a learnership completion record and a qualification achievement record is determined based on the Learnership ID value on the qualification achievement record which is submitted to the NLRD by the ETQE. Learnership records, qualification records and the required relationship between a learnership and a qualification are maintained on the NLRD by SAQA.

Achievement of the relevant qualification for a specific learnership completion record is derived from the tables Learnership Completion, Qualification Achievement and Learnership Qualification Link.

There may be instances where learnership records are not linked to qualification records. This would be considered a data capturing error and will be reported to SAQA for further action or interpretation. The analysis of this type of issue falls outside of the scope of this research.

There may also be instances where:

1. No matching qualification enrolment record can be found for the learnership completion record
2. A matching qualification enrolment record can be found for the learnership completion record, the qualification enrolment record however does not have a Learnership ID recorded against it

3. A matching qualification enrolment record can be found for the learnership completion record, the qualification enrolment record however has the incorrect Learnership ID recorded against it

Although these could also be considered data capturing errors, these types of issues do fall within the scope of this research.

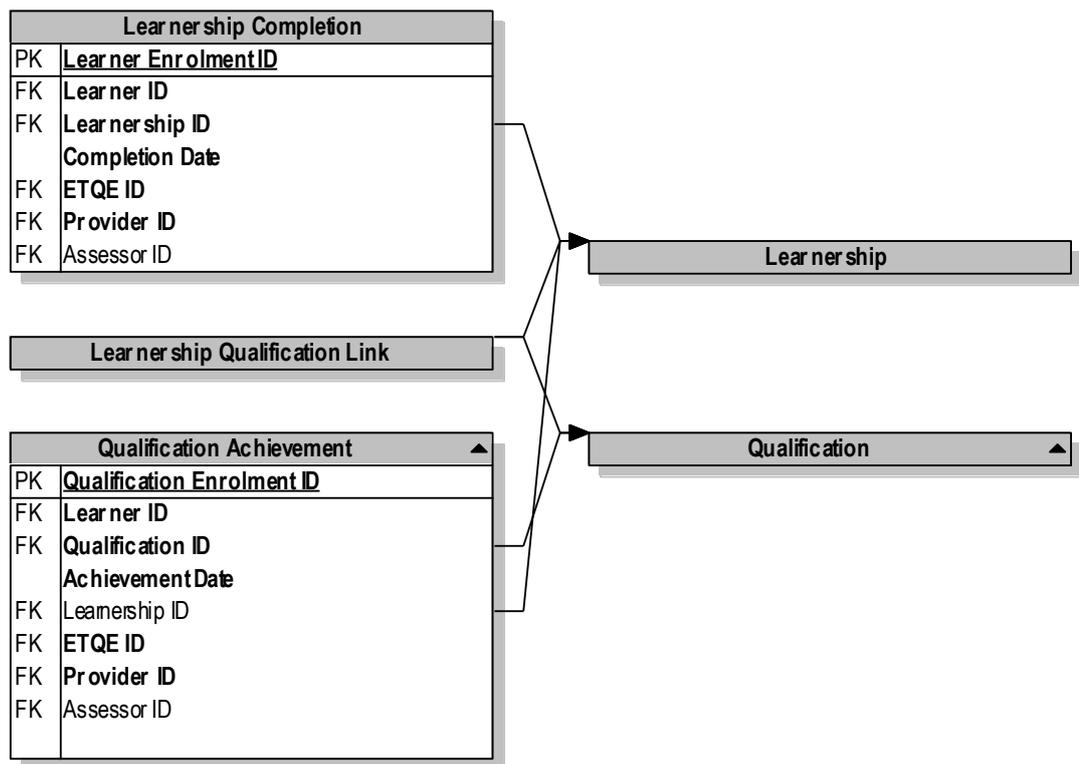


Figure 3.6.3.5.2 Conceptual diagram of the tables and fields that inform business rule 5

6. If the learner has achieved the qualification, and the qualification is a unit standards based qualification
  - a. the learner would have achieved the minimum required number of credits for the qualification, on or before the achievement of the qualification, based on the achievement of unit standards related to the qualification

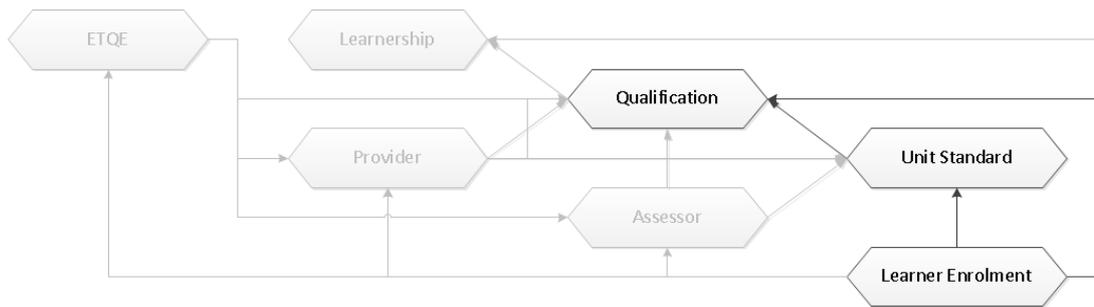


Figure 3.6.3.6.a.1 Conceptual diagram of the NQF concepts that participate in business rule 6.a

The linkage between a qualification achievement record and a unit standard achievement record is determined based on the Qualification ID value on the unit standard achievement record which is submitted to the NLRD by the ETQE. Qualification records, unit standard records and the required relationship between a qualification and a unit standard are maintained on the NLRD by SAQA.

Achievement of the required minimum number of credits for a qualification (recorded in the Qualification table as Minimum Credits) is derived from the tables Qualification Achievement, Qualification, Qualification Unit Standard Link, Unit Standard Achievement and Unit Standard, utilizing the Total Credits in Unit Standard to calculate the actual credits achieved.

There may be instances where the total number of credits for a qualification calculated based on the number of credits available as derived from the unit standards linked to the qualification, is not greater than or equal to the minimum number of credits required for the qualification. This would be considered a data capturing error and will be reported to SAQA for further action or interpretation. The analysis of this type of issue falls outside of the scope of this research.

There may also be instances where:

1. No matching unit standard enrolment records can be found for the qualification achievement record

2. Matching unit standard enrolment records can be found for the qualification achievement record; however the unit standard enrolment records do not have the Qualification ID recorded against them
3. Matching unit standard enrolment records can be found for the qualification achievement record; however the unit standard enrolment records have the incorrect Qualification ID recorded against them

Although these could also be considered data capturing errors, these types of issues do fall within the scope of this research.

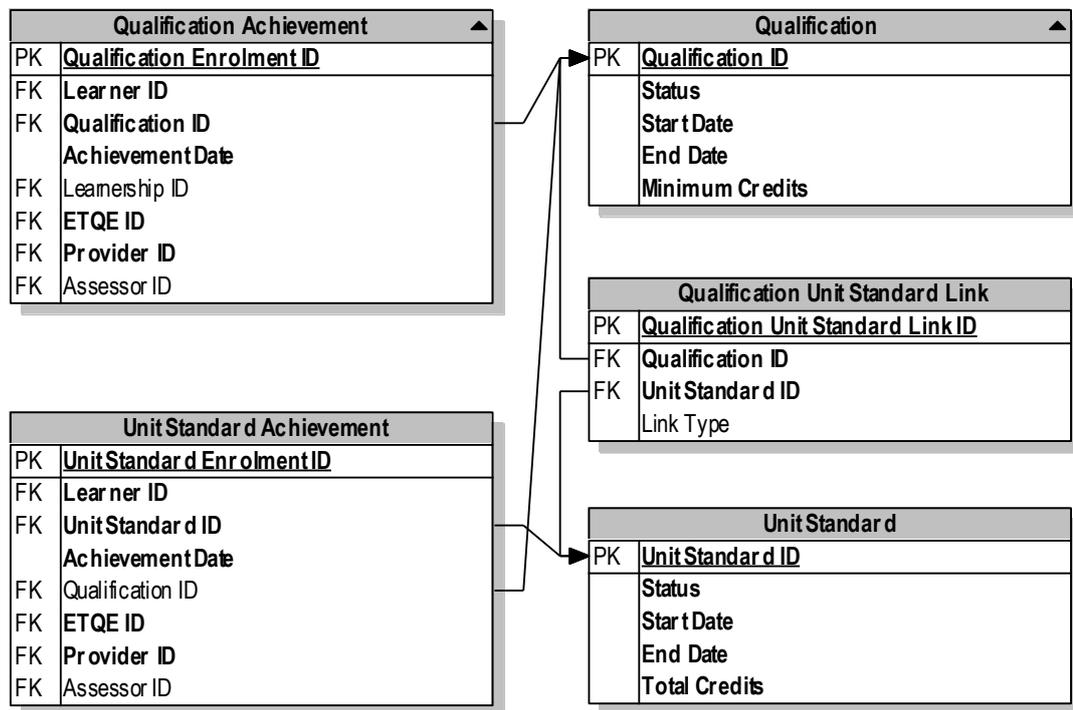


Figure 3.6.3.6.a.2 Conceptual diagram of the tables and fields that inform business rule 6.a

- b. the learner would have achieved the correct range of credits for the qualification, on or before the achievement of the qualification, based on the achievement of unit standards that have been defined as core, fundamental and elective unit standards for the qualification

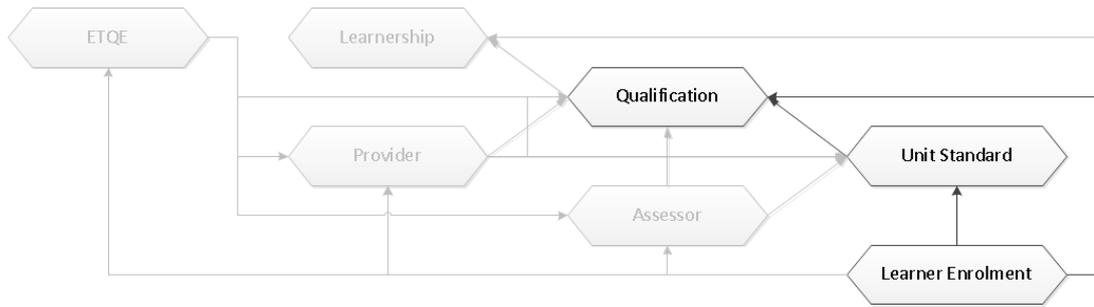


Figure 3.6.3.6.b.1 Conceptual diagram of the NQF concepts that participate in business rule 6.b

As indicated previously, the linkage between a qualification achievement record and a unit standard achievement record is determined based on the Qualification ID value on the unit standard achievement record which is submitted to the NLRD by the ETQE. Qualification records, unit standard records and the required relationship between a qualification and a unit standard are maintained on the NLRD by SAQA.

Achievement of the correct range of credits required for a qualification is derived from the tables Qualification Achievement, Qualification, Qualification Unit Standard Link, Unit Standard Achievement and Unit Standard, utilizing the Link Type in Qualification Unit Standard Link.

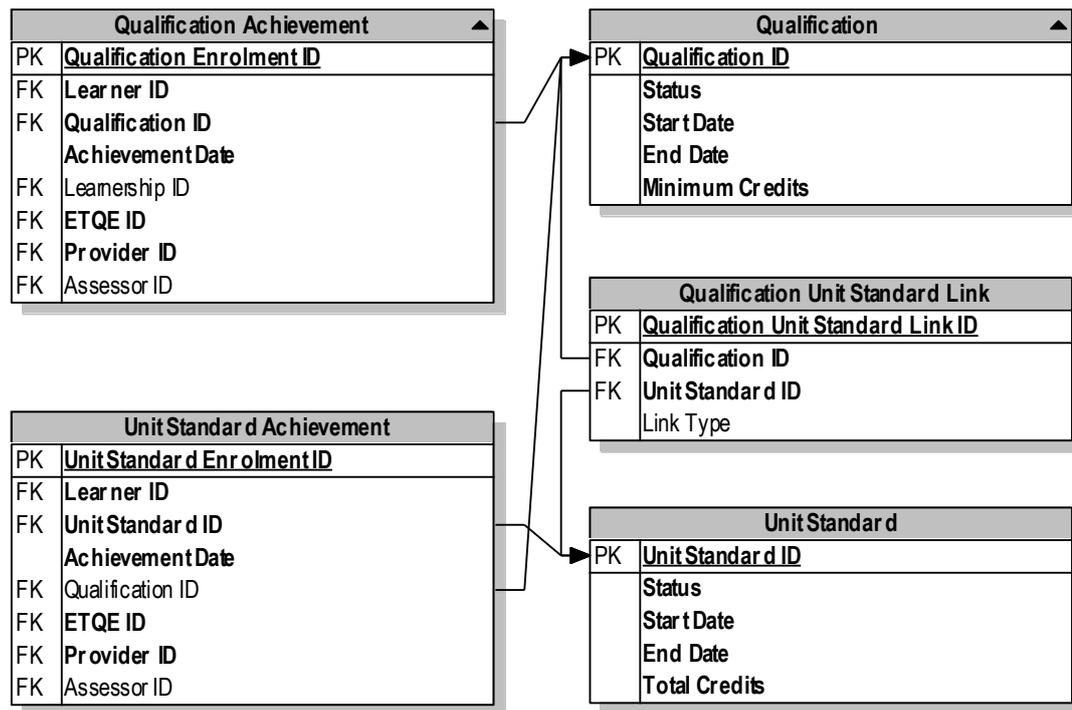


Figure 3.6.3.6.b.2 Conceptual diagram of the tables and fields that inform business rule 6.b

This section provided a review of the data structures that store data related to the semantic business rules that are core to this research. The comprehension gained by completing such a review assisted in the development of an understanding of the nature and scope of the data that is required from the NLRD for this research.

### 3.6.4 Considerations that impact the analysis of the data

Having gained an understanding of how the data which inform the semantic business rules are stored in the NLRD, four (4) idiosyncrasies as to how this data would need to be analysed have been recognised, namely; the active enrolment time period of the learner’s enrolment, active accreditation and registration time periods of ETQEs, providers and assessors, the representation of a learner enrolment record’s compliance to a semantic business rule and the representation of temporal data. This section details how these idiosyncrasies will need to be accommodated for in the analysis of the data in this research.

#### 1. The learner’s active enrolment time period

The NLRD currently only collects the date of enrolment and the date of completion for the learnership or achievement for the qualification/unit standard enrolment record. The compulsory provision of an enrolment date only came into effect at the beginning of

2008 and the provision of a date of completion is only compulsory when the learner has completed the learnership and the provision of a date of achievement is only compulsory when the learner has achieved the qualification/unit standard. Further, there is no additional data value that indicates that the learner is actively participating on the learnership/qualification/unit standard between the date of enrolment and the date of completion/achievement.

The following mechanisms will need to be implemented to derive an active enrolment period when analysing the data:

- a. For records where an enrolment date has not been supplied a derived “start date” will need to be implemented. For records that have not been completed/achieved the “start date” will need to be implemented based on the date stamp (this value denotes the last date on which the record was updated on the ETQE information system and is a compulsory data value) of the first record submitted to the NLRD for the specific enrolment (the first instance of a record may be found in either the data table or audit table for learner enrolments).

For records that have been completed/achieved and as a result have an completion/achievement date, the “start date” will need to be implemented based on the expected duration of the learnership/qualification/unit standard based on the credit for the learnership/qualification/unit standard (the reader should note that learnerships are not allocated a credit value; they are however intrinsically linked to a qualification and as a result can assume the credit value of the linked qualification).

To ensure that on analysis of the data, derived enrolment dates and actual enrolment dates are not confused, both actual enrolment dates and derived enrolment dates should be saved in a data field with a name other than enrolment date (for example the data field could be called start date) and a dichotomous variable must be implemented that discerns the one from the other.

- b. For records where the learner has not completed the learnership or achieved the qualification/unit standard, and as a result has no completion/achievement date available, a derived “completion/achievement date” will need to be implemented based on the expected duration of the learnership/qualification/unit standard based on the credit for the learnership/qualification/unit standard.

To ensure that on analysis of the data, derived completion/achievement dates and actual completion/achievement dates are not confused, both actual completion/achievement dates and derived completion/achievements dates should be saved in a data field with a name other than completion/achievement date (for example the data field could be called end date). A dichotomous variable would not be required in this instance because the data structure of the enrolment record already contains a nominal variable that describes the status of the enrolment record.

Using the derived start date and end date will allow for the implementation of an active enrolment period for each record regardless of whether the record has data values for the enrolment date and/or completion/achievement date. The implementation of the dichotomous variable that further describes the nature of the start date and the already implemented nominal variable that describes the status of the enrolment will allow for the accurate selection of data where the analysis requires that only records with actual enrolment dates or completion/achievement dates must be analysed.

## 2. Active accreditation and registration time periods

As stated in the Section 3.6.3, the NLRD does not store the history of ETQE and provider accreditations or qualification, unit standard and assessor registrations in designated history tables. The history of these types of records is deduced from the audit tables for the respective table. As a result, the history of active accreditations or registrations may contain gaps in the data.

As an example the following active accreditations/registrations could be found in the data and “history” table (see Figure 3.6.4.2.1):

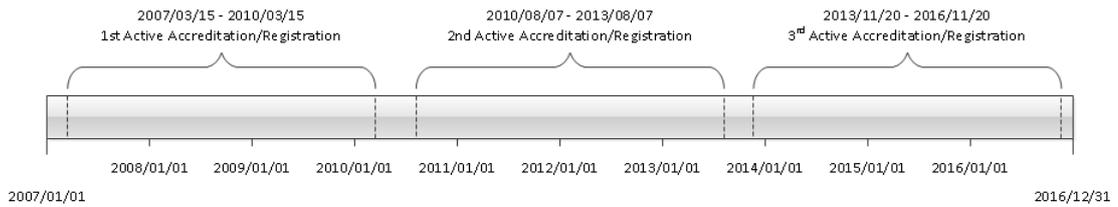


Figure 3.6.4.2.1 Figure depicting active accreditations/registrations with gaps

The gaps in the data are generally considered to be as a result of administrative type issues such as a delay in the request for re-accreditation/re-registration. Including these gaps as times in which the ETQE and provider or qualification, unit standard and assessor do not have an active accreditation or registration would be meaningless and would skew the results of the research. As a result when deducing the active accreditation or registration time period of the ETQE and provider or qualification, unit standard and assessor, only the start date of the first accreditation/registration and end date of the last accreditation/registration must be considered as depicted in Figure 3.6.4.2.2:

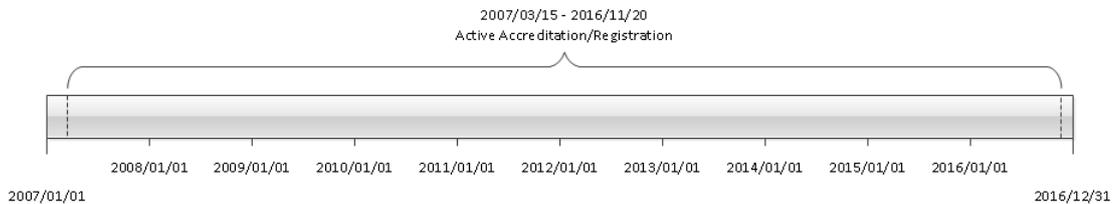


Figure 3.6.4.2.2 Figure depicting active accreditation/registration with gaps removed

### 3. Representation of compliance to a rule

Although the semantic business rules seem to require a simple yes or no answer in regard to whether the learner enrolment record is compliant to a rule, some consideration must be given to the nature of the information that the data stored in the NLRD represents.

As an example, the ETQE ID is a compulsory field on the learner enrolment record. As a result, determining whether the ETQE was accredited for the duration of the learner's active enrolment on the learnership/qualification/unit standard could be expressed as dichotomous values "Yes" and "No". However neither the ETQE's accreditation nor the learner's active enrolment constitute single points in time, rather they both have a time span. As a result an ETQE's accreditation time span could cover the entirety of the

learner's active enrolment time span, or fall before or after the learner's active enrolment time span or cover only the beginning or the end of the learner's active enrolment time span.

The level of detail described in the example above can be expected in the results of each semantic business rule and is important for the analysis of the data. As a result nominal categories are required to describe the results of a test against a specific semantic business rule.

#### 4. Representation of temporal data

There are two temporal data aspects that need to be considered in regard to the semantic business rules.

The first is the representation of data values such as the derived start and end dates for each learner enrolment record. The representation of this data must contain sufficient levels of aggregation in order to have the correct level of meaning for the data mining algorithm and the correct level of detail in order to bring meaning to the analysis of the data.

A straightforward analysis of the data based on "start date" and "end date" combined with nominal values that describe compliance to a rule will bring some insight into the data being analysed. This approach however ignores the salient consideration noted at the end of Section 3.6.2 that make reference to the NQF and all its supporting structures and policies having been new structures, and that a common understanding of what these policies and structures entail would only have been gained over time, and that the NLRD stores both current and legacy learner enrolment and achievement records.

The second representation of temporal data therefore needs to address the representation of data in relation to a point in time. Examples are the initial implementation of the NQF, the start or end of the active accreditation of an ETQE or provider, the start or end of the active registration of a learnership/qualification/unit standard or assessor.

This section detailed four (4) idiosyncrasies in how the data from the NLRD would need to be analysed for this research. These idiosyncrasies included the active enrolment time

period of the learner's enrolment, active accreditation and registration time periods of ETQEs, providers and assessors, the representation of a learner enrolment's compliance to a semantic business rule and the representation of temporal data. The understanding gained of these idiosyncrasies will guide the manner in which the data that is sourced from the NLRD will be derived in order to ensure that the data can be properly analysed in relation to the semantic business rules.

### ***3.6.5 Selection of additional data for the analysis***

Section 3.6.3 and Section 3.6.4 provide an indication as to the minimum data that would need to be extracted from the NLRD in order to conduct this research. This section explores whether there are any additional data fields that should be obtained from the NLRD that may prove valuable to the analysis of learner enrolment records in regard to their compliance to the semantic business rules of this research.

Most data mining software applications have built-in functionality that allows for the automatic selection of relevant data for the data mining activity. This study however has a very specific focus in regard to which data in the NLRD will be mined, and as a result some parameters for the selection of additional data that is included into the analysis must be defined.

Thus far a limited number of data fields have been identified which will provide data with which to test the compliance of an enrolment record against the semantic business rules. The relevance of these data fields, or the information derived from these data fields, are deemed an essential aspect to the data to be mined. However, in order to produce an analysis of the data that is meaningful to SAQA further data fields will be incorporated into the data to be analysed. The selection of such data fields must exclude the selection of any biographical data, but may include data that further describes the learner enrolment record, an ETQE, provider, assessor, learnership, qualification and unit standard.

A review of the relevant data structures indicates that:

1. The data field achievement type would further describe learner enrolment records.
2. The only additional data field that could bring value to the research in regard to ETQE related data is the data field organisation type. The scope of ETQEs that submit learner

enrolment data to the NLRD is however limited to ETQAs, QCs and QPs and its inclusion into the data may result in a redundant correlation.

3. Additional nominal data variables that further describe a provider include the ETQE of the provider (this may be different from the ETQE that submitted the enrolment record to the NLRD), the type of provider, the class of the provider and the province that the provider is situated in.
4. No additional data fields, that are non-biographical, can be found that can further describe an assessor.
5. The data field NQF level would further describe learnerships.
6. The data fields NQF level, qualification type, qualification class, field and learning subfield would further describe qualifications.
7. The data fields NQF level, unit standard type, field and learning subfield would further describe unit standards.

All of the additional data fields listed above are data fields that hold nominal data variable values.

This section determines that there are 15 additional data fields that may prove valuable to the analysis of the learner enrolment records in regard to their compliance to the semantic business rules of this research.

### ***3.6.6 Conclusion***

This section is instrumental in: determining the semantic business rules for this research; gaining an understanding of the physical data structures in the NLRD that store data related to the semantic business rules; appreciating the manner in which some of the data would need to be prepared for analysis and determining any additional data fields that might prove useful for this research.

The numerous reviews and their resultant considerations allows for the establishment of a fundamental understanding of the type and nature of the data that would be mined for this research. This provides a clear framework that guides the selection of data from the NLRD. Further, the insight gained from this section guided the literature review in regard to the type and nature of the data mining that is conducted for this research.

### 3.7 Data collection

This research is conducted on the pre-existing data stored in the NLRD therefore no new data was collected for the purposes of this research.

### 3.8 Data analysis procedure

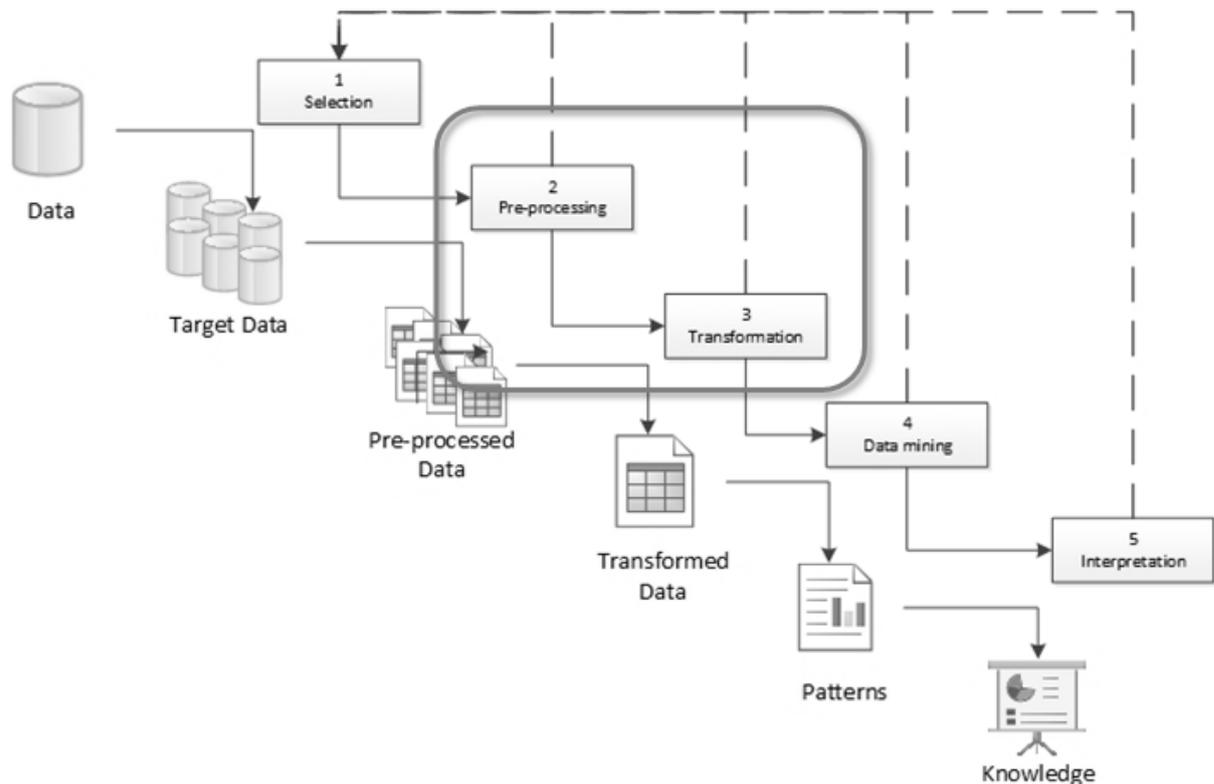


Figure 3.8.1 KDD phases – Pre-processing and Transformation

#### 3.8.1 Obtaining data from the NLRD

As indicated in Section 3.6.1, the NLRD database comprises of two discrete functionalities;

- an operational information system aspect that allows SAQA to manage and maintain data related to the NQF, and
- a data warehouse aspect which is populated with data, describing learner enrolment and related records.

Data from the NLRD is published for the general public. For example the SAQA website publishes a searchable database of all the qualifications and part qualifications at <http://www.saqa.org.za/show.php?id=5677>. The descriptive data in regard to the

qualifications and part qualifications are sourced from the operational information system aspect of the NLRD. When searching for a specific qualification or part qualification the user is given the option to search for an accredited provider or to view the providers accredited to offer a specific qualification or part qualification that has been selected. These provider accreditation details are sourced from the data warehouse aspect of the NLRD.

The NLRD, which is considered a national resource, is currently the most comprehensive repository of data that describes learner enrolment and related records in the Republic of South Africa. In all instances where data is required for the general public, from either aspect of the NLRD, SAQA enforces a strict data push policy that ensures that data is pushed to the recipient system. Further, no system or application available to the general public has direct access to the NLRD. The technical implementation of the push policy prevents any data being pulled from the NLRD in an unauthorized manner. Additionally, due to the detailed and personal nature of the records stored in the NLRD data warehouse, SAQA does not publish any data related to learner enrolments in a detailed format. Data related to learner enrolments is always published in an aggregated format as publications.

This research requires, by its very nature, unit record data. However, as illustrated in Section 3.6.3 this research does not require any data that characterises or details personal information stored in the NLRD. The results of the research could however potentially reveal sensitive information in regard to the origin of the data: such as the integrity of the information system and/or administrative processes deployed at its source.

Obtaining data for this research was impacted by the considerable and understandable security concerns described above. As a result obtaining data from the NLRD for this research was achieved in the following manner:

1. Specific permission was obtained from SAQA for the data and the utilization of the data for this research.
2. The data for this research was obtained as data extracts from the NLRD.
3. All data that could possibly be used to identify people or organisations was de-identified. The de-identification preserved identifying information that would allow SAQA to relink data if required.

### **3.8.2 Raw data obtained from the NLRD**

The initial review of the physical NLRD data structures in relation to the type of data required, documented in Sections 3.6.3, 3.6.4 and 3.6.5, revealed that the research would require sourcing data from 13 data tables, 12 audit-related tables and 14 lookup tables.

Although only a limited number of data fields were required from each table, the development of “active” records as described in Section 3.6.4.1 and Section 3.6.4.2 requires an almost complete extract of all of the data records stored in the data tables and their audit tables. The development of “active” records is further complicated in that the NLRD contains numerous statuses that denote an “active” status for most record types. As a result, sourcing the data records in raw format from the NLRD requires a considerable understanding of the specific statuses that need to be included in the development of “active” records.

In consultation with the Director of the NLRD, it was determined that obtaining data from the NLRD in such a granular format would bring no benefit to the research. As a result the request for data to be extracted from the NLRD included the derivation of data by the NLRD database administrator (DBA) as follows:

- A single active record was provided, in the manner as described in Section 3.6.4.1.a for all learner enrolment records.
- A single active record was provided in the manner as described in 3.6.4.2 for all records that describe accreditations and registrations. Only records that had been derived in this manner with both a start date and an end date, where the start date was not equal to the end date, were provided.
- The description for all lookup values requested was included in the data tables provided, thereby eliminating the need to extract the lookup tables as separate data tables. In any instances where the lookup value was not defined as required on the original data table, NULL values were recoded to a lookup code of 0 with a description of “Undefined”.

By eliminating the need for a full audit trail, both the number of tables and the volume of data provided is greatly reduced. The overall number of tables required is further reduced by including the description of any lookup values in the data tables. As a result the final request for data from the NLRD is reduced to 18 tables, a technical description of these tables has been provided in A.1. A non-technical description of the 18 tables follows:

1. DM\_ASOR  
This table stores records that describe assessor registrations.
2. DM\_ASOR\_REGSTR  
This table stores details in regard to the assessor's registrations to assess learnerships, qualifications and/or unit standards.
3. DM\_ETQE  
This table stored records that describe ETQE accreditations.
4. DM\_ETQE\_ACCRED  
This table stores details in regard to ETQE accreditations to quality assure qualifications and unit standards.
5. DM\_ETQE\_START  
This table stores records that describe the first date on which an ETQE submitted a full data submission to the NLRD. Further information in regard to the requirement for this table can be found in Section 1.
6. DM\_LSHP  
This table stores records that describe learnerships.
7. DM\_LSHP\_ENROL  
This table stores details in regard to learnership enrolments.
8. DM\_LSHP\_ETQE  
This table is not addressed directly in the analysis of the NLRD data structures in Section 3.6.3. This table stores a record of each ETQE that was mandated to implement a specific learnership. Further information in regard to the requirement for this table can be found in Section 3.8.3.3.
9. DM\_PROV  
This table stores records that describe provider accreditations.
10. DM\_PROV\_ACCRED  
This table stores details in regard to provider accreditations to offer learnerships, qualifications and unit standards.
11. DM\_QUAL  
This table stores records that describe qualifications. This table contains two additional fields that are not addressed directly in the analysis of the NLRD data structures in Section 3.6.3:

- The transition period for the qualification which is an extension of the end date of the registration of a qualification because new learners may still be enrolled on the qualification during this time period (South African Qualifications Authority, 2007).
- The train-out period for the qualification which is an extension of the end date of the registration of a qualification (South African Qualifications Authority, 2007, p. 1).  
The data value in both of these fields has bearing on the analysis of qualification enrolments when determining whether the qualification was registered for the duration of the learner's active enrolment on the qualification.

#### 12. DM\_QUAL\_ENROL

This table stores details in regard to qualification enrolments.

#### 13. DM\_QUAL\_LSHP

This table stores data that describes the relationship between qualifications and learnerships.

#### 14. DM\_QUAL\_REPL

This table was not addressed directly in the analysis of the NLRD data structures in Section 3.6.3. At the end of a qualification's normal lifespan (generally 3 years) the qualification is reviewed and depending on the results of the review process will either be "*...re-registered, significantly changed or replaced by a newly developed qualification.*" (South African Qualifications Authority, 2007, p. 1). This table stores data related to the replacement of one qualification with another qualification and has bearing on the analysis of learnership enrolments in relation to qualification enrolments.

#### 15. DM\_USTD

This table stores records that describe unit standards. As with the table DM\_QUAL, this table also contains the two additional fields, transition period and train-out period, which are not addressed directly in the analysis of the NLRD data structures in Section 3.6.3.

#### 16. DM\_USTD\_ENROL

This table stores details in regard to unit standard enrolments.

#### 17. DM\_USTD\_QUAL

This table stores data that describes the relationship between unit standards and qualifications.

#### 18. DM\_USTD\_REPL

As with the table DM\_QUAL\_REPL, this table is not addressed directly in the analysis of the NLRD data structures in Section 3.6.3. This table stores data related to the

replacement of one unit standard with another unit standard and has bearing on the analysis of qualification enrolments in relation to unit standard enrolments.

### **3.8.3 *Overarching data derivation considerations***

This section provides additional information in regard to specific contextual aspects of the data received from the NLRD that needs to be accommodated during the derivation of the data in preparation of the data mining activity.

#### **1. Missing histories**

On review of the data in the NLRD it was found that in all instances a considerable amount of time elapsed between the time at which an ETQE was established and when the ETQE submitted its first full data submission to the NLRD. During this time period providers/assessors may have been accredited/registered and then re-accredited and re-registered. The NLRD only receives the most recent version of the accreditation/registration record for the provider/assessor and as a result the history of initial accreditation or registration would not have been submitted to the NLRD, thereby resulting in a missing history for the provider/assessor accreditation/registration.

The following is an illustration of the issue:

An ETQE is established in the beginning of 2000 and accredits its first provider, Provider A, to offer qualifications from 2001/01/01 to 2001/12/31. During the time period 2001/01/01 to 2001/12/31 Provider A, enrolls 300 learners on qualifications, of which 150 learners achieve their qualification before the end of Provider A's accreditation. Provider A is re-accredited by the ETQE from 2002/01/01 to 2002/12/31. During the time period 2002/01/01 to 2002/12/31, 100 of the remaining learners achieve their qualification. The ETQE only submits its first full data submission to the NLRD at the end of 2002. As a result, the submission only contains a record describing the accreditation of Provider A from 2002/01/01 to 2002/12/31.

An analysis of the qualification enrolment records in the above example, based on the data that exists in the NLRD data tables and audit trails, shows that:

- 150 learners enrolled on and achieved their qualification before Provider A was accredited,

- 100 learners enrolled on their qualification before Provider A was accredited and achieved their qualification whilst Provider A was accredited, and
- 50 learners enrolled on their qualification before Provider A was accredited and have yet to achieve the qualification.

The above type of scenario results in a false negative for the semantic business rule “... the provider was accredited for the duration of the learner’s active enrolment on the learnership/qualification/unit standard” (see Section 3.6.2.2.a).

In order to eliminate any false negatives generated as a result of this type of scenario, it was decided that nominal indicators that describe a record’s compliance to a specific semantic business rule must differentiate between enrolment records with a start date that precedes the date on which an ETQE submitted its first full data submission. Consequently the request for data from the NLRD was amended to include the table DM\_ETQE\_START as described in Section 3.8.2.5. The START\_DATE value in this table was utilized to determine whether the start date of an enrolment record preceded the first full data submission from the ETQE, to the NLRD. Nominal variables describing the records compliance to a specific semantic business rule are amended accordingly.

## 2. Unpredictable futures

The end of the active enrolment time period for learnership enrolment records that have not been completed, and qualification/unit standard enrolment records that had not been achieved needs to be derived for the purposes of this research (Section 3.6.4.1). These types of records can be categorized as

- legacy (the derived end date for the active enrolment is in the past), and
- current (the derived end date for the active enrolment is in the future) enrolment records.

The conditions around a current enrolment record may change at some point in the future. As a result, a record that was found to have failed compliance against a specific semantic business rule at the time of this research may in fact be found to be compliant against the same semantic business rule at a later date.

The following example illustrates why current enrolment records are problematic.

- The enrolment: A learner is enrolled on a qualification with a start date of one year ago, on a qualification that takes three years to complete.
- The provider accreditation: The provider at which the learner is completing his/her studies was accredited four and a half years ago, and the provider's accreditation will end in six months' time. The provider is in the process of applying for re-accreditation.

An analysis of the qualification enrolment record in the above example, based on the data that exists in the NLRD, would show that the active enrolment time period is longer than the accreditation time period of the provider.

The above type of scenario would result in a false negative for the semantic business rule "... the provider was accredited for the duration of the learner's active enrolment on the learnership/qualification/unit standard" (see Section 3.6.2.2.a).

In order to eliminate any false negatives generated as a result of this type of scenario it was decided that nominal indicators that describe a record's compliance to a specific semantic business rule must differentiate between enrolment records that have their start and end dates at some time in the past and enrolment records have their start date at some time in the past and their end date at some time in the future.

Bearing in mind that the ETQEs only submit data to the NLRD twice a year (South African Qualifications Authority, 2011, p. 1) and that these submissions are bound by specific dates on which the data submission must reach the NLRD, the data derivation logic was amended to include a variable that could be set to the most recent data submission cycle date. This date was then utilized to determine whether the end date of an enrolment record is subsequent to the most recent data submission cycle.

### 3. Incorporating ETQE amalgamations

Over the course of the life cycle of the NLRD a number of ETQEs have been established at different points in time. When new ETQEs are created they may take over the jurisdiction of one or more existing ETQEs and as a result the accreditation to quality assure specific qualifications and unit standards. The analysis of learner enrolment records must incorporate accreditations of amalgamated ETQEs in order to ensure that the

research does not produce false positives when evaluating semantic business rules related to the accreditation of the ETQE.

ETQE amalgamations can prove problematic if not considered during the analysis of the data as illustrated below:

ETQE A is accredited by SAQA both as an ETQE and to quality assure qualifications X and Y from 2000/01/01 to 2004/12/31. In the time period 2001/01/01 to 2001/12/31 ETQE A submits 3000 learner enrolment records linked to qualification X to the NLRD.

ETQE A is amalgamated into existing ETQE B and new ETQE C on 2002/01/01. ETQE C is accredited by SAQA both as an ETQE and to quality assure qualification X from 2002/01/01 to 2005/12/31. In the time period 2003/01/01 to 2003/12/31 ETQE C submits the same 3000 learners as mentioned above as having achieved qualification X to the NLRD.

An analysis of the qualification enrolment records in the above example, based on the data that exists in the NLRD, would show that ETQE C was not accredited as an ETQE or accredited to quality assure qualification X when the 3000 learners enrolled on qualification X.

The above type of scenario would result in a false negative for the semantic business rules "... the ETQE that submitted the record was accredited for the duration of the learner's active enrolment on the learnership/qualification/unit standard" (see Section 3.6.2.1.a) and "... the ETQE that submitted the record was accredited to quality assure the qualification/unit standard for the duration of the learner's active enrolment on the learnership/qualification/unit standard" (see Section 3.6.2.1.b).

In order to eliminate any false negatives generated as a result of this type of scenario it was decided that nominal indicators that describe a record's compliance to an ETQE related semantic business rule must consider the amalgamation of ETQEs and differentiate between records where:

- only the submitting ETQE's active accreditation time period was considered, and
- where the amalgamated ETQE's active accreditation time period was considered.

For learnership enrolment records this was done by recording the accreditation details of the ETQE that submitted the learnership enrolment record and the accreditation details of any other ETQE that had/has been accredited to quality assure the learnership against the enrolment record. Using this data the logic developed could:

- test compliance of a business rule against the details of the ETQE that submitted the record to the NLRD and record the results, and
- if the record failed compliance, retest the record using the accreditation details of the amalgamated ETQE and record the results.

The same type of logic is used to assume an active accreditation time period of an amalgamated ETQE for qualification and unit standard related enrolment records.

#### 4. Incorporating the evolution of qualifications and unit standards

The normal lifespan of qualifications and unit standards is three years. After three years a qualification or unit standard is reviewed and dependent on the outcome of the review process the qualification or unit standard is either re-registered, changed or replaced respectively by a new qualification or unit standard (South African Qualifications Authority, 2007, p. 1).

In instances where qualifications that are replaced are linked to learnerships, the linkage between the learnership and the replaced qualification is updated with a linkage between the learnership and the qualification that replaces the initial qualification. As a result, the historical linkage between the replaced qualification and the learnership is no longer immediately evident. The same applies in instances where unit standards that are replaced are linked to qualifications.

In both instances the research must take into consideration the historical linkages of both learnerships and qualifications whilst testing the compliance of any semantic business rules that relate to:

- the relationship between learnership enrolments and their respective qualification enrolments, and
- qualification enrolments and their respective unit standard enrolments.

The following example illustrates why qualification/unit standard replacements must be taken into consideration during the analysis of the data:

When learnership A was initially designed in 2000 it was linked to qualification X. Qualification X was registered from 2000/01/01 to 2002/12/31. As qualification X neared the end of its registration time period it was reviewed, and a decision was made to replace qualification X with qualification Y. Qualification X expired on 2002/12/31 and qualification Y was registered from 2003/01/01 to 2005/12/31.

ETQE K has been given the mandate to implement learnership A and submits 100 completed learnership enrolments against learnership A, with 100 corresponding achieved qualification X enrolments in 2002. In 2003 ETQE K submits a further 100 completed learnership enrolments against learnership A, with 100 corresponding achieved qualification Y enrolments.

An analysis of learnership A completions in relation to the corresponding qualification X achievements in the above example would show that the learners that completed learnership A do not have qualification enrolment records for the qualification that is linked to learnership A.

The above type of scenario would result in a false negative for the semantic business rule "... if the learner has completed the learnership, then due to the intrinsic nature of a learnership and qualification the learner would have achieved the qualification on or before the completion of the learnership" (see Section 3.6.2.5).

In order to eliminate any false negatives generated as a result of this type of scenario it was decided that any rules that seek to establish a relationship between a:

- Learnership enrolment record and a qualification enrolment record (see Section 3.6.2.5) would utilize the qualification replacement table (DM\_QUAL\_REPL) to ensure that the matching of records includes replacement qualification enrolment records.
- Qualification enrolment record and unit standard enrolment records (see Section 3.6.2.6) would utilize the unit standard replacement table (DM\_USTD\_REPL) to

ensure that the matching of records includes replacement unit standard enrolment records.

#### 5. Primary ETQE of a provider and the ‘ETQE provider’

In certain situations, ETQEs are forced to create a placeholder provider record (known as an ‘ETQE provider’ from this point forward) in order to submit data records to the NLRD. An ‘ETQE provider’ may not be submitted to the NLRD as accredited and may not be accredited to offer qualifications/unit standards. The analysis of learner enrolment records must ensure that the existence of an ‘ETQE provider’ does not produce false positives when evaluating semantic business rules related to the accreditation of the provider.

Although a provider may be accredited to offer qualifications by more than one ETQE, a provider may only have one primary ETQE. In the NLRD specifications, SAQA defines that a record that describes the provider may only be submitted to the NLRD by the primary ETQE of the provider (South African Qualifications Authority, 2013, p. 5).

- The record that describes a provider includes an identifier for the provider, called a provider code, which is unique to the ETQE, and unique on a national level when combined with the ETQE’s identifier.
- All providers are able to identify a primary ETQE either by the level of the qualifications that they are accredited to offer (for example a public university has the CHE as its primary ETQE) or the sector that they predominantly offer training in (for example an accounting software training provider has the Finance and Accounting Services Sector Education and Training Authority (FASSET) as its primary ETQE).

When an ETQE makes reference to a provider in its data submissions to the NLRD it must use the provider code (as issued by the primary ETQE of the provider) in combination with the provider’s primary ETQE’s identifier.

- The NLRD data warehouse is relational and as a result the provider code issued by primary ETQE can only be used in a data submission of a non-primary ETQE once the primary ETQE of the provider has successfully submitted the provider record to the NLRD.

- In the NLRD specifications, SAQA further defines that a valid combination of ETQE identifier and provider identifier must be provided for each learner enrolment record (South African Qualifications Authorityd, 2013, pp. 16, 17 and 18).

There are a number of situations, of an administrative nature, that exasperate an ETQE’s efforts to conform to both of these above mentioned requirements when submitting data to the NLRD. In order to ensure that enrolment records are not prevented from reaching the NLRD in these types of situations, SAQA allows each ETQE to submit one provider record that records the ETQE as a provider i.e. an ‘ETQE provider’.

An ETQE provider record must have an accreditation status of “unknown” and may not be linked to accreditation records to offer qualifications and/unit standards. As a result any enrolment record that has an ‘ETQE provider’ would trigger a false negative for the semantic business rules:

- “... the provider was accredited for the duration of the learner’s active enrolment on the learnership/qualification/unit standard” (see Section 3.6.2.2.a), and
- “... the provider was accredited to offer the qualification/unit standard for the duration of the learner’s active enrolment on the learnership/qualification/unit standard” (see Section 3.6.2.2.b).

In order to eliminate any false negatives generated as a result of ‘ETQE providers’ the results for all semantic business rules that are related to the accreditation of a provider must have a separate category that denotes an ‘ETQE provider’.

## 6. Rounding date differences

The evaluation of a number of the semantic business rules requires a comparison between two dates. A suitable level of detail in regard to the difference between two dates needs to be selected in order to make sure that the results of date comparisons are meaningful. It was decided, in consultation with the Director of the NLRD, that the difference between two dates would be expressed as a whole number representing the difference in months between the two dates being compared.

For example the semantic business rule “...the ETQE that submitted the record is accredited for the duration of the learner’s active enrolment on the

learnership/qualification/unit standard” (see Section 3.6.2.1.a) required a comparison between the start date of the ETQE and the start date of the enrolment record and a comparison between the end date of the ETQE and the end date of the enrolment record.

As an example, the difference between '2003/01/01' and '2003/03/14' would return a value of 2.41935483870968. It was decided that the comparison between two dates should be expressed as a whole number of months only. Further, the calculation must be lenient and must round down if the result was a positive value and round up if the result was a negative value. To illustrate:

- The difference between a start date '2003/07/01' and a start date '2003/03/14' generates a result of -3.58064516129032 which when rounded up is -3.
- The difference between a start '2003/03/14' and a start date '2003/07/01' generates a result of 3.58064516129032 which when rounded down is 3.

## 7. Learnership registration problems

The registration of learnerships was managed by the Department of Labour from 1998 to 2010, thereafter this role was taken over by the Department of Higher Education and Training (DHET) (Ministry in the Office of the President, Skills Development Act, Act 97 of 1998, p. 11).

The overall development of learnerships is the responsibility of SETAs (Ministry in the Office of the President, Skills Development Act, Act 97 of 1998, p. 11). The envisaged development of learnerships by a SETA as defined by the DoL (Department of Labour, 2013, p. 3) included the participation of a standards generating body and working on the development of unit standards and the qualification for the learnership. Once completed the learnership was submitted to DoL for registration.

The initial analysis of qualification enrolment records immediately highlighted a problem in the data in regard to the amount of records that infringed on the semantic business rules. Further review of the problem shows that qualification enrolment records that are linked to a learnership (i.e. the LEARNERSHIP\_ID value in the table DM\_QUAL\_ENROL was not NULL) were extremely likely to infringe the semantic business rules. These types of records show that a number of learnership based qualification enrolment records were either linked to qualifications outside of the registration time period of the qualification.

The matter was discussed with the Director of the NLRD and it was determined that the manner in which learnerships were established and managed had contributed to this problem. Unfortunately, the process of the development of learnerships by SETAs and the registration of learnerships by DoL had not taken into consideration the registration time period of the qualification that was/is linked to the learnership. This ultimately resulted in the publication of learnerships against qualifications in a manner that would result in the infringement of the semantic business rules.

It was decided, in consultation with the Director of the NLRD, that the further analysis of these types of records will not form part of this research. The percentage of records for this specific issue is however illustrated for all of the semantic business rules in order to provide an understanding of how this specific issue has impacted the data in the NLRD.

#### ***3.8.4 Learnership enrolment data selection, pre-processing and derivation***

The initial selection, pre-processing and derivation of the learnership enrolment records, received from the NLRD in the table DM\_LSHP\_ENROL, into a format that is suitable for data mining is described in Appendix C.

The specific semantic business rules that are applicable to learnership enrolment records are identified in Appendix C.2. The analysis and data mining of these semantic business rules requires the implementation of four (4) semantic business rule indicators. Appendix C.3 describes the selection, pre-processing and derivation steps required for the implementation of these semantic business rule indicators.

The selection, pre-processing and derivation logic resulted in the implementation of a final version of the learnership enrolment data as a new table called DM\_LSHP\_ENROL\_FINAL. A technical description of the table DM\_LSHP\_ENROL\_FINAL is provided in C.1.

#### ***3.8.5 Qualification enrolment data selection, pre-processing and derivation***

The initial selection, pre-processing and derivation of the qualification enrolment records, received from the NLRD in the table DM\_QUAL\_ENROL, into a format that is suitable for data mining is described in Appendix E.

The specific semantic business rules that are applicable to qualification enrolment records are identified in Appendix E.2. The analysis and data mining of these semantic business rules requires the implementation of eight (8) semantic business rule indicators. Appendix E.3 describes the selection, pre-processing and derivation steps required for the implementation of these semantic business rule indicators.

The selection, pre-processing and derivation logic resulted in the implementation of a final version of the qualification enrolment data as a new table called DM\_QUAL\_ENROL\_FINAL. A technical description of the table DM\_QUAL\_ENROL\_FINAL is provided in E.1.

### ***3.8.6 Unit Standard enrolment data selection, pre-processing and derivation***

The initial selection, pre-processing and derivation of the unit standard enrolment records, received from the NLRD in the table DM\_USTD\_ENROL, into a format that is suitable for data mining is described in Appendix G.

The specific semantic business rules that are applicable to qualification enrolment records are identified in Appendix G.2. The analysis and data mining of these semantic business rules requires the implementation of seven (7) semantic business rule indicators. Appendix G.3 describes the selection, pre-processing and derivation steps required for the implementation of these semantic business rule indicators.

The selection, pre-processing and derivation logic resulted in the implementation of a final version of the unit standard enrolment data as a new table called DM\_USTD\_ENROL\_FINAL. A technical description of the table DM\_USTD\_ENROL\_FINAL is provided in G.1.

### ***3.8.7 Data mining methods***

Three specific data mining techniques are utilized in this study, each for specific purposes. The most frequently utilized data mining technique is EDM techniques which summarizes the data, thereby allowing for the visualization of the data and the identification of hidden relationships. In some instances, although EDM techniques suggest that a hidden relationship exists, the data is too large or the relationship too diverse and as a result cluster

data mining techniques are implemented in order to better describe the relationships. In contrast, association data mining techniques are implemented to determine whether any relationships exist in the data that could not be discerned by either EDM techniques or cluster data mining techniques.

A description of each of these data mining techniques as implemented in this study is provided in Appendix I.

### **3.9 Chapter summary**

This chapter details the framework, methodology, process, data collection and overarching data derivation considerations for the research. Further, the chapter covers the selection of the variables required for the research and the pre- processing and derivation of the data in preparation for the research. Finally, this chapter describes the data mining techniques applied in this research. The next chapter presents the results related to the research methods and steps within the methodological framework.

## 4 Chapter 4: Data analysis and research findings

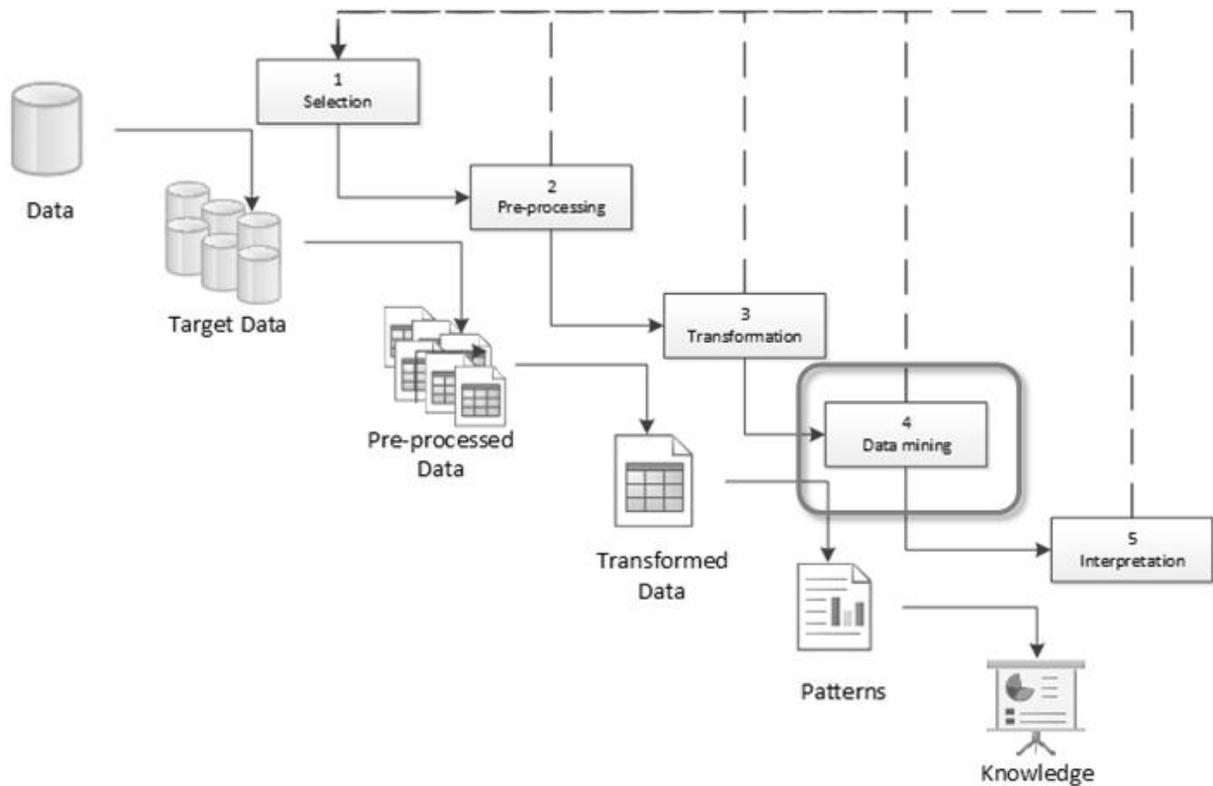


Figure 4.1 KDD phases – Data Mining

### 4.1 Introduction

This chapter presents the results of the analysis of the learnership, qualification and unit standard enrolments in relation to the nineteen (19) applicable semantic business rules as defined in Appendix A.7. The results of each analysis are presented by semantic business rule and enrolment record type.

Each data set is prepared prior to data analysis (see Appendix C, Appendix E and Appendix G) to ensure that the data is in a format that is suitable for data mining. Further, the data is prepared in order to address the overarching data derivation considerations highlighted in Section 3.8.3.

Each enrolment dataset is analysed according to a specific semantic business rule. The analysis of the data starts with EDM in order to provide an overview of the data. Where the number of records that do not comply with a semantic business rule exceeds 5%, and the

results lend themselves to further data mining efforts, the data subset is further analysed utilizing cluster data mining techniques. Where cluster data mining techniques are applied both a description of the most pertinent aspects of the resultant clusters and the technical description of the cluster are provided. Further, a summary of semantic infringements by ETQE is also provided in order to provide clarity in regard to the results when compared to an overall view of the percentage of infringements by ETQE.

Finally, the data is further analysed using association data mining techniques where associations are sought amongst records that have one or more semantic business rule infringement. These analyses are conducted by each type of enrolment record.

## **4.2 ETQE accreditation**

This section presents the results of the analysis of learner enrolment records in relation to whether the ETQE was accredited for the duration of the learner's active enrolment on a learnership, qualification or unit standard. The section therefore focuses on the nominal data value ETQE\_IND which contains a value denoting the record's compliance in regard to whether the ETQE was accredited for the duration of the learner's active enrolment on the learnership, qualification or unit standard.

This section presents the results of the analysis of this data field for learnership enrolment records, qualification enrolment records and unit standard enrolment records.

### ***4.2.1 Learnership enrolments***

As defined in Appendix C.2, the indicator ETQE\_IND denotes whether the ETQE was accredited for the duration of the learner's active enrolment on the learnership. The manner in which the categories in this indicator are derived for learnership enrolment records is detailed in Appendix C.3.4. An overview of the derived categories, with ETQE\_IND\_DESC shown as Description and the frequency of these values as a percentage, is presented in Table 4.2.1.1:

Table 4.2.1.1 ETQE accreditation categories for learnership enrolments

Description	% Records
OK	95.21%
Start Before, End Before	0.05%
Start Before, End During	0.47%
Start During, End After	0.00%
Start During, End After Predicted	3.23%
Submitting ETQE: Start Before, End During, Other ETQE: Start During, End After	0.55%
Submitting ETQE: Start During, End After, Other ETQE: Start After, End After Predicted	0.47%
Submitting ETQE: Start During, End After, Other ETQE: Start Before, End During	0.00%
Submitting ETQE: Start During, End After, Other ETQE: Start During, End After Predicted	0.00%
<b>Total</b>	<b>100.00%</b>

An overview of what the text in these categories denotes is as follows:

- ‘OK’ indicates that the ETQE was accredited for the duration of the learner’s active enrolment on the learnership,
- ‘Start After’ indicates that the active time period of the learnership enrolment record started after the ETQE’s active accreditation time period,
- ‘Start Before’ indicates that the active time period of the learnership enrolment record started before the ETQE’s active accreditation time period,
- ‘Start During’ indicates that the active time period of the learnership enrolment record started during the ETQE’s active accreditation time period,
- ‘End After’ indicates that the active time period of the learnership enrolment record ended after the ETQE’s active accreditation time period,
- ‘End Before’ indicates that the active time period of the learnership enrolment record ended before the ETQE’s active accreditation time period,
- ‘End During’ indicates that the active time period of the learnership enrolment record ended during the ETQE’s active accreditation time period,
- ‘Submitting ETQE’ and ‘Other ETQE’ indicate an ETQE amalgamation, and
- ‘Predicted’ indicates a current learnership enrolment record that has not yet been completed and the expected active enrolment on the learnership has not yet expired.

Any record with a category that ends with the text ‘Predicted’ is a current learnership enrolment. The data of the learnership enrolment record or the data in the ETQE table for

these types of records may change before the learnership enrolment record's active time period expires. As a result of this uncertainty it was decided in consultation with the Director of the NLRD that these types of records will be assumed as correct for the purposes of this research.

Categories that contain text like 'Submitting ETQE' and 'Other ETQE' indicate an ETQE amalgamation. The following types of categories indicate a situation that describes a normal progression of a learnership enrolment found in an ETQE amalgamation and as a result were considered correct for the purposes of this research:

- 'Submitting ETQE: Start Before, End During, Other ETQE: Start During, End After'
- 'Submitting ETQE: Start During, End After, Other ETQE: Start Before, End During'

A preliminary investigation was conducted on the remaining categories of records namely 'Start Before, End Before', 'Start Before, End During' and 'Start During, End After'.

It is found that the records in the category 'Start Before, End Before' denote specific situations in which the ETQE, that results after an amalgamation, has found that a previous ETQE had not submitted data in regard to a specific learnership and related learnership enrolment records to the NLRD. The current ETQE has, on request of the Director of the NLRD, submitted the missing learnership and learnership enrolment records to the NLRD. In these specific cases there is no data in the NLRD that defines a relationship between the learnership, the previous ETQE and the current ETQE. This issue is limited to 8 learnerships. In consultation with the Director of the NLRD it was decided that these types of records need to be assumed as correct for the purposes of this research.

However records that fall into the categories 'Start Before, End During' and 'Start During, End After' denote records that infringe on this particular semantic business rule.

As a result the only categories of records that are considered for this research have a description of 'Start Before, End During' or 'Start During, End After'. Figure 4.2.1.1 presents an overview of the percentage of records that infringe on the semantic business rule that requires that the ETQE must be accredited for the duration of the learner's active enrolment on the learnership.

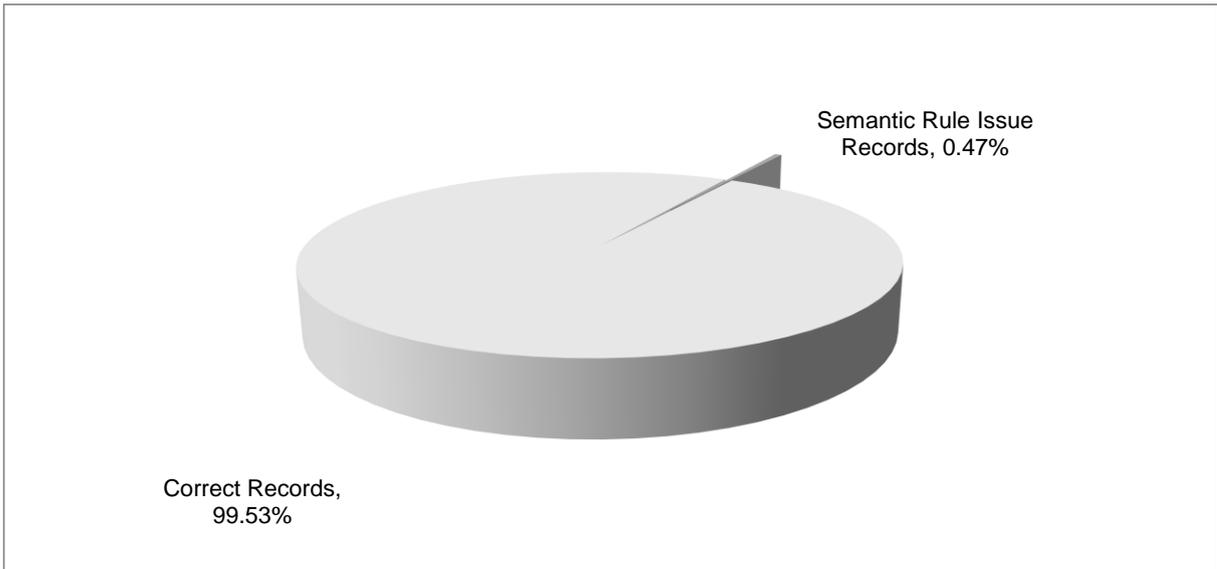


Figure 4.2.1.1 % records according to the semantic business rule that requires that the ETQE must be accredited for the duration of the learner’s active enrolment on the learnership

The total percentage of records that infringe on this semantic business rule is very low, namely 0.47%. The low infringement incidence rate could be attributed to the fact that SAQA manages the data that describes ETQEs on the NLRD (see Section 3.6.3.1.a). The records that infringe on this semantic business rule are comprised of two categories:

- Start Before, End During (99.89%)

This category indicates that the learnership enrolment started before and either was completed or expired whilst the ETQE was accredited.

Of the 27 discrete ETQEs in the dataset, 7 ETQEs are linked to this category. The majority of these records (95.93%) belong to a single ETQE and learnership. On further investigation it is found that these records represent learnership enrolments against a learnership which is noted as a special case whilst analysing the ‘Start Before, End Before’ category for this semantic business rule.

These records fall into this category for the same reasons as stated in the analysis of the ‘Start Before, End Before’ category. The ETQE that results after an amalgamation has found that a previous ETQE had not submitted data in regard to a specific learnership and related learnership enrolment records to the NLRD. The current ETQE has, on request of the Director of the NLRD, submitted the missing learnership and learnership

enrolment records to the NLRD. In these specific cases there is no data in the NLRD that defines a relationship between the learnership, the previous ETQE and the current ETQE. As in the case of the ‘Start Before, End Before’ records, these specific records are also assumed to be correct.

The remaining records (4.07%) found in this category are shared by six different ETQEs across 14 different learnerships. The low incidence of records that fall into this category, combined with the distribution of these records across so many ETQEs and learnerships suggests that these records are found in this category as a result of incorrect data capturing at the source of the data.

- **Start During, End After (0.11%)**

This category indicates that the learnership enrolment started whilst the ETQE was accredited and either was completed or expired after the ETQE was no longer accredited.

All of these records belong to a single ETQE and learnership. These records represent 0.02% of the records submitted by the ETQE to the NLRD and 0.06% of the records submitted to the NLRD for this specific learnership. The low percentage of the overall number of records submitted to the NLRD for this specific learnership suggests that these records are found in this category as a result of incorrect data capturing at the source of the data.

Overall the results for this semantic business rule indicate very few issues exist in regard to whether the ETQE is accredited for the duration of the learner’s active enrolment on the learnership. As already stated this result could be attributed to the fact that SAQA maintains ETQE related data in the NLRD. The incidence of infringements against this rule is so low and/or so diverse as to suggest that the issues arise as a result of data capturing problems rather than systemic issues with ETQE accreditation records.

#### ***4.2.2 Qualification enrolments***

As defined in Appendix E.2, the indicator ETQE\_IND denotes whether the ETQE was accredited for the duration of the learner’s active enrolment on the qualification. The manner in which the categories in this indicator are derived for qualification enrolment

records is detailed in Appendix E.3.4. An overview of the derived categories, with ETQE\_IND\_DESC shown as Description and the frequency of these values as a percentage, is presented in Table 4.2.2.1:

Table 4.2.2.1 ETQE accreditation categories for qualification enrolments

Description	% Records
OK	97.03%
Start Before, End Before	0.41%
Start Before, End Before (Qual Linked to Lshp)	0.02%
Start Before, End During	0.97%
Start Before, End During (Qual Linked to Lshp)	0.01%
Start During, End After Predicted	0.28%
Start During, End After Predicted (Qual Linked to Lshp)	0.00%
Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End After (Qual Linked to Lshp)	0.00%
Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End Before	0.00%
Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End Before (Qual Linked to Lshp)	0.00%
Submitting ETQE: Start Before, End During, Other ETQE: Start During, End After	1.12%
Submitting ETQE: Start Before, End During, Other ETQE: Start During, End After (Qual Linked to Lshp)	0.12%
Submitting ETQE: Start During, End After, Other ETQE: Start After, End After Predicted	0.02%
Submitting ETQE: Start During, End After, Other ETQE: Start After, End After Predicted (Qual Linked to Lshp)	0.00%
Submitting ETQE: Start During, End After, Other ETQE: Start Before, End During	0.00%
Submitting ETQE: Start During, End After, Other ETQE: Start During, End After Predicted	0.00%
<b>Total</b>	<b>100.00%</b>

An overview of what the text in these categories denotes is as follows:

- ‘OK’ indicates that the ETQE was accredited for the duration of the learner’s active enrolment on the qualification,
- ‘Start After’ indicates that the active time period of the qualification enrolment record started after the ETQE’s active accreditation time period,
- ‘Start Before’ indicates that the active time period of the qualification enrolment record started before the ETQE’s active accreditation time period,
- ‘Start During’ indicates that the active time period of the qualification enrolment record started during the ETQE’s active accreditation time period,
- ‘End After’ indicates that the active time period of the qualification enrolment record ended after the ETQE’s active accreditation time period,
- ‘End Before’ indicates that the active time period of the qualification enrolment record ended before the ETQE’s active accreditation time period,
- ‘End During’ indicates that the active time period of the qualification enrolment record ended during the ETQE’s active accreditation time period,
- ‘Submitting ETQE’ and ‘Other ETQE’ indicate an ETQE amalgamation,

- ‘Predicted’ indicates a current qualification enrolment record that has not yet been achieved and the expected active enrolment on the qualification has not yet expired, and
- ‘(Qual Linked to Lshp)’ indicates that the qualification is linked to a learnership and as a result the problem may be related to the issue described in Section 3.8.3.7.

Any record with a category that ends with the text ‘(Qual Linked to Lshp)’ may have infringed on this semantic business rule as a result of the contextual issues described in Section 3.8.3.7. It was decided, in consultation with the Director of the NLRD, that the further analysis of these types of records will not form part of this research. The percentage of records for this specific issue is however illustrated for all of the semantic business rules in order to provide an understanding of how the specific issue impacts the data in the NLRD.

Any record with a category that ends with the text ‘Predicted’ is a current qualification enrolment. The data of the qualification enrolment record or the data in the ETQE table for these types of records may change before the qualification enrolment record’s active time period expires. As a result of this uncertainty it was decided in consultation with the Director of the NLRD that these types of records will be assumed as correct for the purposes of this research.

Categories that contain text like ‘Submitting ETQE’ and ‘Other ETQE’ indicate an ETQE amalgamation. The following types of categories indicate a situation that describes a normal progression of a qualification enrolment found in an ETQE amalgamation and as a result are considered correct for the purposes of this research:

- ‘Submitting ETQE: Start Before, End During, Other ETQE: Start During, End After’
- ‘Submitting ETQE: Start During, End After, Other ETQE: Start Before, End During’

As a result the only categories of records that are considered for this research have a description of ‘Start Before, End During’, ‘Start Before, End Before’ or ‘Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End Before’. Figure 4.2.2.1 presents an overview of the percentage of records that infringe on the semantic business

rule that requires that the ETQE must be accredited for the duration of the learner's active enrolment on the qualification.

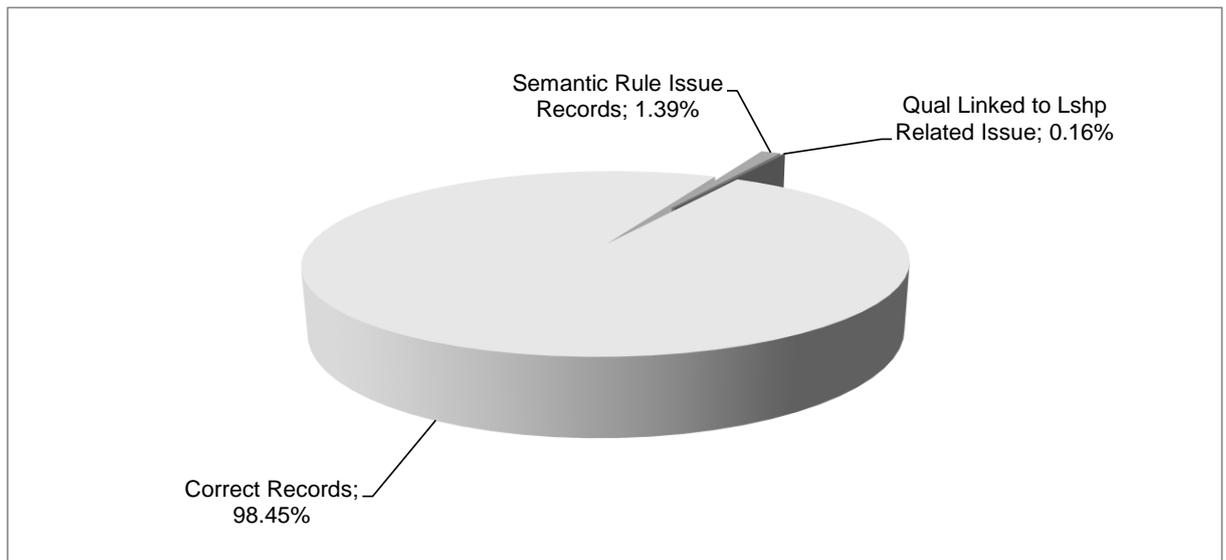


Figure 4.2.2.1 % records according to the semantic business rule that requires that the ETQE must be accredited for the duration of the learner's active enrolment on the qualification

The total percentage of records that infringe on this semantic business rule is very low, namely 1.39%. The low infringement incidence rate could be attributed to the fact that SAQA manages the data that describes ETQEs on the NLRD (see Section 3.6.3. 1.a). The records that infringe on this semantic business rule are comprised of three categories:

- Start Before, End During (70.00%)

This category indicates that the qualification enrolment started before and either was achieved or expired whilst the ETQE was accredited.

Of the 29 discrete ETQEs in the dataset, 7 ETQEs are linked to this category. The majority of these records (85.35%) belong to a single ETQE and qualification. It is found that these specific records denote a specific situation in which the ETQE, which results after an amalgamation, has found that a previous ETQE had not submitted data in regard to specific qualification enrolment records to the NLRD. The current ETQE has, on request of the Director of the NLRD, submitted the missing qualification enrolment records to the NLRD. In these specific cases there is no data in the NLRD

that defines a relationship between the qualification, the previous ETQE and the current ETQE. This issue was limited to 5 qualifications. In consultation with the Director of the NLRD it was decided that these types of records need to be assumed as correct for the purposes of this research.

The remaining records (14.65%) found in this category are shared by six different ETQEs across 43 different qualifications. The low incidence of records that fall into this category, combined with the distribution of these records across so many ETQEs and qualifications suggests that these records are found in this category as a result of incorrect data capturing at the source of the data.

- Start Before, End Before (29.83%)

This category indicates that the qualification enrolment started before and either was achieved or expired before the ETQE was accredited.

Of the 29 discrete ETQEs in the dataset, 4 ETQEs are linked to this category. The majority of these records (96.42%) belong to a single ETQE and 3 qualifications.

As was found for the records that exist in the ‘Start Before, End During’ category for the same ETQE, these records represent qualification enrolments which are noted as a special case.

The ETQE that results after an amalgamation has found that a previous ETQE had not submitted data in regard to specific qualification enrolment records to the NLRD. The current ETQE has, on request of the Director of the NLRD, submitted the missing qualification enrolment records to the NLRD. In these specific cases there is no data in the NLRD that defines a relationship between the qualification, the previous ETQE and the current ETQE. As in the case of the ‘Start Before, End During’ records, these specific records are also assumed to be correct.

The remaining records (3.58%) in this category are shared by three different ETQEs across 24 different qualifications. The low incidence of records that fall into this category, combined with the distribution of these records across so many ETQEs and

qualifications suggests that these records are found in this category as a result of incorrect data capturing at the source of the data.

- Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End Before (0.17%)

This category indicates that the qualification enrolment:

- started before one of the members of an ETQE amalgamation was accredited and either was achieved or expired whilst the same ETQE was accredited, and
- started before the other member of an ETQE amalgamation was accredited and either was achieved or expired before the same ETQE was accredited.

All of these records belong to a single ETQE and 8 qualifications. These records represent 0.11% of the records submitted by the ETQE to the NLRD. The low percentage of the overall number of records submitted to the NLRD for this specific ETQE spread over 8 qualifications suggests that these records are found in this category as a result of incorrect data capturing at the source of the data.

Overall the results for this semantic business rule indicate very few issues exist in regard to whether the ETQE was accredited for the duration of the learner's active enrolment on the qualification. As already stated this result could be attributed to the fact that SAQA maintains ETQE related data in the NLRD. The incidence of infringements against this rule is so low and/or so diverse as to suggest that the issues arise as a result of data capturing problems rather than systemic issues with ETQE accreditation records.

#### ***4.2.3 Unit Standard enrolments***

As defined in Appendix G.2, the indicator ETQE\_IND denotes whether the ETQE was accredited for the duration of the learner's active enrolment on the unit standard. The manner in which the categories in this indicator are derived for unit standard enrolment records is detailed in Appendix G.3.4. An overview of the derived categories, with ETQE\_IND\_DESC shown as Description and the frequency of these values as a percentage, is presented in Table 4.2.3.1:

Table 4.2.3.1 ETQE accreditation categories for unit standard enrolments

Description	% Records
OK	95.01%
Start Before, End Before	2.60%
Start Before, End Before (UStd Linked to Lshp)	0.01%
Start Before, End During	2.38%
Start Before, End During (UStd Linked to Lshp)	0.00%
Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End Before	0.00%
Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End Before (UStd Linked to Lshp)	0.00%
Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End During	0.00%
Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End During	0.00%
Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End During (UStd Linked to Lshp)	0.00%
Submitting ETQE: Start Before, End During, Other ETQE: Start During, End After	0.00%
<b>Total</b>	<b>100.00%</b>

An overview of what the text in these categories denotes is as follows:

- ‘OK’ indicates that the ETQE was accredited for the duration of the learner’s active enrolment on the unit standard,
- ‘Start After’ indicates that the active time period of the unit standard enrolment record started after the ETQE’s active accreditation time period,
- ‘Start Before’ indicates that the active time period of the unit standard enrolment record started before the ETQE’s active accreditation time period,
- ‘Start During’ indicates that the active time period of the unit standard enrolment record started during the ETQE’s active accreditation time period,
- ‘End After’ indicates that the active time period of the unit standard enrolment record ended after the ETQE’s active accreditation time period,
- ‘End Before’ indicates that the active time period of the unit standard enrolment record ended before the ETQE’s active accreditation time period,
- ‘End During’ indicates that the active time period of the unit standard enrolment record ended during the ETQE’s active accreditation time period,
- ‘Submitting ETQE’ and ‘Other ETQE’ indicate an ETQE amalgamation,
- ‘Predicted’ indicates a current unit standard enrolment record that has not yet been achieved and the expected active enrolment on the unit standard has not yet expired, and
- ‘(UStd Linked to Lshp)’ indicates that the unit standard is linked to a learnership and as a result the problem may be related to the issue described in Section 3.8.3.7.

Any record with a category that ends with the text ‘(UStd Linked to Lshp)’ may have infringed on this semantic business rule as a result of the contextual issues described in Section 3.8.3.7. It was decided, in consultation with the Director of the NLRD, that the further analysis of these types of records will not form part of this research. The percentage of records for this specific issue is however illustrated for all of the semantic business rules in order to provide an understanding of how the specific issue impacts the data in the NLRD.

Any record with a category that ends with the text ‘Predicted’ is a current unit standard enrolment. The data of the unit standard enrolment record or the data in the ETQE table for these types of records may change before the unit standard enrolment record’s active time period expires. As a result of this uncertainty it was decided in consultation with the Director of the NLRD that these types of records will be assumed as correct for the purposes of this research.

Categories that contain text like ‘Submitting ETQE’ and ‘Other ETQE’ indicate an ETQE amalgamation. The following types of categories indicate a situation that describes a normal progression of a unit standard enrolment found in an ETQE amalgamation and as a result were considered correct for the purposes of this research:

- ‘Submitting ETQE: Start Before, End During, Other ETQE: Start During, End After’
- ‘Submitting ETQE: Start During, End After, Other ETQE: Start Before, End During’

As a result the only categories of records that are considered for this research have a description of ‘Start Before, End Before’, ‘Start Before, End During’, ‘Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End During’, ‘Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End During’ or ‘Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End Before’. Figure 4.2.3.1 presents an overview of the percentage of records that infringe on the semantic business rule that requires that the ETQE must be accredited for the duration of the learner’s active enrolment on the unit standard.

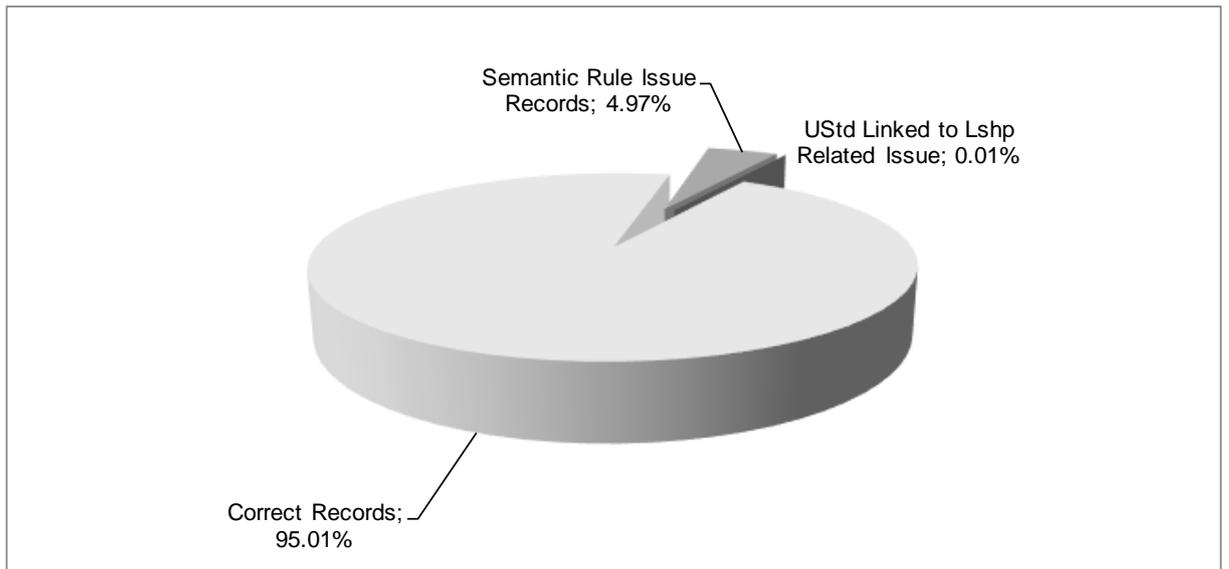


Figure 4.2.3.1 % records according to the semantic business rule that requires that the ETQE must be accredited for the duration of the learner’s active enrolment on the unit standard

The total percentage of records that infringe on this semantic business rule is relatively low, namely 4.97%. The low infringement incidence rate could be attributed to the fact that SAQA manages the data that describes ETQEs on the NLRD (see Section 3.6.3.1.a). The records that infringe on this semantic business rule are comprised of five categories:

- Start Before, End Before (52.22%)

This category indicates that the unit standard enrolment started before and either was achieved or expired before the ETQE was accredited.

Of the 29 discrete ETQEs in the dataset, 16 ETQEs are linked to this category. The majority of these records (98.30%) belong to 4 ETQEs and 557 unit standards. It is found that these specific records denote a specific situation in which the ETQE, which results after an amalgamation, has found that a previous ETQE had not submitted data in regard to specific qualification enrolment records to the NLRD. The current ETQE has, on request of the Director of the NLRD, submitted the missing qualification enrolment and their related unit standard enrolment records to the NLRD. In these specific cases there is no data in the NLRD that defines a relationship between the qualification, the previous ETQE and the current ETQE. Consequently, the relationship that describes the current ETQE and the unit standard is also missing. In

consultation with the Director of the NLRD it was decided that these types of records need to be assumed as correct for the purposes of this research.

The remaining records (1.70%) found in this category are shared by 12 different ETQEs across 1067 different unit standards. The low incidence of records that fall into this category, combined with the distribution of these records across so many ETQEs and unit standards suggests that these records are found in this category as a result of incorrect data capturing at the source of the data.

- Start Before, End During (47.77%)

This category indicates that the unit standard enrolment started before and either was achieved or expired whilst the ETQE was accredited.

Of the 29 discrete ETQEs in the dataset, 14 ETQEs are linked to this category. The majority of these records (97.47%) belong to 3 ETQEs and 395 unit standards.

As is found for the records that exist in the ‘Start Before, End Before’ category for the same ETQEs, these records represent unit standard enrolments which have been noted as a special case.

The ETQE that results after an amalgamation has found that a previous ETQE had not submitted data in regard to specific qualification enrolment records to the NLRD. The current ETQE has, on request of the Director of the NLRD, submitted the missing qualification enrolment records and their related unit standard enrolment records to the NLRD. In these specific cases there is no data in the NLRD that defines a relationship between the qualification, the previous ETQE and the current ETQE. Consequently, the relationship that describes the current ETQE and the unit standard is also missing. As in the case of the ‘Start Before, End Before’ records, these specific records are also assumed to be correct.

The remaining records (2.53%) in this category are shared by 11 different ETQEs across 1252 different unit standards. The low incidence of records that fall into this category, combined with the distribution of these records across so many ETQEs and

unit standards suggests that these records are found in this category as a result of incorrect data capturing at the source of the data.

- Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End During (0.01%)

This category indicates that the unit standard enrolment:

- started before one of the members of an ETQE amalgamation was accredited and either was achieved or expired before the same ETQE was accredited, and
- started before the other member of an ETQE amalgamation was accredited and either was achieved or expired whilst the same ETQE was accredited.

These records belong to 2 ETQEs and 2 unit standards. These records represent 0.02% of the records submitted by these ETQEs to the NLRD. The low percentage of the overall number of records submitted to the NLRD for these specific ETQEs spread over 2 unit standards suggests that these records are found in this category as a result of incorrect data capturing at the source of the data.

- Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End During (0.00%)

This category indicates that the unit standard enrolment:

- started before one of the members of an ETQE amalgamation was accredited and either was achieved or expired whilst the same ETQE was accredited, and
- started before the other member of an ETQE amalgamation was accredited and either was achieved or expired whilst the same ETQE was accredited.

These records belong to 2 ETQEs and 2 unit standards. These records represent 0.01% of the records submitted by these ETQEs to the NLRD. The low percentage of the overall number of records submitted to the NLRD for these specific ETQEs spread over 2 unit standards suggests that these records are found in this category as a result of incorrect data capturing at the source of the data.

- Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End Before (0.00%)

This category indicates that the unit standard enrolment:

- started before one of the members of an ETQE amalgamation was accredited and either was achieved or expired before the same ETQE was accredited, and
- started before the other member of an ETQE amalgamation was accredited and either was achieved or expired before the same ETQE was accredited.

These records belong to 2 ETQEs and 2 unit standards. These records represent 0.02% of the records submitted by these ETQEs to the NLRD. The low percentage of the overall number of records submitted to the NLRD for these specific ETQEs spread over 2 unit standards suggests that these records are found in this category as a result of incorrect data capturing at the source of the data.

Overall the results for this semantic business rule indicate very few issues exist in regard to whether the ETQE was accredited for the duration of the learner's active enrolment on the unit standard. As already stated, this result could be attributed to the fact that SAQA maintains ETQE related data in the NLRD. The incidence of infringements against this rule is so low and/or so diverse as to suggest that the issues arise as a result of data capturing problems rather than systemic issues with ETQE accreditation records.

#### ***4.2.4 Conclusion***

This section focuses on the analysis of the nominal data value ETQE\_IND which contains a value denoting the record's compliance in regard to whether the ETQE was accredited for the duration of the learner's active enrolment on the learnership, qualification or unit standard.

Overall the results for this semantic business rule indicate very few issues exist in regard to whether the ETQE was accredited for the duration of the learner's active enrolment. This result could be attributed to the fact that SAQA maintains ETQE related data in the NLRD. Generally, the incidence of infringements against this rule is so low and/or so diverse as to suggest that the issues arise as a result of data capturing problems rather than systemic issues with ETQE accreditation records.

The analysis of the learnership and qualification enrolment records show that very few issues exist in regard to whether the ETQE was accredited for the duration of the learner's

active enrolment. The number of qualification enrolment records that are not compliant with this rule is higher than that of learnership enrolment records. Further investigation shows that this is as a result of missing qualification enrolment records that were loaded into the NLRD, on request of the Director of the NLRD, after an ETQE amalgamation. These records fail compliance to this rule because the NLRD does not contain the required ETQE accreditation record for the enduring ETQE. Given that unit standard enrolment records are generally linked to qualification enrolment records, this issue is further propagated in unit standard enrolment records which in turn had an even larger number of records that are not compliant with this rule.

Specific recommendations in regard to data records that are not compliant to this semantic business rule are provided in Appendix P.1.1 for learnership enrolments, Appendix P.1.2 for qualification enrolments and Appendix P.1.3 for unit standard enrolments.

### **4.3 ETQE accreditation to quality assure the qualification or unit standard**

This section presents the results of the analysis of learner enrolment records in relation to whether the ETQE was accredited to quality assure the qualification/unit standard for the duration of the learner's active enrolment on the qualification/unit standard. The section therefore focuses on the nominal data value ETQE\_ACCRED\_IND which contains a value denoting the record's compliance in regard to whether the ETQE was accredited to quality assure the qualification/unit standard for the duration of the learner's active enrolment on the qualification or unit standard.

This section presents the results of the analysis of this data field for qualification enrolment records and unit standard enrolment records.

#### ***4.3.1 Qualification enrolments***

As defined in Appendix E.2, the indicator ETQE\_ACCRED\_IND denotes whether the ETQE was accredited to quality assure the qualification for the duration of the learner's active enrolment on the qualification. The manner in which the categories in this indicator are derived for qualification enrolment records is detailed in Appendix E.3.5. An overview of the derived categories, with ETQE\_ACCRED\_IND\_DESC shown as Description and the frequency of these values as a percentage, is presented in Table 4.3.1.1:

Table 4.3.1.1 ETQE accreditation to quality assure the qualification categories

Description	% Records
No Accreditation	0.04%
No Accreditation (Qual Linked to Lshp)	0.00%
OK	96.80%
Start After, End After	0.81%
Start After, End After (Qual Linked to Lshp)	0.01%
Start After, End After Predicted (Qual Linked to Lshp)	0.00%
Start Before, End After (Qual Linked to Lshp)	0.00%
Start Before, End Before	0.00%
Start Before, End Before (Qual Linked to Lshp)	0.18%
Start Before, End During	0.16%
Start Before, End During (Qual Linked to Lshp)	0.87%
Start During, End After	0.47%
Start During, End After (Qual Linked to Lshp)	0.00%
Start During, End After Predicted	0.13%
Submitting ETQE: Start After, End After, Other ETQE: Start After, End After (Qual Linked to Lshp)	0.00%
Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End Before	0.00%
Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End Before (Qual Linked to Lshp)	0.02%
Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End After (Qual Linked to Lshp)	0.00%
Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End During	0.02%
Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End During (Qual Linked to Lshp)	0.44%
Submitting ETQE: Start During, End After, Other ETQE: Start After, End After	0.02%
Submitting ETQE: Start During, End After, Other ETQE: Start During, End After	0.02%
Submitting ETQE: Start During, End After, Other ETQE: Start During, End After (Qual Linked to Lshp)	0.00%
<b>Total</b>	<b>100.00%</b>

An overview of what the text in these categories denotes is as follows:

- ‘OK’ indicates that the ETQE was accredited to quality assure the qualification for the duration of the learner’s active enrolment on the qualification,
- ‘Start After’ indicates that the active time period of the qualification enrolment record started after the ETQE’s active accreditation to quality assure the qualification time period,
- ‘Start Before’ indicates that the active time period of the qualification enrolment record started before the ETQE’s active accreditation to quality assure the qualification time period,
- ‘Start During’ indicates that the active time period of the qualification enrolment record started during the ETQE’s active accreditation to quality assure the qualification time period,
- ‘End After’ indicates that the active time period of the qualification enrolment record ended after the ETQE’s active accreditation to quality assure the qualification time period,
- ‘End Before’ indicates that the active time period of the qualification enrolment record ended before the ETQE’s active accreditation to quality assure the qualification time period,

- ‘End During’ indicates that the active time period of the qualification enrolment record ended during the ETQE’s active accreditation to quality assure the qualification time period,
- ‘Submitting ETQE’ and ‘Other ETQE’ indicate an ETQE amalgamation,
- ‘Predicted’ indicates a current qualification enrolment record that has not yet been achieved and the expected active enrolment on the qualification has not yet expired, and
- ‘(Qual Linked to Lshp)’ indicates that the qualification is linked to a learnership and as a result the problem may be related to the issue described in Section 3.8.3.7.

Any record with a category that ends with the text ‘(Qual Linked to Lshp)’ may have infringed on this semantic business rule as a result of the contextual issues described in Section 3.8.3.7. It was decided, in consultation with the Director of the NLRD, that the further analysis of these types of records will not form part of this research. The percentage of records for this specific issue is however illustrated for all of the semantic business rules in order to provide an understanding of how the specific issue impacts the data in the NLRD.

Any record with a category that ends with the text ‘Predicted’ is a current qualification enrolment. The data of the qualification enrolment record or the data in the ETQE Accreditation table for these types of records may change before the qualification enrolment record’s active time period expires. As a result of this uncertainty it was decided in consultation with the Director of the NLRD that these types of records are assumed as correct for the purposes of this research.

Categories that contain text like ‘Submitting ETQE’ and ‘Other ETQE’ indicate an ETQE amalgamation. The following types of categories indicate a situation that describes a normal progression of a qualification enrolment found in an ETQE amalgamation and as a result were considered correct for the purposes of this research:

- ‘Submitting ETQE: Start Before, End During, Other ETQE: Start During, End After’
- ‘Submitting ETQE: Start During, End After, Other ETQE: Start Before, End During’

As a result the only categories of records that are considered for this research have a description of 'No Accreditation', 'Start After, End After', 'Start Before, End Before', 'Start Before, End During', 'Start During, End After', 'Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End Before', 'Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End During', 'Submitting ETQE: Start During, End After, Other ETQE: Start After, End After' or 'Submitting ETQE: Start During, End After, Other ETQE: Start During, End After'. Figure 4.3.1.1 presents an overview of the percentage of records that infringe on the semantic business rule that requires that the ETQE must be accredited to quality assure the qualification for the duration of the learner's active enrolment on the qualification.

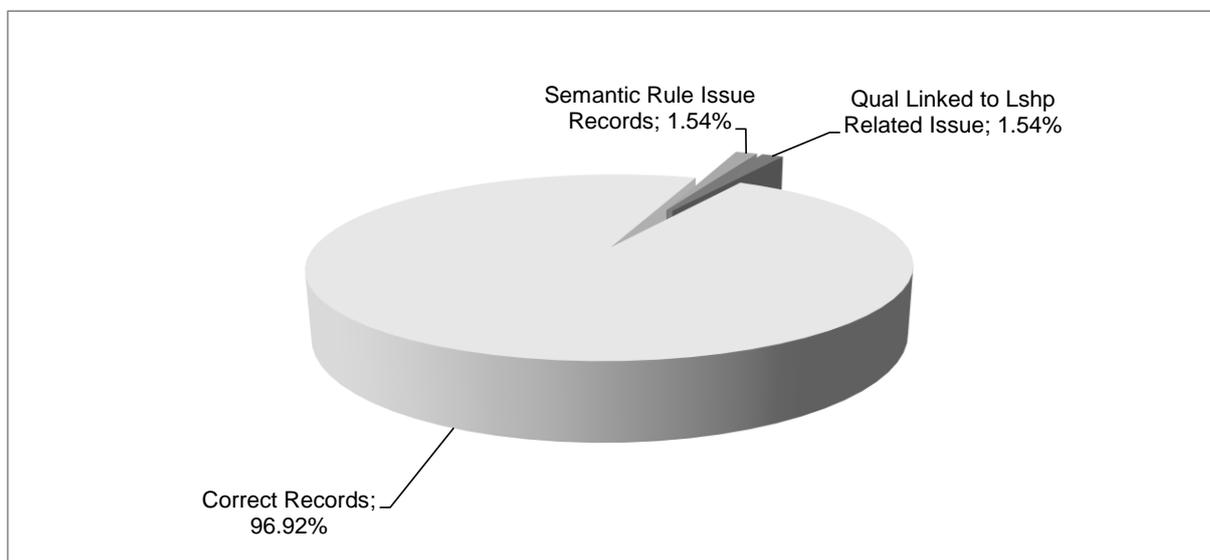


Figure 4.3.1.1 % records according to the semantic business rule that requires that the ETQE must be accredited to quality assure the qualification for the duration of the learner's active enrolment on the qualification

The total percentage of records that infringe on this semantic business rule is very low, namely 1.54%. The low infringement incidence rate could be attributed to the fact that SAQA manages the data that describes ETQEs on the NLRD (see Section 3.6.3.1.a).

As indicated in the ETQE accreditation semantic business rule analysis (see Section 4.2.2) there is a specific situation in which the ETQE, which results after an amalgamation, has found that a previous ETQE had not submitted data in regard to specific qualification enrolment records to the NLRD. The current ETQE has, on request of the Director of the

NLRD, submitted the missing qualification enrolment records to the NLRD. In these specific cases there is no data in the NLRD that defines a relationship between the qualification, the previous ETQE and the current ETQE. In consultation with the Director of the NLRD it was decided that these types of records need to be assumed as correct for the purposes of this research.

The records that infringe on this semantic business rule are comprised of the following 9 categories:

- Start After, End After (52.53%)

This category indicates that the qualification enrolment started after and either was achieved or expired after the ETQE was accredited to quality assure the qualification.

Of the 29 discrete ETQEs in the dataset, 2 ETQEs are linked to this category. The majority of these records (99.23%) belong to a single ETQE (ETQE Identifier 1103). Most notably these records constitute 10.82% of the records submitted to the NLRD by this ETQE.

Of the 861 discrete qualifications in the dataset, 14 qualifications are linked to this category. Of these 14 qualifications, 10 qualifications contribute to 97.64% of records in this category.

- Start During, End After (30.68%)

This category indicates that the qualification enrolment started during and either was achieved or expired after the ETQE was accredited to quality assure the qualification.

Of the 29 discrete ETQEs in the dataset, 5 ETQEs are linked to this category. The majority of these records (77.11%) belong to a single ETQE (ETQE Identifier 1103). Most notably these records constitute 4.91% of the records submitted to the NLRD by this ETQE.

Of the 861 discrete qualifications in the dataset, 18 qualifications are linked to this category. Of these 18 qualifications, 10 qualifications contribute to 96.82% of records in this category.

- Start Before, End During (10.22%)

This category indicates that the qualification enrolment started before and either was achieved or expired whilst the ETQE was accredited to quality assure the qualification.

Of the 29 discrete ETQEs in the dataset, 11 ETQEs are linked to this category. Of these records, 90.53% were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 15 qualifications are linked to this category. Of these 15 qualifications, 10 qualifications contribute to 99.55% of records in this category. Most notably, although one of the 10 qualifications only constitute 0.13% of the records; the records for this qualification represents 100% of the qualification enrolment records submitted to the NLRD for the qualification.

- No Accreditation (2.27%)

This category indicates that the ETQE that is linked to the qualification enrolment has never had an active accreditation to quality assure the qualification and this category is of greatest concern to SAQA.

Of the 29 discrete ETQEs in the dataset, 2 ETQEs are linked to this category. Of the 861 discrete qualifications in the dataset, 4 qualifications are linked to this category. The majority of these records (68.29%) belong to a single ETQE (ETQE Identifier 1104) and a single qualification. Most notably these records constitute 5.03% of the records submitted to the NLRD by this ETQE.

- Submitting ETQE: Start During, End After, Other ETQE: Start After, End After (1.53%)

This category indicates that the qualification enrolment:

- started during one of the members of an ETQE amalgamation was accredited to quality assure the qualification and either was achieved or expired whilst the same ETQE was no longer accredited to quality assure the qualification, and

- started after the other member of an ETQE amalgamation was accredited to quality assure the qualification and either was achieved or expired after the same ETQE was accredited to quality assure the qualification.

Of the 29 discrete ETQEs in the dataset only one ETQE (ETQE identifier 1125) and two qualifications are linked to this category.

- Submitting ETQE: Start During, End After, Other ETQE: Start During, End After (1.31%)

This category indicates that the qualification enrolment:

- started during one of the members of an ETQE amalgamation was accredited to quality assure the qualification and either was achieved or expired whilst the same ETQE was no longer accredited to quality assure the qualification, and
- started during the other member of an ETQE amalgamation accredited to quality assure the qualification and either was achieved or expired after the same ETQE was accredited to quality assure the qualification.

Of the 29 discrete ETQEs in the dataset only 1 ETQE (ETQE identifier 1125) and four qualifications are linked to this category.

- Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End During (1.05%)

This category indicates that the qualification enrolment:

- started before one of the members of an ETQE amalgamation was accredited to quality assure the qualification and either was achieved or expired whilst the same ETQE was accredited to quality assure the qualification, and
- started before the other member of an ETQE amalgamation was accredited to quality assure the qualification and either was achieved or expired whilst the same ETQE was accredited to quality assure the qualification.

Of the 29 discrete ETQEs in the dataset, 5 ETQEs are linked to this category. Of these records, 93.83% were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 10 qualifications are linked to this category. Most notably, although one of the 10 qualifications only constitute 11.73% of the records; the records for this qualification represents 100% of the qualification enrolment records submitted to the NLRD for the qualification.

- Start Before, End Before (0.31%)

This category indicates that the qualification enrolment started before and either was achieved or expired before the ETQE was accredited to quality assure the qualification.

Of the 29 discrete ETQEs in the dataset, 5 ETQEs are linked to this category. Of these records, 95.74% were submitted to the NLRD by 3 ETQEs. Of the 861 discrete qualifications in the dataset, 7 qualifications are linked to this category.

- Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End Before (0.08%)

This category indicates that the qualification enrolment:

- started before one of the members of an ETQE amalgamation was accredited to quality assure the qualification and either was achieved or expired before the same ETQE was accredited to quality assure the qualification, and
- started before the other member of an ETQE amalgamation was accredited to quality assure the qualification and either was achieved or expired before the same ETQE was accredited to quality assure the qualification.

Of the 29 discrete ETQEs in the dataset, 3 ETQEs are linked to this category. Of the 861 discrete qualifications in the dataset, 4 qualifications are linked to this category. Most notably, although two of the 4 qualifications constitute 76.92% of the records; the records for these qualifications represent 100% of the qualification enrolment records submitted to the NLRD for these qualifications.

Overall the results for this semantic business rule indicate very few issues exist in regard to whether the ETQE was accredited to quality assure the qualification for the duration of the learner's active enrolment on the qualification. As already stated this result could be

attributed to the fact that SAQA maintains ETQE related data in the NLRD. The incidence of infringements against this rule is so low and/or so diverse as to suggest that the issues arise as a result of data capturing problems rather than systemic issues with ETQE accreditation records. The generally high incidence of these types of records, when compared against the overall number of records submitted to the NLRD, for ETQE Identifier 1104 and 1103 may however indicate issues of a systemic nature at these two ETQEs.

#### ***4.3.2 Unit Standard enrolments***

As defined in Appendix G.2, the indicator ETQE\_ACCRED\_IND denotes whether the ETQE was accredited to quality assure the unit standard for the duration of the learner's active enrolment on the unit standard. The manner in which the categories in this indicator are derived for unit standard enrolment records is detailed in Appendix G.3.5. An overview of the derived categories, with ETQE\_ACCRED\_IND\_DESC shown as Description and the frequency of these values as a percentage, is presented in Table 4.3.2.1:

Table 4.3.2.1 ETQE accreditation to quality assure the unit standard categories

Description	% Records
No Accreditation	1.97%
No Accreditation (UStd Linked to Lshp)	0.00%
OK	96.05%
Start After, End After	0.00%
Start Before, End After (UStd Linked to Lshp)	0.00%
Start Before, End Before	1.06%
Start Before, End Before (UStd Linked to Lshp)	0.09%
Start Before, End During	0.46%
Start Before, End During (UStd Linked to Lshp)	0.07%
Start During, End After	0.00%
Start During, End After (UStd Linked to Lshp)	0.00%
Submitting ETQE: No Accreditation, Other ETQE: Start After, End After	0.00%
Submitting ETQE: No Accreditation, Other ETQE: Start After, End After (UStd Linked to Lshp)	0.00%
Submitting ETQE: No Accreditation, Other ETQE: Start Before, End Before	0.00%
Submitting ETQE: No Accreditation, Other ETQE: Start Before, End Before (UStd Linked to Lshp)	0.00%
Submitting ETQE: No Accreditation, Other ETQE: Start Before, End During	0.02%
Submitting ETQE: No Accreditation, Other ETQE: Start Before, End During (UStd Linked to Lshp)	0.00%
Submitting ETQE: Start After, End After, Other ETQE: Start After, End After	0.00%
Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End Before	0.11%
Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End Before (UStd Linked to Lshp)	0.00%
Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End After	0.00%
Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End After (UStd Linked to Lshp)	0.00%
Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End Before	0.00%
Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End Before (UStd Linked to Lshp)	0.00%
Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End During	0.14%
Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End During (UStd Linked to Lshp)	0.01%
Submitting ETQE: Start During, End After, Other ETQE: Start After, End After	0.00%
Submitting ETQE: Start During, End After, Other ETQE: Start During, End After	0.00%
<b>Total</b>	<b>100.00%</b>

An overview of what the text in these categories denotes is as follows:

- ‘OK’ indicates that the ETQE was accredited to quality assure the unit standard for the duration of the learner’s active enrolment on the unit standard,
- ‘Start After’ indicates that the active time period of the unit standard enrolment record started after the ETQE’s active accreditation to quality assure the unit standard time period,
- ‘Start Before’ indicates that the active time period of the unit standard enrolment record started before the ETQE’s active accreditation to quality assure the unit standard time period,
- ‘Start During’ indicates that the active time period of the unit standard enrolment record started during the ETQE’s active accreditation to quality assure the unit standard time period,
- ‘End After’ indicates that the active time period of the unit standard enrolment record ended after the ETQE’s active accreditation to quality assure the unit standard time period,

- ‘End Before’ indicates that the active time period of the unit standard enrolment record ended before the ETQE’s active accreditation to quality assure the unit standard time period,
- ‘End During’ indicates that the active time period of the unit standard enrolment record ended during the ETQE’s active accreditation to quality assure the unit standard time period,
- ‘Submitting ETQE’ and ‘Other ETQE’ indicate an ETQE amalgamation,
- ‘Predicted’ indicates a current unit standard enrolment record that has not yet been achieved and the expected active enrolment on the unit standard has not yet expired, and
- ‘(UStd Linked to Lshp)’ indicates that the unit standard is linked to a learnership and as a result the problem may be related to the issue described in Section 3.8.3.7.

Any record with a category that ends with the text ‘(UStd Linked to Lshp)’ may have infringed on this semantic business rule as a result of the contextual issues described in Section 3.8.3.7. It was decided, in consultation with the Director of the NLRD, that the further analysis of these types of records will not form part of this research. The percentage of records for this specific issue is however illustrated for all of the semantic business rules in order to provide an understanding of how the specific issue impacts the data in the NLRD.

Any record with a category that ends with the text ‘Predicted’ is a current unit standard enrolment. The data of the unit standard enrolment record or the data in the ETQE Accreditation table for these types of records may change before the unit standard enrolment record’s active time period expires. As a result of this uncertainty it was decided in consultation with the Director of the NLRD that these types of records will be assumed as correct for the purposes of this research.

Categories that contain text like ‘Submitting ETQE’ and ‘Other ETQE’ indicate an ETQE amalgamation. The following types of categories indicate a situation that describes a normal progression of a unit standard enrolment found in an ETQE amalgamation and as a result are considered correct for the purposes of this research:

- ‘Submitting ETQE: Start Before, End During, Other ETQE: Start During, End After’

- ‘Submitting ETQE: Start During, End After, Other ETQE: Start Before, End During’

As a result the only categories of records that are considered for this research have a description of 'No Accreditation', 'Start Before, End Before', 'Start Before, End During', 'Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End During', 'Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End Before', 'Submitting ETQE: No Accreditation, Other ETQE: Start Before, End During', 'Submitting ETQE: Start During, End After, Other ETQE: Start During, End After', 'Submitting ETQE: Start During, End After, Other ETQE: Start After, End After', 'Start During, End After', 'Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End After', 'Submitting ETQE: No Accreditation, Other ETQE: Start Before, End Before', 'Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End Before', 'Start After, End After', 'Submitting ETQE: Start After, End After, Other ETQE: Start After, End After' or 'Submitting ETQE: No Accreditation, Other ETQE: Start After, End After'. Figure 4.3.2.1 presents an overview of the percentage of records that infringe on the semantic business rule that requires that the ETQE must be accredited to quality assure the unit standard for the duration of the learner’s active enrolment on the unit standard.

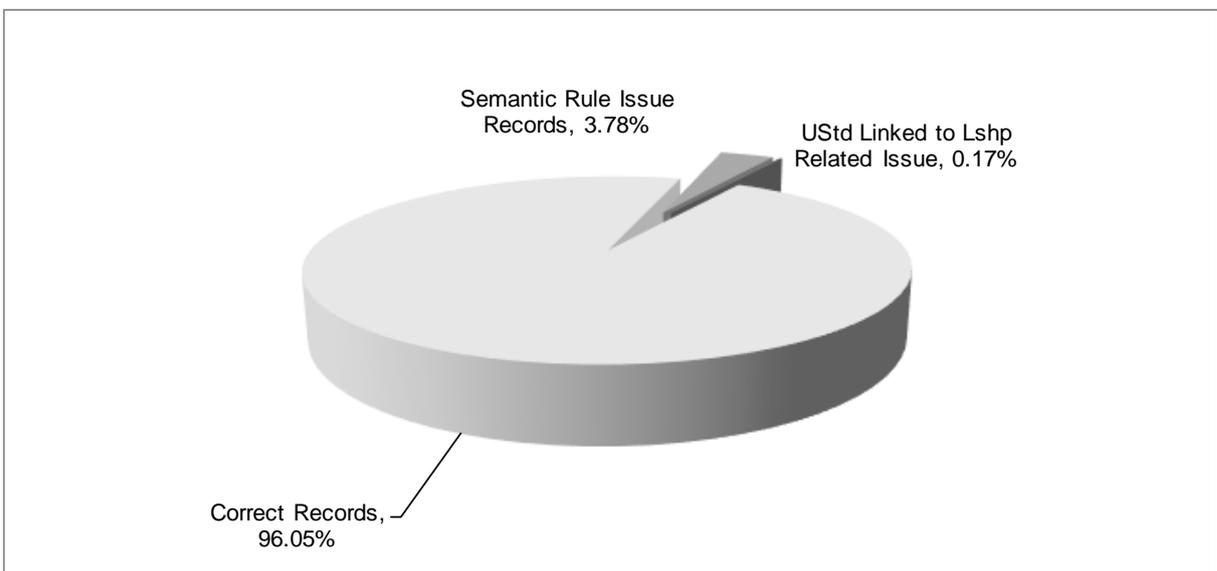


Figure 4.3.2.1 % records according to the semantic business rule that requires that the ETQE must be accredited to quality assure the unit standard for the duration of the learner’s active enrolment on the unit standard

The total percentage of records that infringe on this semantic business rule is relatively low, namely 3.78%. The low infringement incidence rate could be attributed to the fact that SAQA manages the data that describes ETQEs on the NLRD (see Section 3.6.3.1.a).

As indicated in the ETQE accreditation semantic business rule analysis (see Section 4.2.3) there is a specific situation in which the ETQE, which results after an amalgamation, has found that a previous ETQE had not submitted data in regard to specific qualification enrolment records to the NLRD. The current ETQE has, on request of the Director of the NLRD, submitted the missing qualification enrolment records and their related unit standard enrolment records to the NLRD. In these specific cases there is no data in the NLRD that defines a relationship between the qualification, the previous ETQE and the current ETQE. In consultation with the Director of the NLRD it was decided that these types of records need to be assumed as correct for the purposes of this research.

The records that infringe on this semantic business rule are comprised of the following 15 categories:

- No Accreditation (52.21%)

This category indicates that the ETQE that is linked to the unit standard enrolment has never had an active accreditation to quality assure the unit standard and this category is of greatest concern to SAQA.

Of the 29 discrete ETQEs in the dataset, 26 ETQEs are linked to this category. The majority of these records (42.23%) belong to a single ETQE (ETQE Identifier 1126). Of the 9124 discrete unit standards in the dataset, 688 are linked to this category. Most notably, although 249 of the 688 unit standards only constitutes 18.91% of the records; the records for these unit standards represents 100% of the unit standard enrolment records submitted to the NLRD for the unit standard.

- Start Before, End Before (28.11%)

This category indicates that the unit standard enrolment started before and either was achieved or expired before the ETQE was accredited to quality assure the unit standard.

Of the 29 discrete ETQEs in the dataset, 27 ETQEs are linked to this category. Of these records, 90.57% were submitted to the NLRD by 3 ETQEs. Of the 9124 discrete unit standards in the dataset, 1968 are linked to this category. Most notably, although 2 of the 1968 unit standards only constitutes 0.06% of the records; the records for these unit standards represents 100% of the unit standard enrolment records submitted to the NLRD for the unit standard.

- Start Before, End During (12.19%)

This category indicates that the unit standard enrolment started before and either was achieved or expired whilst the ETQE was accredited to quality assure the unit standard.

Of the 29 discrete ETQEs in the dataset, 25 ETQEs are linked to this category. Of these records, 76.79% were submitted to the NLRD by 3 ETQEs. Of the 9124 discrete unit standards in the dataset, 2296 are linked to this category.

- Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End During (3.82%)

This category indicates that the unit standard enrolment:

- started before one of the members of an ETQE amalgamation was accredited to quality assure the unit standard and either was achieved or expired whilst the same ETQE was accredited to quality assure the unit standard, and
- started before the other member of an ETQE amalgamation was accredited to quality assure the unit standard and either was achieved or expired whilst the same ETQE was accredited to quality assure the unit standard.

Of the 29 discrete ETQEs in the dataset, 12 ETQEs are linked to this category. Of these records, 98.59% were submitted to the NLRD by 3 ETQEs. Of the 9124 discrete unit standards in the dataset, 346 are linked to this category.

- Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End Before (3.03%)

This category indicates that the unit standard enrolment:

- started before one of the members of an ETQE amalgamation was accredited to quality assure the unit standard and either was achieved or expired before the same ETQE was accredited to quality assure the unit standard, and
- started before the other member of an ETQE amalgamation was accredited to quality assure the unit standard and either was achieved or expired before the same ETQE was accredited to quality assure the unit standard.

Of the 29 discrete ETQEs in the dataset, 17 ETQEs are linked to this category. Most notably 90.32% of the records were submitted to the NLRD by one ETQE. Of the 9124 discrete unit standards in the dataset, 399 are linked to this category. Most notably, although 5 of the 399 unit standards constitute only 0.02% of the records; the records for these unit standards represent 100% of the unit standard enrolment records submitted to the NLRD for these unit standards.

- Submitting ETQE: No Accreditation, Other ETQE: Start Before, End During (0.41%)  
This category indicates that the unit standard enrolment:
  - started during one of the members of an ETQE amalgamation was not accredited to quality assure the unit standard, and
  - started before the other member of an ETQE amalgamation was accredited to quality assure the unit standard and either was achieved or expired whilst the same ETQE was accredited to quality assure the unit standard.

Of the 29 discrete ETQEs in the dataset, 3 ETQEs are linked to this category. Of these records, 99.88% were submitted to the NLRD by one ETQE. Of the 9124 discrete unit standards in the dataset, 26 are linked to this category.

- Submitting ETQE: Start During, End After, Other ETQE: Start During, End After (0.11%)  
This category indicates that the unit standard enrolment:
  - started during one of the members of an ETQE amalgamation was accredited to quality assure the unit standard and either was achieved or expired whilst the same ETQE was no longer accredited to quality assure the unit standard, and

- started during the other member of an ETQE amalgamation accredited to quality assure the unit standard and either was achieved or expired after the same ETQE was accredited to quality assure the unit standard.

Of the 29 discrete ETQEs in the dataset 2 ETQEs and 34 are linked to this category. Most notably, although 3 of the 34 unit standards constitute only 2.35% of the records; the records for these unit standards represent 100% of the unit standard enrolment records submitted to the NLRD for these unit standards.

- Submitting ETQE: Start During, End After, Other ETQE: Start After, End After (0.05%)

This category indicates that the unit standard enrolment:

- started during one of the members of an ETQE amalgamation was accredited to quality assure the unit standard and either was achieved or expired whilst the same ETQE was no longer accredited to quality assure the unit standard, and
- started after the other member of an ETQE amalgamation was accredited to quality assure the unit standard and either was achieved or expired after the same ETQE was accredited to quality assure the unit standard.

Of the 29 discrete ETQEs in the dataset only one ETQE (ETQE identifier 1125) and 14 are linked to this category.

- Start During, End After (0.04%)

This category indicates that the unit standard enrolment started during and either was achieved or expired after the ETQE was accredited to quality assure the unit standard.

Of the 29 discrete ETQEs in the dataset, 3 ETQEs are linked to this category. The majority of these records (95.45%) belong to a single ETQE (ETQE Identifier 1107). Of the 9124 discrete unit standards in the dataset, 16 are linked to this category. Of these 16 unit standards, 5 unit standards contribute to 95.45% of records in this category.

- Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End After (0.01%)

This category indicates that the unit standard enrolment:

- started during one of the members of an ETQE amalgamation was accredited to quality assure the unit standard and either was achieved or expired whilst the same ETQE was accredited to quality assure the unit standard, and
- started before the other member of an ETQE amalgamation accredited to quality assure the unit standard and either was achieved or expired after the same ETQE was accredited to quality assure the unit standard.

Of the 29 discrete ETQEs in the dataset only one ETQE (ETQE Identifier 1125) and 21 are linked to this category. These records constitute 0.04% of the records submitted to the NLRD by this specific ETQE.

- Submitting ETQE: No Accreditation, Other ETQE: Start Before, End Before (0.01%)

This category indicates that the unit standard enrolment:

- started during one of the members of an ETQE amalgamation was not accredited to quality assure the unit standard, and
- started before the other member of an ETQE amalgamation was accredited to quality assure the unit standard and either was achieved or expired before the same ETQE was accredited to quality assure the unit standard.

Of the 29 discrete ETQEs in the dataset, 5 ETQEs are linked to this category. Of these records, 86.54% were submitted to the NLRD by one ETQE. Of the 9124 discrete unit standards in the dataset, 33 are linked to this category.

- Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End Before (0.01%)

This category indicates that the unit standard enrolment:

- started during one of the members of an ETQE amalgamation was accredited to quality assure the unit standard and either was achieved or expired whilst the same ETQE was accredited to quality assure the unit standard, and

- started before the other member of an ETQE amalgamation accredited to quality assure the unit standard and either was achieved or expired before the same ETQE was accredited to quality assure the unit standard.

Of the 29 discrete ETQEs in the dataset 3 ETQEs and 20 are linked to this category.

- Start After, End After (0.00%)

This category indicates that the unit standard enrolment started after and either was achieved or expired after the ETQE was accredited to quality assure the unit standard.

Of the 29 discrete ETQEs in the dataset, 4 ETQEs are linked to this category. The majority of these records (73.91%) belong to a single ETQE (ETQE Identifier 1107). Of the 9124 discrete unit standards in the dataset, 10 are linked to this category.

- Submitting ETQE: Start After, End After, Other ETQE: Start After, End After (0.00%)

This category indicates that the unit standard enrolment:

- started after one of the members of an ETQE amalgamation was accredited to quality assure the unit standard and either was achieved or expired after the same ETQE was accredited to quality assure the unit standard, and
- started after the other member of an ETQE amalgamation accredited to quality assure the unit standard and either was achieved or expired after the same ETQE was accredited to quality assure the unit standard.

Of the 29 discrete ETQEs in the dataset only one ETQE (ETQE Identifier 1125) and 5 are linked to this category.

- Submitting ETQE: No Accreditation, Other ETQE: Start After, End After (0.00%)

This category indicates that the unit standard enrolment:

- started during one of the members of an ETQE amalgamation was not accredited to quality assure the unit standard, and
- started after the other member of an ETQE amalgamation accredited to quality assure the unit standard and either was achieved or expired after the same ETQE was accredited to quality assure the unit standard.

Of the 29 discrete ETQEs in the dataset only one ETQE (ETQE Identifier 1100) and 9 are linked to this category.

Overall the results for this semantic business rule indicate very few issues exist in regard to whether the ETQE was accredited to quality assure the unit standard for the duration of the learner's active enrolment on the unit standard. As already stated this result could be attributed to the fact that SAQA maintains ETQE related data in the NLRD. The incidence of infringements against this rule is so low and/or so diverse as to suggest that the issues arise as a result of data capturing problems rather than systemic issues with ETQE accreditation records. The generally high incidence of these types of records, when compared against the overall number of records submitted to the NLRD, for ETQE Identifier 1100 and 1104 may however indicate issues of a systemic nature at these two ETQEs.

#### **4.3.3 Conclusion**

This section focuses on the analysis of the nominal data value ETQE\_ACCRED\_IND which contains a value denoting the record's compliance in regard to whether the ETQE was accredited to quality assure the qualification/unit standard for the duration of the learner's active enrolment on the qualification or unit standard.

Overall the results for this semantic business rule indicate very few issues exist in regard to whether the ETQE was accredited to quality assure the qualification/unit standard for the duration of the learner's active enrolment. This result could be attributed to the fact that SAQA maintains ETQE related data in the NLRD. Generally, the incidence of infringements against this rule is so low and/or so diverse as to suggest that the issues arise as a result of data capturing problems rather than systemic issues with ETQE accreditation records.

The analysis of the qualification enrolment records show that very few issues exist in regard to whether the ETQE was accredited for the duration of the learner's active enrolment. The number of qualification enrolment records that are not compliant with this rule is slightly higher than expected. Further investigation shows that, as with ETQE accreditation records (see Section 4.2) this is as a result of missing qualification enrolment

records that were loaded into the NLRD, on request of the Director of the NLRD, after an ETQE amalgamation. These records fail compliance to this rule because the NLRD does not contain the required ETQE accreditation record for the enduring ETQE. Given that unit standard enrolment records are generally linked to qualification enrolment records, this issue was further propagated in unit standard enrolment records which in turn had an even larger number of records that were not compliant with this rule.

Specific recommendations in regard to data records that are not compliant to this semantic business rule are provided in Appendix P.2.1 for qualification enrolments and Appendix P.2.2 for unit standard enrolments.

#### **4.4 Provider accreditation**

This section presents the results of the analysis of learner enrolment records in relation to whether the provider was accredited for the duration of the learner's active enrolment on the learnership, qualification or unit standard. The section therefore focuses on the nominal data value PROV\_IND which contains a value denoting the record's compliance in regard to whether the provider was accredited for the duration of the learner's active enrolment on the learnership, qualification or unit standard.

This section presents the results of the analysis of this data field for learnership enrolment records, qualification enrolment records and unit standard enrolment records.

##### ***4.4.1 Learnership enrolments***

As defined in Appendix C.2, the indicator PROV\_IND denotes whether the provider was accredited for the duration of the learner's active enrolment on the learnership. The manner in which the categories in this indicator is derived is detailed in Appendix C.3.5. An overview of the derived categories, with PROV\_IND\_DESC shown as Description and the frequency of these values as a percentage, is presented in Table 4.4.1.1:

Table 4.4.1.1 Provider accreditation categories for learnership enrolments

Description	% Records
ETQE Provider	6.27%
ETQE Provider Predicted	0.03%
No Accreditation	3.02%
No Accreditation Predicted	0.06%
OK	66.06%
Pre First Submission ETQE Provider	3.51%
Pre First Submission Start After, End After	0.43%
Pre First Submission Start Before, End After	0.02%
Pre First Submission Start Before, End Before	6.70%
Pre First Submission Start Before, End During	2.59%
Pre First Submission Start During, End After	0.30%
Start After, End After	1.20%
Start After, End After Predicted	0.16%
Start Before, End After	0.00%
Start Before, End After Predicted	0.00%
Start Before, End Before	5.41%
Start Before, End During	2.77%
Start Before, End During Predicted	0.04%
Start During, End After	0.59%
Start During, End After Predicted	0.82%
<b>Total</b>	<b>100.00%</b>

An overview of what the text in these categories denotes is as follows:

- ‘ETQE Provider’ denotes a record that has been submitted with an ETQE as the provider (Section 3.8.3.5),
- ‘No Accreditation’ denotes a record where the provider indicated on the learnership enrolment record has never had an active accreditation,
- ‘OK’ indicates that the provider was accredited for the duration of the learner’s active enrolment on the learnership,
- ‘Start After’ indicates that the active time period of the learnership enrolment record started after the provider’s active accreditation time period,
- ‘Start Before’ indicates that the active time period of the learnership enrolment record started before the provider’s active accreditation time period,
- ‘Start During’ indicates that the active time period of the learnership enrolment record started during the provider’s active accreditation time period,
- ‘End After’ indicates that the active time period of the learnership enrolment record ended after the provider’s active accreditation time period,
- ‘End Before’ indicates that the active time period of the learnership enrolment record ended before the provider’s active accreditation time period,

- ‘End During’ indicates that the active time period of the learnership enrolment record ended during the provider’s active accreditation time period,
- ‘Pre First Submission’ indicates a record where the learner enrolled on the learnership prior to the first full data submission from the ETQE to the NLRD (Section 3.8.3.1), and
- ‘Predicted’ indicates a current learnership enrolment record that has not yet been completed and the expected active enrolment on the learnership has not yet expired.

Any record with a category of ‘ETQE provider’ is practically diverting the requirements for accreditation from the provider to the ETQE. A determination of the ETQE’s accreditation time period has already been conducted as part of the analysis of ETQE\_IND (Section 4.2.1). In order to avoid the duplication of these results, records that have this category are assumed to be correct for the purposes of this research.

Any record with a category that starts with the text ‘Pre First Submission’ is a learnership enrolment record with an enrolment date that precedes the date on which the primary ETQE of the provider made its first full data submission to the NLRD. As a result the history of the provider’s active accreditation time period may not be complete (Section 3.8.3.1). As a result of this uncertainty it was decided in consultation with the Director of the NLRD that these types of records will be assumed as correct for the purposes of this research.

Any record with a category that ends with the text ‘Predicted’ is a current learnership enrolment. The data of the learnership enrolment record or the data in the Provider table for these types of records may change before the learnership enrolment record’s active time period expires. As a result of this uncertainty it was decided in consultation with the Director of the NLRD that these types of records will be assumed as correct for the purposes of this research.

As a result the only categories of records that are considered for this research have a description of ‘No Accreditation’, ‘Start After, End After’, ‘Start Before, End After’, ‘Start Before, End Before’, ‘Start Before, End During’ and ‘Start During, End After’. Figure 4.4.1.1 presents an overview of the percentage of records that infringe on the semantic business rule that requires that the provider must be accredited for the duration of the learner’s active enrolment on the learnership.

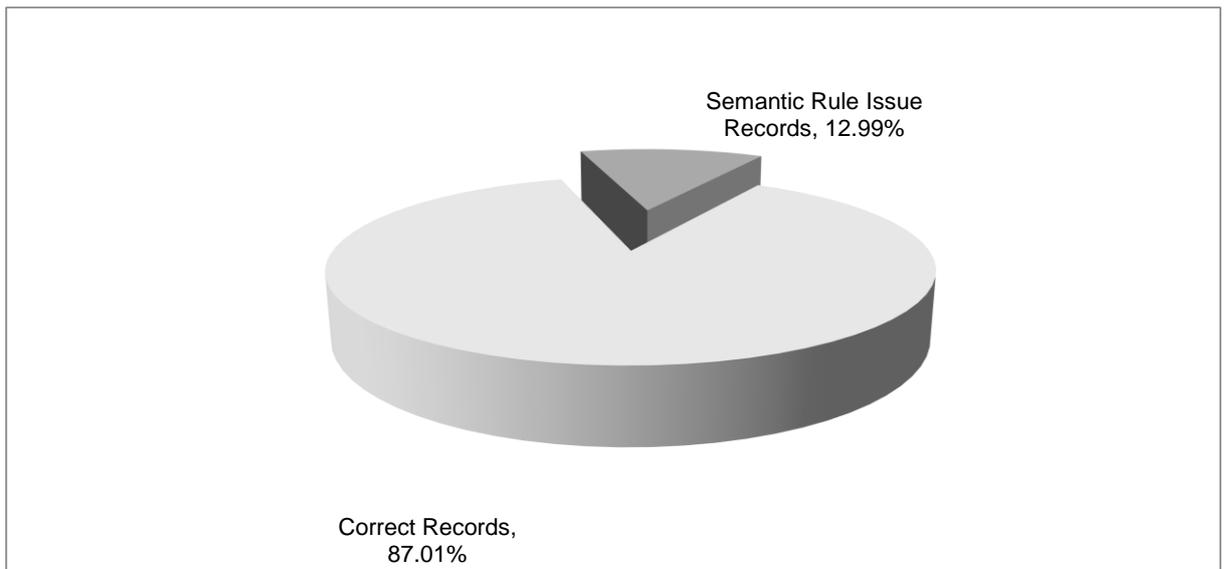


Figure 4.4.1.1 % records according to the semantic business rule that requires that the provider must be accredited for the duration of the learner’s active enrolment on the learnership

The total percentage of records that infringe on this semantic business rule is 12.99%. The reader should note that ETQEs manage data that describe providers on the NLRD (see Section 3.6.3.2.a). The 12.99% records that infringe on this semantic business rule are comprised of 6 categories. Figure 4.4.1.2 provides an overview of the percentage of records found in each of these categories:

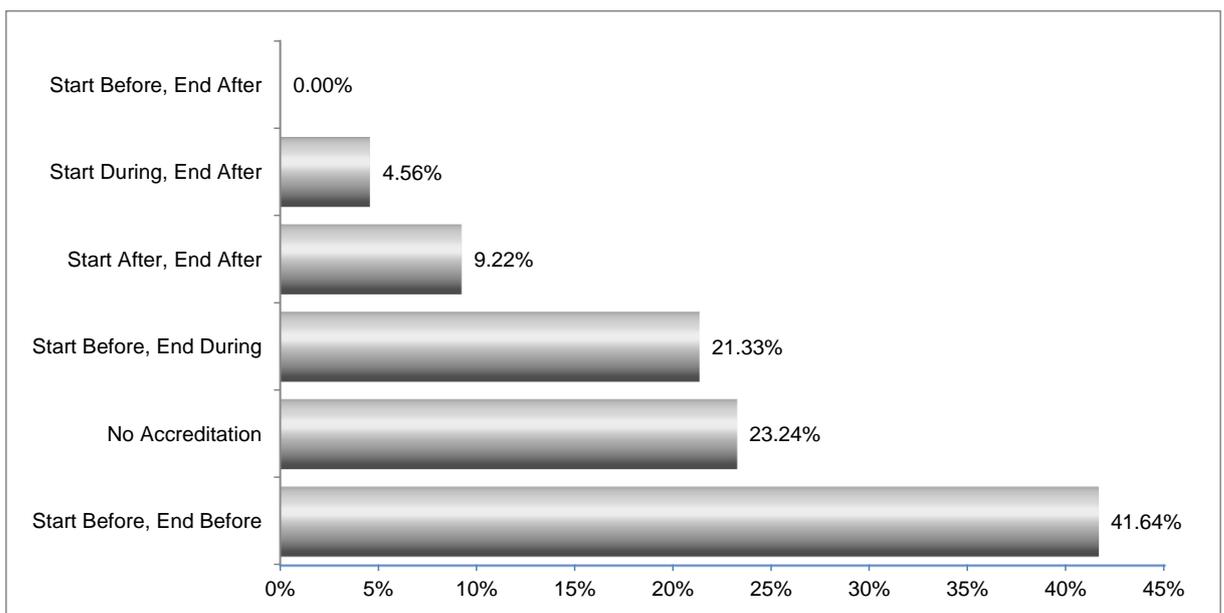


Figure 4.4.1.2 % records that infringe the semantic business rule that requires that the provider must be accredited for the duration of the learner's active enrolment on the learnership by category

The scope and volume of records that infringe on this semantic business rule required a detailed review of each of these categories. This review can be found in Appendix J.1. The analysis of learnership enrolment records in regard to whether the provider was accredited for the duration of the learner's active enrolment highlights the possibility of systemic issues in regard to provider accreditations.

The cluster analysis for the 'Start Before, End Before or End During' and 'Start During, Start After and End After' categories is able to provide a clear description of the data in the categories. Further, a comparison across the two cluster analyses shows that ETQE identifiers 1103, 1115, 1116 and 1126 are featured in both categories. The analysis of the 'No Accreditation' category highlights possible systemic issues in regard to provider accreditations as implemented by ETQE identifiers 1119 and 1115.

The analysis of all three of these categories also show that the utilization of providers that are not accredited by the submitting ETQE has a remarkable impact on adherence in regard to this semantic business rule.

The cluster analysis of both the 'Start Before, End Before or End During' and 'Start During, Start After and End After' categories is able to identify records that may exist in these categories as a result of incorrect data capturing on the learnership enrolment record. The analysis of the 'Start Before, End After' category in turn allows for the identification of a provider record that has possibly been captured incorrectly.

Finally, the summary of semantic infringements by ETQE, which shows the percentage of infringements of this semantic business rule calculated as a percentage of the number of learnership enrolment records submitted to the NLRD by ETQE, shows clear trends of a systemic nature at some ETQEs.

#### **4.4.2 *Qualification enrolments***

As defined in Appendix E.2, the indicator PROV\_IND denotes whether the provider was accredited for the duration of the learner's active enrolment on the qualification. The manner in which the categories in this indicator is derived is detailed in Appendix E.3.6. An overview of the derived categories, with PROV\_IND\_DESC shown as Description and the frequency of these values as a percentage, is presented in Table 4.4.2.1:

Table 4.4.2.1 Provider accreditation categories for qualification enrolments

Description	% Records
ETQE Provider	17.16%
No Accreditation	2.26%
No Accreditation (Qual Linked to Lshp)	0.28%
OK	60.07%
Pre First Submission Start After, End After	0.22%
Pre First Submission Start After, End After (Qual Linked to Lshp)	0.00%
Pre First Submission Start Before, End After	0.01%
Pre First Submission Start Before, End After (Qual Linked to Lshp)	0.04%
Pre First Submission Start Before, End Before	5.83%
Pre First Submission Start Before, End Before (Qual Linked to Lshp)	0.45%
Pre First Submission Start Before, End During	2.81%
Pre First Submission Start Before, End During (Qual Linked to Lshp)	0.51%
Pre First Submission Start During, End After	0.29%
Pre First Submission Start During, End After (Qual Linked to Lshp)	0.04%
Start After, End After	0.79%
Start After, End After (Qual Linked to Lshp)	0.13%
Start After, End After Predicted	0.08%
Start After, End After Predicted (Qual Linked to Lshp)	0.00%
Start Before, End After	0.02%
Start Before, End After (Qual Linked to Lshp)	0.00%
Start Before, End After Predicted	0.01%
Start Before, End Before	4.07%
Start Before, End Before (Qual Linked to Lshp)	0.31%
Start Before, End During	2.87%
Start Before, End During (Qual Linked to Lshp)	0.44%
Start Before, End During Predicted	0.01%
Start During, End After	0.64%
Start During, End After (Qual Linked to Lshp)	0.11%
Start During, End After Predicted	0.55%
Start During, End After Predicted (Qual Linked to Lshp)	0.00%
<b>Total</b>	<b>100.00%</b>

An overview of what the text in these categories denotes is as follows:

- ‘ETQE Provider’ denotes a record that has been submitted with an ETQE as the provider (Section 3.8.3.5),
- ‘No Accreditation’ denotes a record where the provider indicated on the qualification enrolment record has never had an active accreditation,
- ‘OK’ indicates that the provider was accredited for the duration of the learner’s active enrolment on the qualification,
- ‘Start After’ indicates that the active time period of the qualification enrolment record started after the provider’s active accreditation time period,
- ‘Start Before’ indicates that the active time period of the qualification enrolment record started before the provider’s active accreditation time period,
- ‘Start During’ indicates that the active time period of the qualification enrolment record started during the provider’s active accreditation time period,
- ‘End After’ indicates that the active time period of the qualification enrolment record ended after the provider’s active accreditation time period,

- ‘End Before’ indicates that the active time period of the qualification enrolment record ended before the provider’s active accreditation time period,
- ‘End During’ indicates that the active time period of the qualification enrolment record ended during the provider’s active accreditation time period,
- ‘Pre First Submission’ indicates a record where the learner enrolled on the qualification prior to the first full data submission from the ETQE to the NLRD (Section 3.8.3.1),
- ‘Predicted’ indicates a current qualification enrolment record that has not yet been achieved and the expected active enrolment on the qualification has not yet expired, and
- ‘(Qual Linked to Lshp)’ indicates that the qualification is linked to a learnership and as a result the problem may be related to the issue described in Section 3.8.3.7.

Any record with a category that ends with the text ‘(Qual Linked to Lshp)’ may have infringed on this semantic business rule as a result of the contextual issues described in Section 3.8.3.7. It was decided, in consultation with the Director of the NLRD, that the further analysis of these types of records will not form part of this research. The percentage of records for this specific issue is however illustrated for all of the semantic business rules in order to provide an understanding of how the specific issue impacts the data in the NLRD.

Any record with a category of ‘ETQE provider’ is practically diverting the requirements for accreditation from the provider to the ETQE. A determination of the ETQE’s accreditation time period has already been conducted as part of the analysis of ETQE\_IND (Section 4.2.2). In order to avoid the duplication of these results, records that have this category are assumed to be correct for the purposes of this research.

Any record with a category that starts with the text ‘Pre First Submission’ is a qualification enrolment record with an enrolment date that precedes the date on which the primary ETQE of the provider made its first full data submission to the NLRD. As a result the history of the provider’s active accreditation time period may not be complete (Section 3.8.3.1). As a result of this uncertainty it was decided in consultation with the Director of the NLRD that these types of records will be assumed as correct for the purposes of this research.

Any record with a category that ends with the text 'Predicted' is a current qualification enrolment. The data of the qualification enrolment record or the data in the Provider table for these types of records may change before the qualification enrolment record's active time period expires. As a result of this uncertainty it was decided in consultation with the Director of the NLRD that these types of records will be assumed as correct for the purposes of this research.

As a result the only categories of records that are considered for this research have a description of 'No Accreditation', 'Start After, End After', 'Start Before, End After', 'Start Before, End Before', 'Start Before, End During' or 'Start During, End After'. Figure 4.4.2.1 presents an overview of the percentage of records that infringe on the semantic business rule that requires that the provider must be accredited for the duration of the learner's active enrolment on the qualification.

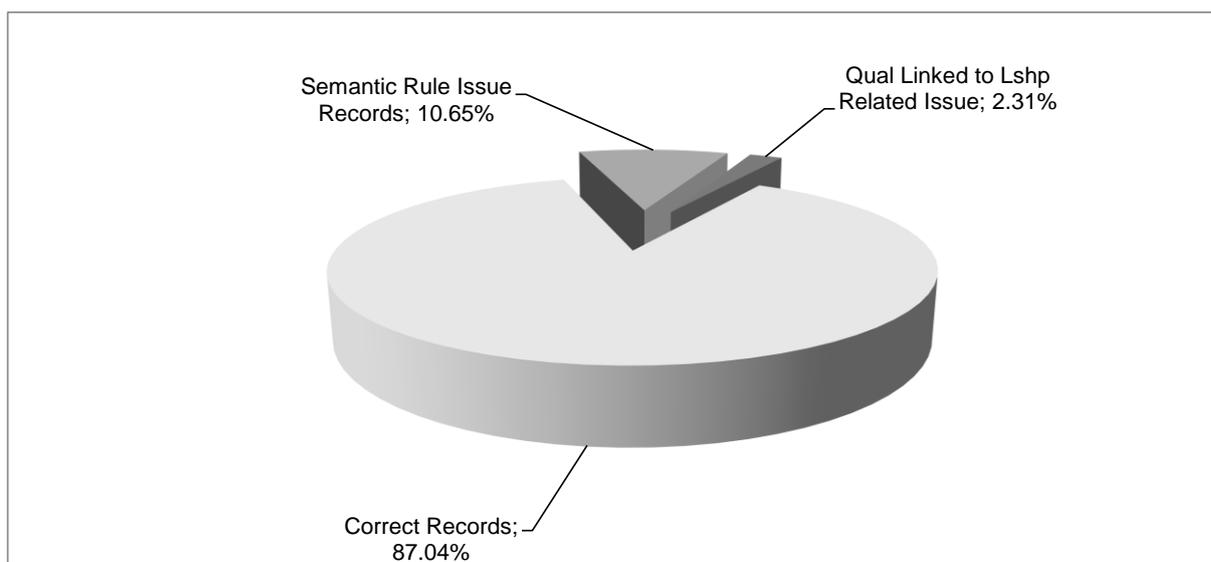


Figure 4.4.2.1 % records according to the semantic business rule that requires that the provider must be accredited for the duration of the learner's active enrolment on the qualification

The total percentage of records that infringe on this semantic business rule is 10.65%. The reader should note that ETQEs manage data that describe providers on the NLRD (see Section 3.6.3.2.a). The 10.65% records that infringe on this semantic business rule are

comprised of 6 categories. Figure 4.4.2.2 provides an overview of the percentage of records found in each of these categories:

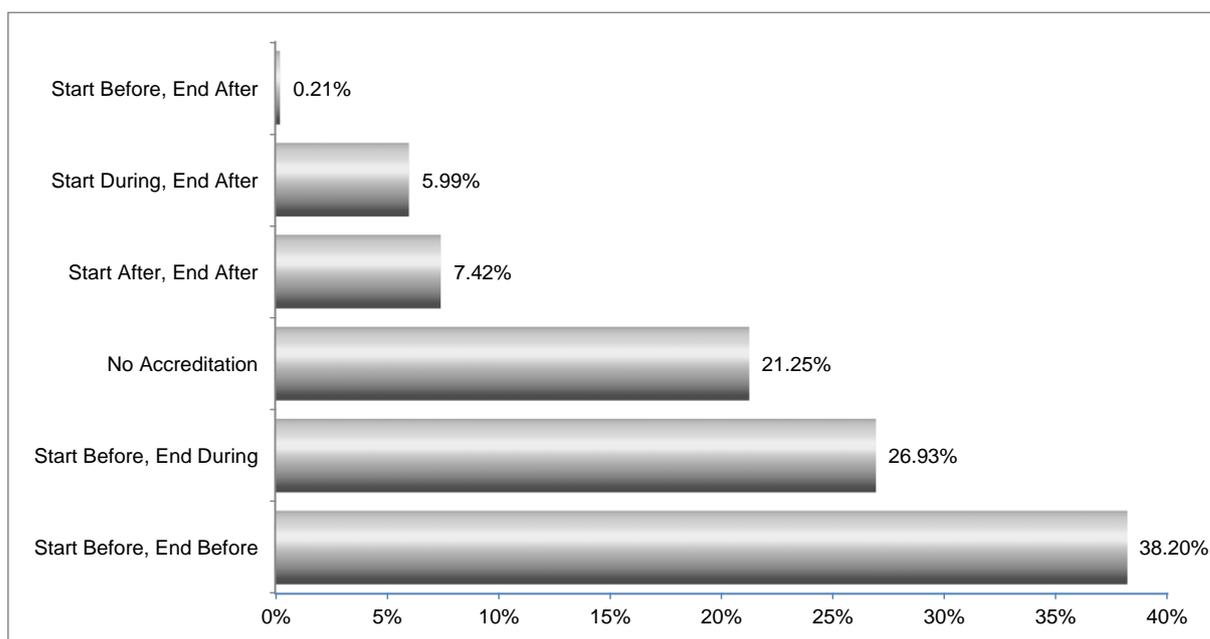


Figure 4.4.2.2 % records that infringe the semantic business rule that requires that the provider must be accredited for the duration of the learner’s active enrolment on the qualification by category

The scope and volume of records that infringe on this semantic business rule require a detailed review of each of these categories. This review can be found in Appendix J.2.

The analysis of qualification enrolment records in regard to whether the provider was accredited for the duration of the learner’s active enrolment highlights the possibility of systemic issues in regard to provider accreditations.

The cluster analysis for the ‘Start Before, End Before or End During’ and ‘Start During, Start After and End After’ categories is able to provide a clear description of the data in the categories. Further, a comparison across the two cluster analyses shows that ETQE identifiers 1105, 1106, 1116 and 1126 are featured in both categories. The analysis of the ‘No Accreditation’ category highlights possible systemic issues in regard to provider accreditations as implemented by ETQE identifiers 1113, 1115 and 1119.

The analysis of the ‘No Accreditation’ and ‘Start Before, End Before or End During’ categories also show that the utilization of providers that are not accredited by the submitting ETQE has a remarkable impact on adherence in regard to this semantic business rule.

The cluster analysis of both the ‘Start Before, End Before or End During’ and ‘Start During, Start After and End After’ categories is able to identify records that may exist in these categories as a result of incorrect data capturing on the qualification enrolment record. The analysis of the ‘Start Before, End After’ category in turn allows for the identification of enrolment records that have possibly been captured incorrectly.

Finally, the summary of semantic infringements by ETQE, which shows the percentage of infringements of this semantic business rule calculated as a percentage of the number of qualification enrolment records submitted to the NLRD by ETQE, shows clear trends of a systemic nature at some ETQEs.

#### ***4.4.3 Unit Standard enrolments***

As defined in Appendix G.2, the indicator PROV\_IND denotes whether the provider was accredited for the duration of the learner’s active enrolment on the unit standard. The manner in which the categories in this indicator is derived is detailed in Appendix G.3.6. An overview of the derived categories, with PROV\_IND\_DESC shown as Description and the frequency of these values as a percentage, is presented in Table 4.4.3.1:

Table 4.4.3.1 Provider accreditation categories for unit standard enrolments

Description	% Records
ETQE Provider	8.51%
No Accreditation	1.52%
No Accreditation (UStd Linked to Lshp)	0.02%
OK	68.76%
Pre First Submission Start After, End After	0.04%
Pre First Submission Start After, End After (UStd Linked to Lshp)	0.00%
Pre First Submission Start Before, End After	0.06%
Pre First Submission Start Before, End After (UStd Linked to Lshp)	0.00%
Pre First Submission Start Before, End Before	6.76%
Pre First Submission Start Before, End Before (UStd Linked to Lshp)	0.07%
Pre First Submission Start Before, End During	2.26%
Pre First Submission Start Before, End During (UStd Linked to Lshp)	0.03%
Pre First Submission Start During, End After	0.77%
Pre First Submission Start During, End After (UStd Linked to Lshp)	0.02%
Start After, End After	1.59%
Start After, End After (UStd Linked to Lshp)	0.01%
Start Before, End After	0.02%
Start Before, End After (UStd Linked to Lshp)	0.00%
Start Before, End Before	5.76%
Start Before, End Before (UStd Linked to Lshp)	0.05%
Start Before, End During	2.85%
Start Before, End During (UStd Linked to Lshp)	0.02%
Start During, End After	0.85%
Start During, End After (UStd Linked to Lshp)	0.04%
<b>Total</b>	<b>100.00%</b>

An overview of what the text in these categories denotes is as follows:

- ‘ETQE Provider’ denotes a record that has been submitted with an ETQE as the provider (Section 3.8.3.5),
- ‘No Accreditation’ denotes a record where the provider indicated on the unit standard enrolment record has never had an active accreditation,
- ‘OK’ indicates that the provider was accredited for the duration of the learner’s active enrolment on the unit standard,
- ‘Start After’ indicates that the active time period of the unit standard enrolment record started after the provider’s active accreditation time period,
- ‘Start Before’ indicates that the active time period of the unit standard enrolment record started before the provider’s active accreditation time period,
- ‘Start During’ indicates that the active time period of the unit standard enrolment record started during the provider’s active accreditation time period,
- ‘End After’ indicates that the active time period of the unit standard enrolment record ended after the provider’s active accreditation time period,
- ‘End Before’ indicates that the active time period of the unit standard enrolment record ended before the provider’s active accreditation time period,

- ‘End During’ indicates that the active time period of the unit standard enrolment record ended during the provider’s active accreditation time period,
- ‘Pre First Submission’ indicates a record where the learner enrolled on the unit standard prior to the first full data submission from the ETQE to the NLRD (Section 3.8.3.1),
- ‘Predicted’ indicates a current unit standard enrolment record that has not yet been achieved and the expected active enrolment on the unit standard has not yet expired, and
- ‘(UStd Linked to Lshp)’ indicates that the unit standard is linked to a learnership and as a result the problem may be related to the issue described in Section 3.8.3.7.

Any record with a category that ends with the text ‘(UStd Linked to Lshp)’ may have infringed on this semantic business rule as a result of the contextual issues described in Section 3.8.3.7. It was decided, in consultation with the Director of the NLRD, that the further analysis of these types of records will not form part of this research. The percentage of records for this specific issue is however illustrated for all of the semantic business rules in order to provide an understanding of how the specific issue impacts the data in the NLRD.

Any record with a category of ‘ETQE provider’ is practically diverting the requirements for accreditation from the provider to the ETQE. A determination of the ETQE’s accreditation time period has already been conducted as part of the analysis of ETQE\_IND (Section 4.2.2). In order to avoid the duplication of these results, records that have this category are assumed to be correct for the purposes of this research.

Any record with a category that starts with the text ‘Pre First Submission’ is a unit standard enrolment record with an enrolment date that precedes the date on which the primary ETQE of the provider made its first full data submission to the NLRD. As a result the history of the provider’s active accreditation time period may not be complete (Section 3.8.3.1). As a result of this uncertainty it was decided in consultation with the Director of the NLRD that these types of records will be assumed as correct for the purposes of this research.

Any record with a category that ends with the text ‘Predicted’ is a current unit standard enrolment. The data of the unit standard enrolment record or the data in the Provider table for these types of records may change before the unit standard enrolment record’s active

time period expires. As a result of this uncertainty it was decided in consultation with the Director of the NLRD that these types of records will be assumed as correct for the purposes of this research.

As a result the only categories of records that are considered for this research have a description of 'Start Before, End Before', 'Start Before, End During', 'Start After, End After', 'No Accreditation', 'Start During, End After' or 'Start Before, End After'. Figure 4.4.3.1 presents an overview of the percentage of records that infringe on the semantic business rule that requires that the provider must be accredited for the duration of the learner's active enrolment on the unit standard.

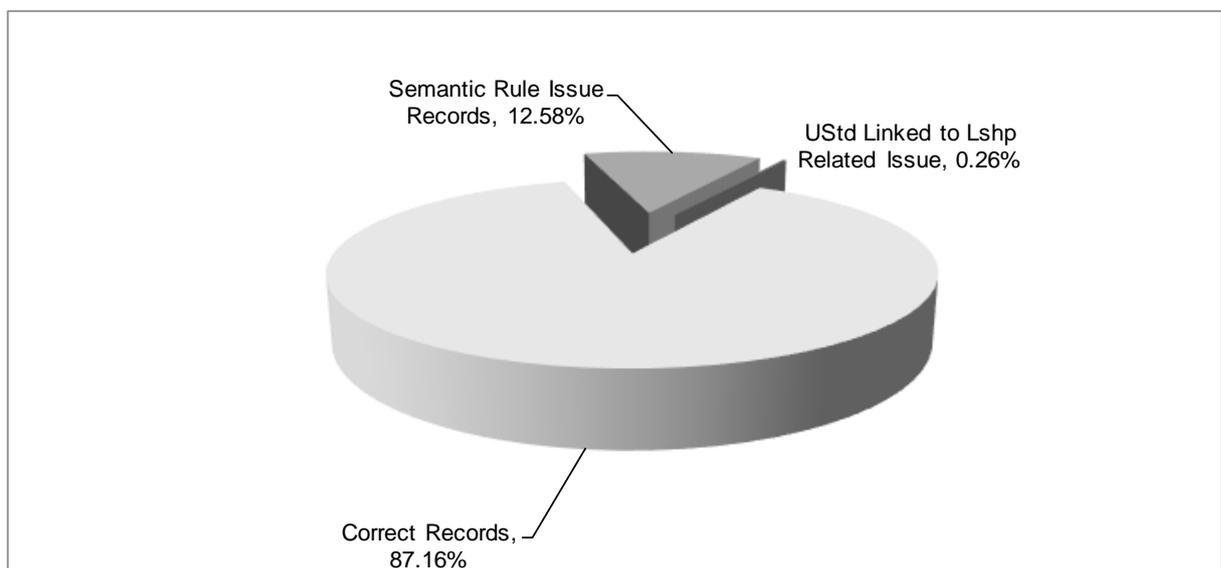


Figure 4.4.3.1 % records according to the semantic business rule that requires that the provider must be accredited for the duration of the learner's active enrolment on the unit standard

The total percentage of records that infringe on this semantic business rule is 12.58%. The reader should note that ETQEs manage data that describe providers on the NLRD (see Section 3.6.3.2.a). The 12.58% records that infringe on this semantic business rule are comprised of 6 categories. Figure 4.4.3.2 provides an overview of the percentage of records found in each of these categories:

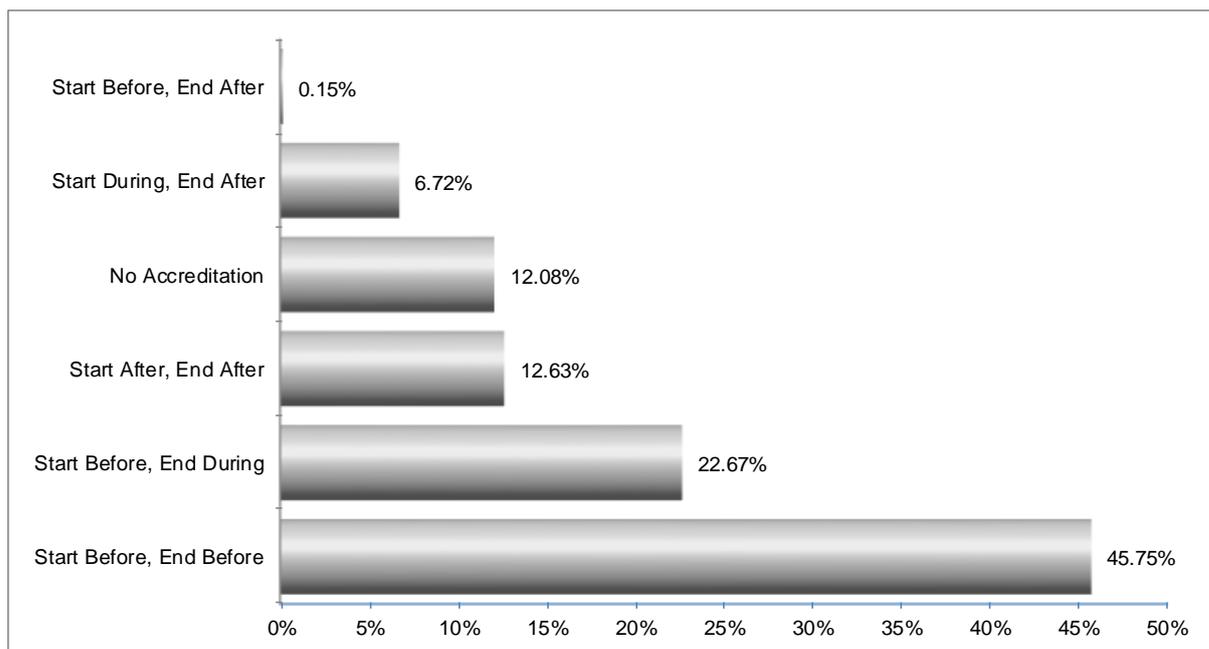


Figure 4.4.3.2 % records that infringe the semantic business rule that requires that the provider must be accredited for the duration of the learner’s active enrolment on the unit standard by category

The scope and volume of records that infringe on this semantic business rule require a detailed review of each of these categories. This review can be found in Appendix J.3.

The analysis of unit standard enrolment records in regard to whether the provider was accredited for the duration of the learner’s active enrolment highlights the possibility of systemic issues in regard to provider accreditations.

The cluster analysis for the ‘Start Before, End Before or End During’ and ‘Start During, Start After and End After’ categories is able to provide a clear description of the data in the categories. Further, a comparison across the two cluster analyses shows that 23 ETQEs are featured in both categories. The cluster analysis of both the ‘Start Before, End Before or End During’ and ‘Start During, Start After and End After’ categories is able to identify records that may exist in these categories as a result of incorrect data capturing on the unit standard enrolment record.

The analysis of the ‘No Accreditation’ category highlights possible systemic issues in regard to provider accreditations as implemented by ETQE identifiers 1105 and 1119. The

analysis of the 'No Accreditation' category also shows that the utilization of providers that are not accredited by the submitting ETQE has a remarkable impact on adherence in regard to this semantic business rule.

Finally, the summary of semantic infringements by ETQE, which shows the percentage of infringements of this semantic business rule calculated as a percentage of the number of unit standard enrolment records submitted to the NLRD by ETQE, shows clear trends of a systemic nature at some ETQEs.

#### **4.4.4 Conclusion**

This section focuses on the analysis of the nominal data value PROV\_IND which contains a value denoting the record's compliance in regard to whether the provider was accredited for the duration of the learner's active enrolment on the learnership, qualification or unit standard.

Overall the results for this semantic business rule highlights the possibility of systemic issues in regard to provider accreditations, with 12.99% learnership, 10.65% qualification and 12.64% unit standard enrolment records infringing on this semantic business rule.

The summary of semantic infringements by ETQE (see Appendix J.1.9, J.2.9 and J.3.9) which provides an overview of the percentage of records, calculated as a percentage of the number of records submitted by the ETQE, which infringe on this semantic business rule gives a clear overview of the ETQEs that most frequently infringe on this semantic business rule:

- ETQE Identifier 1116 shows notable infringements for all three types of enrolments (learnership, qualification and unit standard enrolments),
- ETQE Identifiers 1115, 1110, 1105 and 1075 show notable infringements for two of the three types of enrolments (learnership, qualification and unit standard enrolments), and
- ETQE Identifier 1100 shows pronounced infringements for unit standard enrolments.

For all three types of enrolments it is found that the utilization of providers that are not accredited by the submitting ETQE has a remarkable impact on adherence in regard to this semantic business rule. Of the records that infringe on this semantic business rule, 40.19% learnership, 38.43% qualification and 33.42% unit standard enrolment records were offered

by a provider that was not accredited by the submitting ETQE. This suggests that no significant mechanism exists that facilitates the exchange of information in regard to providers between ETQEs.

Specific recommendations in regard to data records that are not compliant to this semantic business rule are provided in Appendix P.3.1 for learnership enrolments, Appendix P.3.2 for qualification enrolments and P.3.3 for unit standard enrolments.

#### **4.5 Provider accreditation to offer the qualification or unit standard**

This section presents the results of the analysis of learner enrolment records in relation to whether the provider was accredited to offer the qualification or unit standard for the duration of the learner's active enrolment on the qualification or unit standard. The section therefore focuses on the nominal data value PROV\_ACCRED\_IND which contains a value denoting the record's compliance in regard to whether the provider was accredited to offer the qualification or unit standard for the duration of the learner's active enrolment on the qualification or unit standard.

This section presents the results of the analysis of this data field for qualification enrolment records and unit standard enrolment records.

##### ***4.5.1 Qualification enrolments***

As defined in Appendix E.2, the indicator PROV\_ACCRED\_IND denotes whether the provider was accredited to offer the qualification for the duration of the learner's active enrolment on the qualification. The manner in which the categories in this indicator is derived is detailed in Appendix E.3.7. An overview of the derived categories, with PROV\_IND\_DESC shown as Description and the frequency of these values as a percentage, is presented in Table 4.5.1.1:

Table 4.5.1.1 Provider accreditation to offer the qualification categories

Description	% Records
ETQE Provider	16.70%
ETQE Provider (Qual Linked to Lshp)	0.46%
No Accreditation	8.33%
No Accreditation (Qual Linked to Lshp)	0.78%
No Accreditation Predicted	0.09%
No Accreditation Predicted (Qual Linked to Lshp)	0.00%
OK	45.16%
Pre First Submission No Accreditation	3.54%
Pre First Submission No Accreditation (Qual Linked to Lshp)	0.34%
Pre First Submission Start After, End After	0.17%
Pre First Submission Start After, End After (Qual Linked to Lshp)	0.00%
Pre First Submission Start Before, End After	0.07%
Pre First Submission Start Before, End After (Qual Linked to Lshp)	0.02%
Pre First Submission Start Before, End Before	6.24%
Pre First Submission Start Before, End Before (Qual Linked to Lshp)	0.61%
Pre First Submission Start Before, End During	3.08%
Pre First Submission Start Before, End During (Qual Linked to Lshp)	0.54%
Pre First Submission Start During, End After	0.72%
Pre First Submission Start During, End After (Qual Linked to Lshp)	0.07%
Start After, End After	0.74%
Start After, End After (Qual Linked to Lshp)	0.19%
Start After, End After Predicted	0.08%
Start After, End After Predicted (Qual Linked to Lshp)	0.00%
Start Before, End After	0.02%
Start Before, End After (Qual Linked to Lshp)	0.00%
Start Before, End After Predicted	0.01%
Start Before, End Before	4.19%
Start Before, End Before (Qual Linked to Lshp)	0.31%
Start Before, End During	4.75%
Start Before, End During (Qual Linked to Lshp)	0.30%
Start Before, End During Predicted	0.02%
Start During, End After	1.67%
Start During, End After (Qual Linked to Lshp)	0.06%
Start During, End After Predicted	0.72%
Start During, End After Predicted (Qual Linked to Lshp)	0.00%
<b>Total</b>	<b>100.00%</b>

An overview of what the text in these categories denotes is as follows:

- ‘ETQE Provider’ denotes a record that has been submitted with an ETQE as the provider (Section 3.8.3.5),
- ‘No Accreditation’ denotes a record where the provider indicated on the qualification enrolment record has never had an active accreditation to offer the qualification,
- ‘OK’ indicates that the provider was accredited to offer the qualification for the duration of the learner’s active enrolment on the qualification,
- ‘Start After’ indicates that the active time period of the qualification enrolment record started after the provider’s active accreditation to offer the qualification time period,
- ‘Start Before’ indicates that the active time period of the qualification enrolment record started before the provider’s active accreditation to offer the qualification time period,

- ‘Start During’ indicates that the active time period of the qualification enrolment record started during the provider’s active accreditation to offer the qualification time period,
- ‘End After’ indicates that the active time period of the qualification enrolment record ended after the provider’s active accreditation to offer the qualification time period,
- ‘End Before’ indicates that the active time period of the qualification enrolment record ended before the provider’s active accreditation to offer the qualification time period,
- ‘End During’ indicates that the active time period of the qualification enrolment record ended during the provider’s active accreditation to offer the qualification time period,
- ‘Pre First Submission’ indicates a record where the learner enrolled on the qualification prior to the first full data submission from the ETQE to the NLRD (Section 3.8.3.1),
- ‘Predicted’ indicates a current qualification enrolment record that has not yet been achieved and the expected active enrolment on the qualification has not yet expired, and
- ‘(Qual Linked to Lshp)’ indicates that the qualification is linked to a learnership and as a result the problem may be related to the issue described in Section 3.8.3.7.

Any record with a category that ends with the text ‘(Qual Linked to Lshp)’ may have infringed on this semantic business rule as a result of the contextual issues described in Section 3.8.3.7. It was decided, in consultation with the Director of the NLRD, that the further analysis of these types of records will not form part of this research. The percentage of records for this specific issue is however illustrated for all of the semantic business rules in order to provide an understanding of how the specific issue impacts the data in the NLRD.

Any record with a category of ‘ETQE provider’ is practically diverting the requirements for accreditation to offer the qualification from the provider to the ETQE. A determination of the ETQE’s accreditation time period has already been conducted as part of the analysis of ETQE\_IND (Section 4.2.2). In order to avoid the duplication of these results, records that have this category are assumed to be correct for the purposes of this research.

Any record with a category that starts with the text ‘Pre First Submission’ is a qualification enrolment record with an enrolment date that precedes the date on which the ETQE that submitted the enrolment record made its first full data submission to the NLRD. As a result the history of the provider’s active accreditation to offer the qualification time period may not be complete (Section 3.8.3.1). As a result of this uncertainty it was decided in consultation with the Director of the NLRD that these types of records will be assumed as correct for the purposes of this research.

Any record with a category that ends with the text ‘Predicted’ is a current qualification enrolment. The data of the qualification enrolment record or the data in the Provider Accreditation table for these types of records may change before the qualification enrolment record’s active time period expires. As a result of this uncertainty it was decided in consultation with the Director of the NLRD that these types of records will be assumed as correct for the purposes of this research.

As a result the only categories of records that are considered for this research have a description of ‘No Accreditation’, ‘Start After, End After’, ‘Start Before, End After’, ‘Start Before, End Before’, ‘Start Before, End During’ and ‘Start During, End After’. Figure 4.5.1.1 presents an overview of the percentage of records that infringe on the semantic business rule that requires that the provider must be accredited for the duration of the learner’s active enrolment on the qualification.

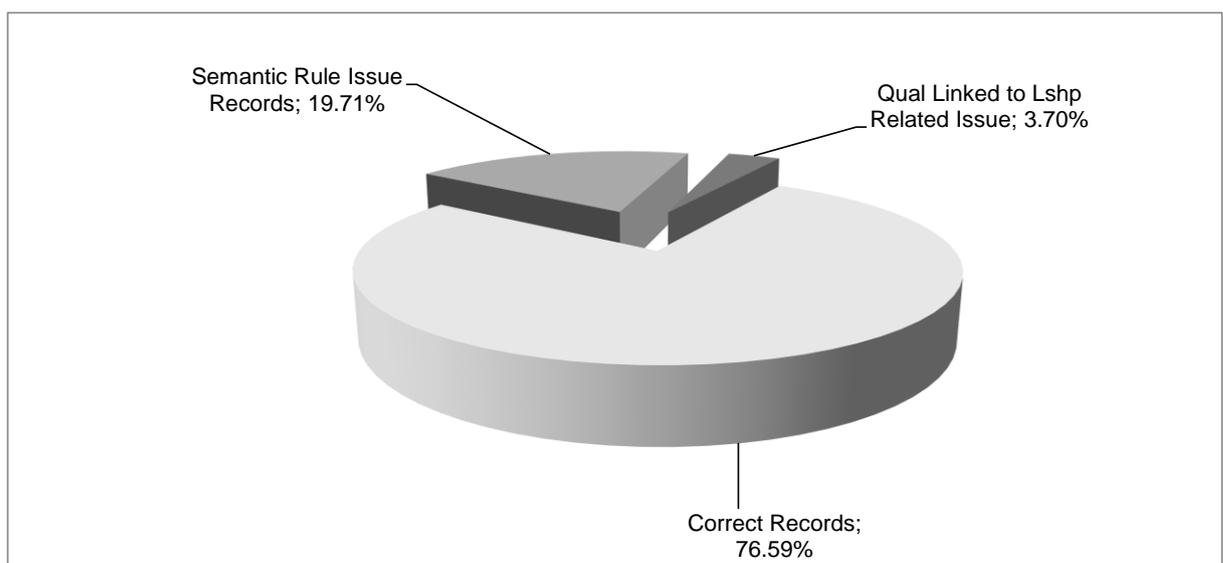


Figure 4.5.1.1 % records according to the semantic business rule that requires that the provider must be accredited to offer the qualification for the duration of the learner's active enrolment on the qualification

The total percentage of records that infringe on this semantic business rule is 19.71%. The reader should note that ETQEs manage data that describe providers on the NLRD (see Section 3.6.3.2.a). The 19.71% records that infringe on this semantic business rule are comprised of 6 categories. Figure 4.5.1.2 provides an overview of the percentage of records found in each of these categories:

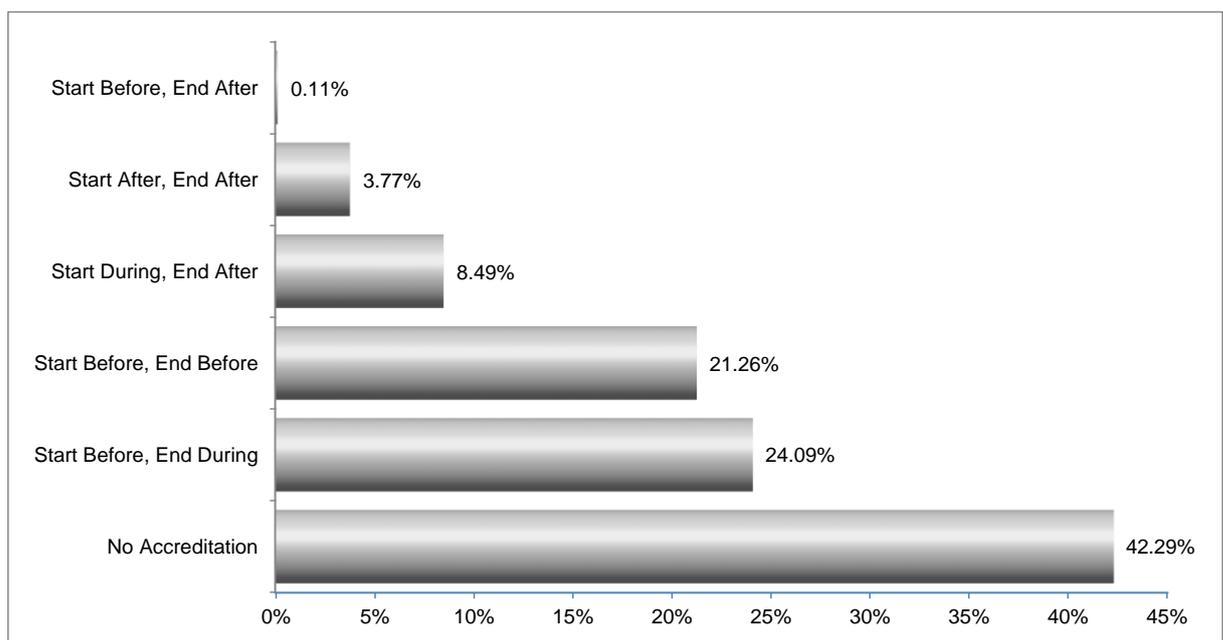


Figure 4.5.1.2 % records that infringe the semantic business rule that requires that the provider must be accredited for the duration of the learner's active enrolment on the qualification by category

The scope and volume of records that infringe on this semantic business rule require a detailed review of each of these categories. This review can be found in Appendix L.1.

The analysis of qualification enrolment records in regard to whether the provider was accredited to offer the qualification for the duration of the learner's active enrolment on the qualification highlights the possibility of systemic issues in regard to provider accreditations.

The analysis of the ‘No Accreditation’ category highlights possible systemic issues in regard to provider accreditations as implemented by ETQE identifiers 1116, 1126 and 1103. The cluster analysis for the ‘Start Before, End Before or End During’ and ‘Start During, Start After and End After’ categories is able to provide a clear description of the data in the categories. Further, a comparison across the two cluster analyses shows that ETQE identifiers 1105, 1106, 1116 and 1126 are featured in both categories.

The analysis of the ‘No Accreditation’ category also shows that the utilization of providers that are not accredited by the submitting ETQE has a remarkable impact on adherence in regard to this semantic business rule.

The cluster analysis of both the ‘Start Before, End Before or End During’ and ‘Start During, Start After and End After’ categories is able to identify records that may exist in these categories as a result of incorrect data capturing on the qualification enrolment record. The analysis of the ‘Start Before, End After’ category in turn allows for the identification of enrolment records that have possibly been captured incorrectly.

Finally, the summary of semantic infringements by ETQE, which shows the percentage of infringements of this semantic business rule calculated as a percentage of the number of qualification enrolment records submitted to the NLRD by ETQE, shows clear trends of a systemic nature at some ETQEs.

#### ***4.5.2 Unit Standard enrolments***

As defined in Appendix G.2, the indicator PROV\_ACCRED\_IND denotes whether the provider was accredited to offer the unit standard for the duration of the learner’s active enrolment on the unit standard. The manner in which the categories in this indicator is derived is detailed in Appendix G.3.7. An overview of the derived categories, with PROV\_IND\_DESC shown as Description and the frequency of these values as a percentage, is presented in Table 4.5.2.1:

Table 4.5.2.1 Provider accreditation to offer the unit standard categories

Description	% Records
ETQE Provider	8.46%
ETQE Provider (UStd Linked to Lshp)	0.06%
No Accreditation	18.74%
No Accreditation (UStd Linked to Lshp)	0.10%
OK	47.33%
Pre First Submission No Accreditation	5.13%
Pre First Submission No Accreditation (UStd Linked to Lshp)	0.07%
Pre First Submission Start After, End After	0.06%
Pre First Submission Start After, End After (UStd Linked to Lshp)	0.00%
Pre First Submission Start Before, End After	0.11%
Pre First Submission Start Before, End After (UStd Linked to Lshp)	0.00%
Pre First Submission Start Before, End Before	5.97%
Pre First Submission Start Before, End Before (UStd Linked to Lshp)	0.08%
Pre First Submission Start Before, End During	2.23%
Pre First Submission Start Before, End During (UStd Linked to Lshp)	0.05%
Pre First Submission Start During, End After	0.36%
Pre First Submission Start During, End After (UStd Linked to Lshp)	0.01%
Start After, End After	1.84%
Start After, End After (UStd Linked to Lshp)	0.02%
Start Before, End After	0.08%
Start Before, End After (UStd Linked to Lshp)	0.00%
Start Before, End Before	4.86%
Start Before, End Before (UStd Linked to Lshp)	0.03%
Start Before, End During	3.14%
Start Before, End During (UStd Linked to Lshp)	0.03%
Start During, End After	1.19%
Start During, End After (UStd Linked to Lshp)	0.02%
<b>Total</b>	<b>100.00%</b>

An overview of what the text in these categories denotes is as follows:

- ‘ETQE Provider’ denotes a record that has been submitted with an ETQE as the provider (Section 3.8.3.5),
- ‘No Accreditation’ denotes a record where the provider indicated on the unit standard enrolment record has never had an active accreditation to offer the unit standard,
- ‘OK’ indicates that the provider was accredited to offer the unit standard for the duration of the learner’s active enrolment on the unit standard,
- ‘Start After’ indicates that the active time period of the unit standard enrolment record started after the provider’s active accreditation to offer the unit standard time period,
- ‘Start Before’ indicates that the active time period of the unit standard enrolment record started before the provider’s active accreditation to offer the unit standard time period,
- ‘Start During’ indicates that the active time period of the unit standard enrolment record started during the provider’s active accreditation to offer the unit standard time period,

- ‘End After’ indicates that the active time period of the unit standard enrolment record ended after the provider’s active accreditation to offer the unit standard time period,
- ‘End Before’ indicates that the active time period of the unit standard enrolment record ended before the provider’s active accreditation to offer the unit standard time period,
- ‘End During’ indicates that the active time period of the unit standard enrolment record ended during the provider’s active accreditation to offer the unit standard time period,
- ‘Pre First Submission’ indicates a record where the learner enrolled on the unit standard prior to the first full data submission from the ETQE to the NLRD (Section 3.8.3.1),
- ‘Predicted’ indicates a current unit standard enrolment record that has not yet been achieved and the expected active enrolment on the unit standard has not yet expired, and
- ‘(UStd Linked to Lshp)’ indicates that the unit standard is linked to a learnership and as a result the problem may be related to the issue described in Section 3.8.3.7.

Any record with a category that ends with the text ‘(UStd Linked to Lshp)’ may have infringed on this semantic business rule as a result of the contextual issues described in Section 3.8.3.7. It was decided, in consultation with the Director of the NLRD, that the further analysis of these types of records will not form part of this research. The percentage of records for this specific issue is however illustrated for all of the semantic business rules in order to provide an understanding of how the specific issue impacts the data in the NLRD.

Any record with a category of ‘ETQE provider’ is practically diverting the requirements for accreditation to offer the unit standard from the provider to the ETQE. A determination of the ETQE’s accreditation time period has already been conducted as part of the analysis of ETQE\_IND (Section 4.2.3). In order to avoid the duplication of these results, records that have this category are assumed to be correct for the purposes of this research.

Any record with a category that starts with the text ‘Pre First Submission’ is a unit standard enrolment record with an enrolment date that precedes the date on which the ETQE that submitted the enrolment record made its first full data submission to the NLRD. As a result the history of the provider’s active accreditation to offer the unit standard time period may

not be complete (Section 3.8.3.1). As a result of this uncertainty it was decided in consultation with the Director of the NLRD that these types of records will be assumed as correct for the purposes of this research.

Any record with a category that ends with the text 'Predicted' is a current unit standard enrolment. The data of the unit standard enrolment record or the data in the Provider Accreditation table for these types of records may change before the unit standard enrolment record's active time period expires. As a result of this uncertainty it was decided in consultation with the Director of the NLRD that these types of records will be assumed as correct for the purposes of this research.

As a result the only categories of records that are considered for this research have a description of 'No Accreditation', 'Start After, End After', 'Start Before, End After', 'Start Before, End Before', 'Start Before, End During' and 'Start During, End After'. Figure 4.5.2.1 presents an overview of the percentage of records that infringe on the semantic business rule that requires that the provider must be accredited for the duration of the learner's active enrolment on the unit standard.

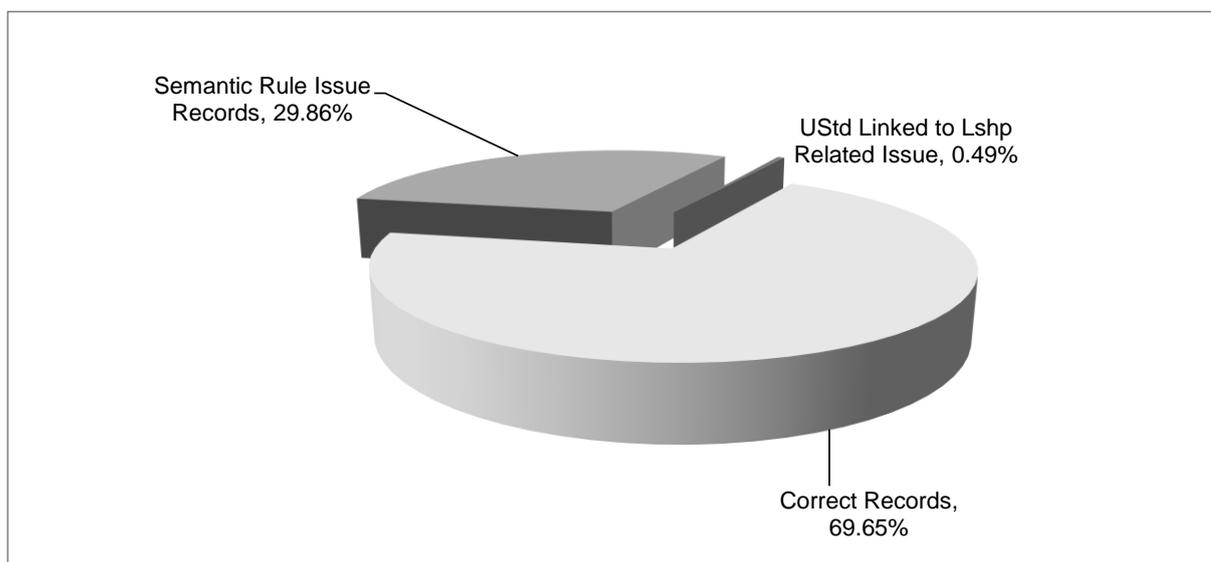


Figure 4.5.2.1 % records according to the semantic business rule that requires that the provider must be accredited to offer the unit standard for the duration of the learner's active enrolment on the unit standard

The total percentage of records that infringe on this semantic business rule is 29.86%. The reader should note that ETQEs manage data that describe providers on the NLRD (see Section 3.6.3.2.a). The 29.86% records that infringe on this semantic business rule are comprised of 6 categories. Figure 4.5.2.2 provides an overview of the percentage of records found in each of these categories:

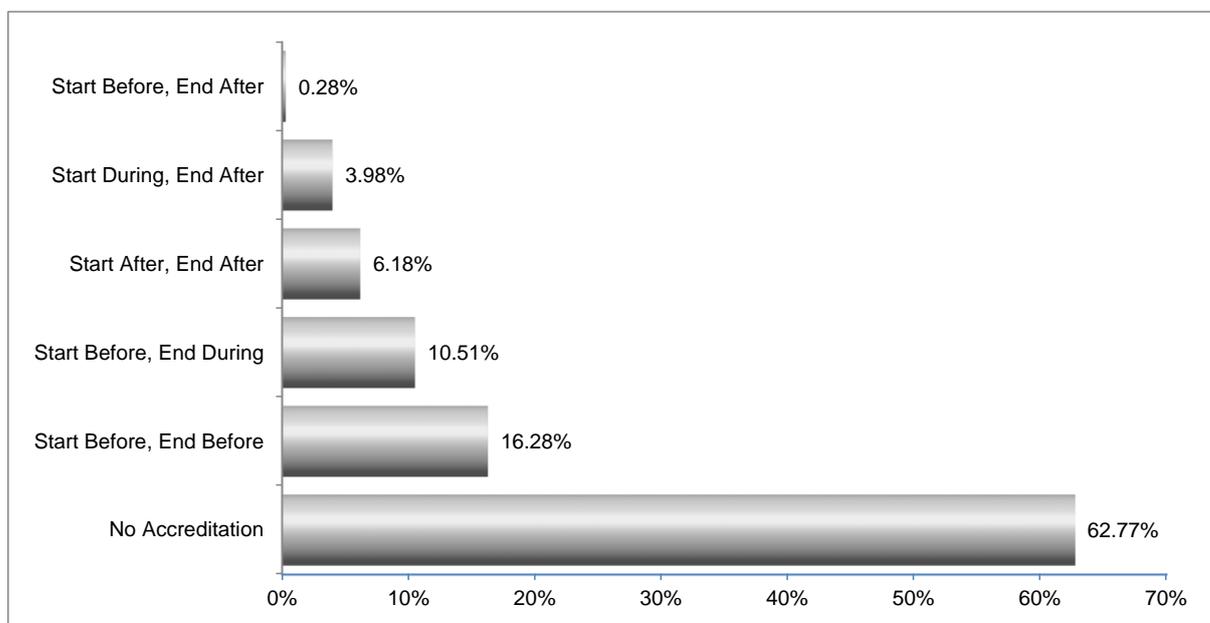


Figure 4.5.2.2 % records that infringe the semantic business rule that requires that the provider must be accredited for the duration of the learner’s active enrolment on the unit standard by category

The scope and volume of records that infringe on this semantic business rule require a detailed review of each of these categories. This review can be found in Appendix L.2.

The analysis of unit standard enrolment records in regard to whether the provider was accredited to offer the unit standard for the duration of the learner’s active enrolment on the unit standard highlights the possibility of systemic issues in regard to provider accreditations.

The analysis of the ‘No Accreditation’ category highlights possible systemic issues in regard to provider accreditations as implemented by ETQE identifiers 1116, 1126 and 1103. The cluster analysis for the ‘Start Before, End Before or End During’ and ‘Start During, Start After and End After’ categories is able to provide a clear description of the data in the

categories. Further, a comparison across the two cluster analyses shows that ETQE identifiers 1105, 1106, 1116 and 1126 are featured in both categories.

The analysis of the 'No Accreditation' category also shows that the utilization of providers that are not accredited by the submitting ETQE has a remarkable impact on adherence in regard to this semantic business rule.

The cluster analysis of both the 'Start Before, End Before or End During' and 'Start During, Start After and End After' categories is able to identify records that may exist in these categories as a result of incorrect data capturing on the unit standard enrolment record. The analysis of the 'Start Before, End After' category in turn allows for the identification of enrolment records that have possibly been captured incorrectly.

Finally, the summary of semantic infringements by ETQE, which shows the percentage of infringements of this semantic business rule calculated as a percentage of the number of unit standard enrolment records submitted to the NLRD by ETQE, shows clear trends of a systemic nature at some ETQEs.

#### **4.5.3 Conclusion**

This section focuses on the analysis of the nominal data value PROV\_ACCRED\_IND which contains a value denoting the record's compliance in regard to whether the provider was accredited to offer the qualification or unit standard for the duration of the learner's active enrolment on the qualification or unit standard.

Overall the results for this semantic business rule highlights the possibility of systemic issues in regard to provider accreditations to offer the qualification or unit standard with 19.71% qualification and 29.86% unit standard enrolment records infringing on this semantic business rule.

The summary of semantic infringements by ETQE (see Appendix L.1.9 and L.2.9) which provides an overview of the percentage of records, calculated as a percentage of the number of records submitted by the ETQE, which infringe on this semantic business rule gives a clear overview of the ETQEs that most frequently infringe on this semantic business rule:

- ETQE Identifier 1116 has the highest rate of infringements for both types of enrolments (qualification and unit standard enrolments),
- ETQE Identifiers 1114, 1116 and 1075 show notable infringements for both types of enrolments (qualification and unit standard enrolments), and
- ETQE Identifier 1116 shows pronounced infringements for unit standard enrolments.

For both types of enrolments it is found that the utilization of providers that are not accredited by the submitting ETQE do not have a remarkable impact on adherence in regard to this semantic business rule. Unlike provider accreditations, the accreditation to offer a specific qualification and most unit standards, can only be conducted by a specific ETQE, there is therefore no expectation by an ETQE that the provider has already been accredited to provide the qualification or unit standard by another ETQE.

Specific recommendations in regard to data records that are not compliant to this semantic business rule are provided in Appendix P.4.1 for qualification enrolments and P.4.2 for unit standard enrolments.

#### **4.6 Assessor registration**

This section presents the results of the analysis of learner enrolment records in relation to whether the assessor was registered at the time of the learner's completion of the learnership or achievement of the qualification/unit standard. The section therefore focuses on the nominal data value ASOR\_IND which contains a value denoting the record's compliance in regard to whether the assessor was registered at the time of the completion of the learnership or achievement of the qualification/unit standard.

This section presents the results of the analysis of this data field for learnership enrolment records, qualification enrolment records and unit standard enrolment records.

##### ***4.6.1 Learnership enrolments***

As defined in Appendix C.2, the indicator ASOR\_IND denotes whether the assessor was registered at the time of the completion of the learnership. The manner in which the categories in this indicator is derived is detailed in Appendix C.3.6. An overview of the derived categories, with ASOR\_IND\_DESC shown as Description and the frequency of these values as a percentage, is presented in Table 4.6.1.1:



Table 4.6.1.1 Assessor registration categories for learnership enrolments

Description	% Records
Lshp Completed After Assessor Registration	0.21%
Lshp Completed Before Assessor Registration	0.19%
No Assessor Provided	23.19%
No Registration	0.07%
Not Completed	36.26%
OK	8.93%
Pre First Submission Lshp Completed After Assessor Registration	0.08%
Pre First Submission Lshp Completed Before Assessor Registration	0.37%
Pre First Submission No Assessor Provided	13.48%
Pre First Submission No Registration	0.02%
Pre First Submission Not Completed	17.21%
<b>Total</b>	<b>100.00%</b>

An overview of what the text in these categories denotes is as follows:

- ‘Lshp Completed After Assessor Registration’ denotes a record where the learnership was completed after the assessor’s active registration time period,
- ‘Lshp Completed Before Assessor Registration’ denotes a record where the learnership was completed before the assessor’s active registration time period,
- ‘No Assessor Provided’ denotes a record where the learnership has been completed but no assessor information was provided with the learnership enrolment record,
- ‘No Registration’ denotes a record where the assessor has never had an active registration,
- ‘Not Completed’ denotes a record where the learnership has not been completed,
- ‘OK’ indicates that the assessor was registered at the time of the completion of the learnership, and
- ‘Pre First Submission’ indicates a record where the learner enrolled on the learnership prior to the first full data submission from the ETQE to the NLRD (Section 3.8.3.1).

Any record with a category of ‘No Assessor Provided’ is a learnership enrolment that has been completed but the details of an assessor have not been provided on the learnership enrolment record. The provision of assessor information against a completed learnership is optional and as a result these records are considered correct for the purposes of this research.

Any record with a category of ‘Not Completed’ is a learnership enrolment that has not been completed. Assessor information may only be provided against a learnership enrolment

record if the learnership enrolment has been completed. For the purposes of this research these records are considered correct.

Any record with a category that starts with the text 'Pre First Submission' is a learnership enrolment record with an enrolment date that precedes the date on which the submitting ETQE made its first full data submission to the NLRD. As a result the history of the assessor's active registration time period may not be complete (Section 3.8.3.1) resulting in the possibility that the data in the NLRD in this regard would be incomplete. It was decided in consultation with the Director of the NLRD that these records will be considered as correct for the purposes of this research.

The only categories of records that are considered for this research have a description of 'Lshp Completed After Assessor Registration', 'Lshp Completed Before Assessor Registration' or 'No Registration'. Figure 4.6.1.1 presents an overview of the percentage of records that infringe on the semantic business rule that requires that the assessor must be registered at the time of the completion of the learnership. For illustrative purposes the figure also includes the categories 'Not Completed' and 'No Assessor Provided'.

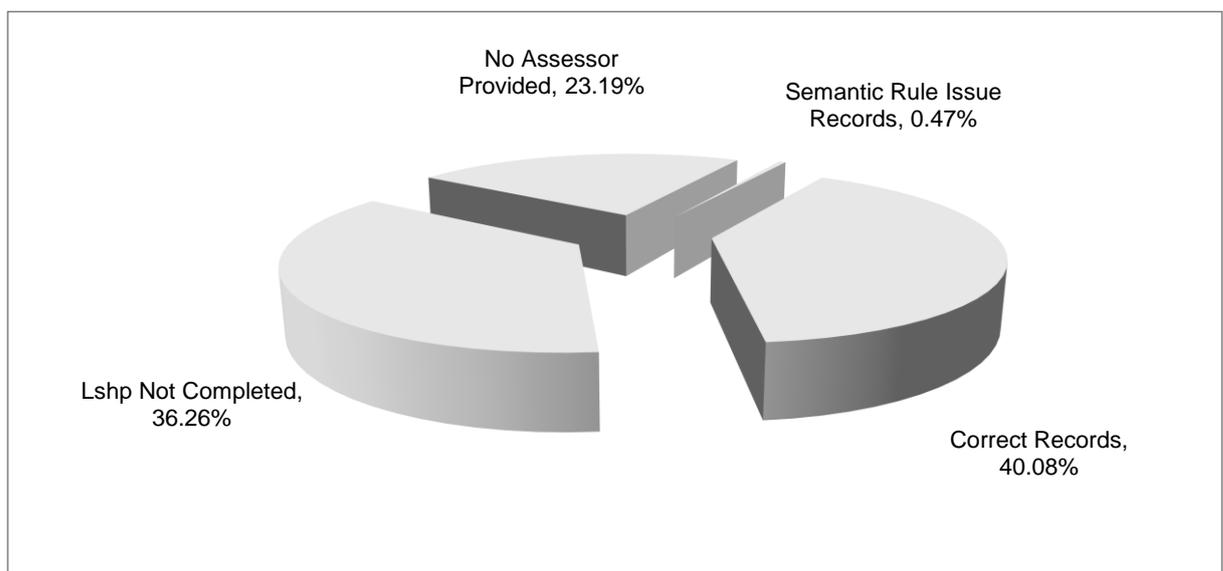


Figure 4.6.1.1 % records according to the semantic business rule that requires that the assessor must be registered at the time of the completion of the learnership

The total percentage of records that infringe on this semantic business rule is very low, namely 0.47%. The records that infringe on this semantic business rule are comprised of three categories:

- Lshp Completed After Assessor Registration (44.46%)

This category indicates that the learnership enrolment was completed and assessed by an assessor after the assessor's registration expired.

Of the 27 discrete ETQEs in the dataset, 5 ETQEs are linked to this category. Of these records, 82.58% were submitted to the NLRD by 3 ETQEs.

Of the 814 discrete learnerships in the dataset, 50 learnerships are linked to this category. Of these 50 learnerships, 10 learnerships contribute to 65.89% of records in this category.

Of the 1767 discrete assessors in the dataset, 80 assessors are linked to this category. Of these 80 assessors, 10 assessors contribute to 48.83% of the records. Most notably, although 41 of the 80 assessors contribute 45.40% of the records; the records for these assessors represent 100% of the learnership enrolment records submitted to the NLRD for the assessor.

- Lshp Completed Before Assessor Registration (40.53%)

This category indicates that the learnership enrolment was completed and assessed by an assessor that was not yet registered.

Of the 27 discrete ETQEs in the dataset, 6 ETQEs are linked to this category. Of these records, 90.85% were submitted to the NLRD by 3 ETQEs.

Of the 814 discrete learnerships in the dataset, 30 learnerships are linked to this category. Of these 30 learnerships, 10 learnerships contribute to 92.60% of records in this category.

Of the 1767 discrete assessors in the dataset, 59 assessors are linked to this category. Of these 59 assessors, 10 assessors contribute to 67.83% of the records. Most notably, although 23 of the 59 assessors contribute 26.92% of the records; the records for these

assessors represent 100% of the learnership enrolment records submitted to the NLRD for the assessor.

- No Registration (15.00%)

This category indicates that the assessor that conducted the assessment of the completed learnership has never had an active registration. This category is of greatest concern to SAQA.

Of the 27 discrete ETQEs in the dataset, 3 ETQEs are linked to this category. Of the 814 discrete learnerships in the dataset, 10 learnerships are linked to this category.

Of the 1767 discrete assessors in the dataset, 24 assessors are linked to this category. Of these 24 assessors, 10 assessors contribute to 90.91% of the records. Most notably, 22 of the 24 assessors contribute 91.64% of the records; the records for these assessors represent 100% of the learnership enrolment records submitted to the NLRD for the assessor.

As already noted, this category indicates that the assessor has never had an active registration. As a result, assessors in this category cannot be reported on in any of the other categories that form part of this research. However, another category in which these assessors can exist in is the 'Pre First Submission No Registration' category that is excluded from this research.

The reader should therefore note that the above statement “...22 of the 24 assessors contribute 91.64% of the records; the records for these assessors represent 100% of the learnership enrolment records submitted to the NLRD for the assessor” must further be interpreted to mean that for the remaining 2 assessors, 100% of the records for the specific assessor fall into the categories 'No Registration' or 'Pre First Submission No Registration'.

Unlike any of the other categories for this semantic business rule, it is unlikely that learnership enrolment records that appear in the 'No Registration' category do so as a result of data capturing issues related to the enrolment record. Rather the data

capturing or data quality issues reside in the assessor record. As a result, the analysis of this category focuses on the assessor records.

Table 4.6.1.2 provides an overview of the records found in this category grouped by submitting ETQE and learnership identifier.

Table 4.6.1.2 ‘No Registration’ records by submitting ETQE identifier and learnership identifier, count of assessors, % learnership enrolment records in the category and records in this category as a % of the records submitted by the ETQE

ETQE Identifier	LSHP Identifier	Count of ASSESOR_ID	% of Rule	% of Records Submitted
<b>1103</b>	729	1	2.17%	0.01%
	730	1	4.33%	0.03%
<b>1103 Total</b>		<b>2</b>	<b>6.50%</b>	<b>0.04%</b>
<b>1109</b>	1093	1	1.81%	0.02%
<b>1109 Total</b>		<b>1</b>	<b>1.81%</b>	<b>0.02%</b>
<b>1116</b>	23	2	2.17%	0.09%
	24	16	20.58%	0.89%
	27	1	0.36%	0.02%
	31	2	3.97%	0.17%
	34	1	0.36%	0.02%
	1325	1	0.36%	0.02%
	1327	2	63.90%	2.75%
<b>1116 Total</b>		<b>25</b>	<b>91.70%</b>	<b>3.95%</b>
<b>Grand Total</b>		<b>28</b>	<b>100.00%</b>	<b>4.01%</b>

Analysis in Table 4.6.1.2 shows the following notable trends:

- ETQE identifier 1116 has the highest incidence of the number of assessors in this category (25 of which 22 are unique assessors). Further analysis shows that the 22 assessors represent 55% of the overall number of assessors that this ETQE references in learnership enrolment records.
- ETQE identifier 1116 also has the highest percentage of records in this category (91.70%).
- ETQE identifier 1116 has the highest percentage of learnership enrolment records in this category when compared to the total number of records submitted to the NLRD by the ETQE (3.95%).

Overall the results for this semantic business rule indicate very few issues exist in regard to whether the assessor was registered at the time of the completion of the learnership. In the

case of the ‘Lshp Completed After Assessor Registration‘ and ‘Lshp Completed Before Assessor Registration‘ categories; the incidence of infringements against this rule is so low and/or so diverse as to suggest that the issues arise as a result of data capturing problems rather than systemic issues with assessor registrations. Although the incidence of records in the ‘No Registration’ category is also very low, the analysis seems to indicate the possibility that there are systemic issues in regard to assessor registrations for ETQE identifier 1116.

#### 4.6.2 Qualification enrolments

As defined in Appendix E.2, the indicator ASOR\_IND denotes whether the assessor was registered at the time of the achievement of the qualification. The manner in which the categories in this indicator is derived is detailed in Appendix E.3.8. An overview of the derived categories, with ASOR\_IND\_DESC shown as Description and the frequency of these values as a percentage, is presented in Table 4.6.2.1:

Table 4.6.2.1 Assessor registration categories for qualification enrolments

Description	% Records
No Assessor Provided	37.75%
No Registration	0.22%
No Registration (Qual Linked to Lshp)	0.00%
Not Achieved	50.20%
OK	9.94%
Pre First Submission Qual Achieved After Assessor Registration	0.12%
Pre First Submission Qual Achieved After Assessor Registration (Qual Linked to Lshp)	0.02%
Pre First Submission Qual Achieved Before Assessor Registration	1.21%
Pre First Submission Qual Achieved Before Assessor Registration (Qual Linked to Lshp)	0.02%
Qual Achieved After Assessor Registration	0.38%
Qual Achieved After Assessor Registration (Qual Linked to Lshp)	0.00%
Qual Achieved Before Assessor Registration	0.14%
Qual Achieved Before Assessor Registration (Qual Linked to Lshp)	0.00%
<b>Total</b>	<b>100.00%</b>

An overview of what the text in these categories denotes is as follows:

- ‘Qual Achieved After Assessor Registration’ denotes a record where the qualification was achieved after the assessor’s active registration time period,
- ‘Qual Achieved Before Assessor Registration’ denotes a record where the qualification was achieved before the assessor’s active registration time period,
- ‘No Assessor Provided’ denotes a record where the qualification has been achieved but no assessor information was provided with the qualification enrolment record,

- ‘No Registration’ denotes a record where the assessor has never had an active registration,
- ‘Not Achieved’ denotes a record where the qualification has not been achieved,
- ‘OK’ indicates that the assessor was registered at the time of the achievement of the qualification,
- ‘Pre First Submission’ indicates a record where the learner enrolled on the qualification prior to the first full data submission from the ETQE to the NLRD (Section 3.8.3.1), and
- ‘(Qual Linked to Lshp)’ indicates that the qualification is linked to a learnership and as a result the problem may be related to the issue described in Section 3.8.3.7.

Any record with a category that ends with the text ‘(Qual Linked to Lshp)’ may have infringed on this semantic business rule as a result of the contextual issues described in Section 3.8.3.7. It was decided, in consultation with the Director of the NLRD, that the further analysis of these types of records will not form part of this research. The percentage of records for this specific issue is however illustrated for all of the semantic business rules in order to provide an understanding of how the specific issue impacts the data in the NLRD.

Any record with a category of ‘No Assessor Provided’ is a qualification enrolment that has been achieved but the details of an assessor have not been provided on the qualification enrolment record. The provision of assessor information against an achieved qualification is optional and as a result these records are considered correct for the purposes of this research.

Any record with a category of ‘Not Achieved’ is a qualification enrolment that has not been achieved. Assessor information may only be provided against a qualification enrolment record if the qualification enrolment has been achieved. For the purposes of this research these records are considered correct.

Any record with a category that starts with the text ‘Pre First Submission’ is a qualification enrolment record with an enrolment date that precedes the date on which the submitting ETQE made its first full data submission to the NLRD. As a result the history of the assessor’s active registration time period may not be complete (Section 3.8.3.1) resulting in

the possibility that the data in the NLRD in this regard would be incomplete. It was decided in consultation with the Director of the NLRD that these records will be considered as correct for the purposes of this research.

The only categories of records that are considered for this research have a description of ‘Qual Achieved After Assessor Registration’, ‘Qual Achieved Before Assessor Registration’ or ‘No Registration’. Figure 4.6.2.1 presents an overview of the percentage of records that infringe on the semantic business rule that requires that the assessor must be registered at the time of the achievement of the qualification. For illustrative purposes the figure also includes the categories ‘Not Achieved’ and ‘No Assessor Provided’.

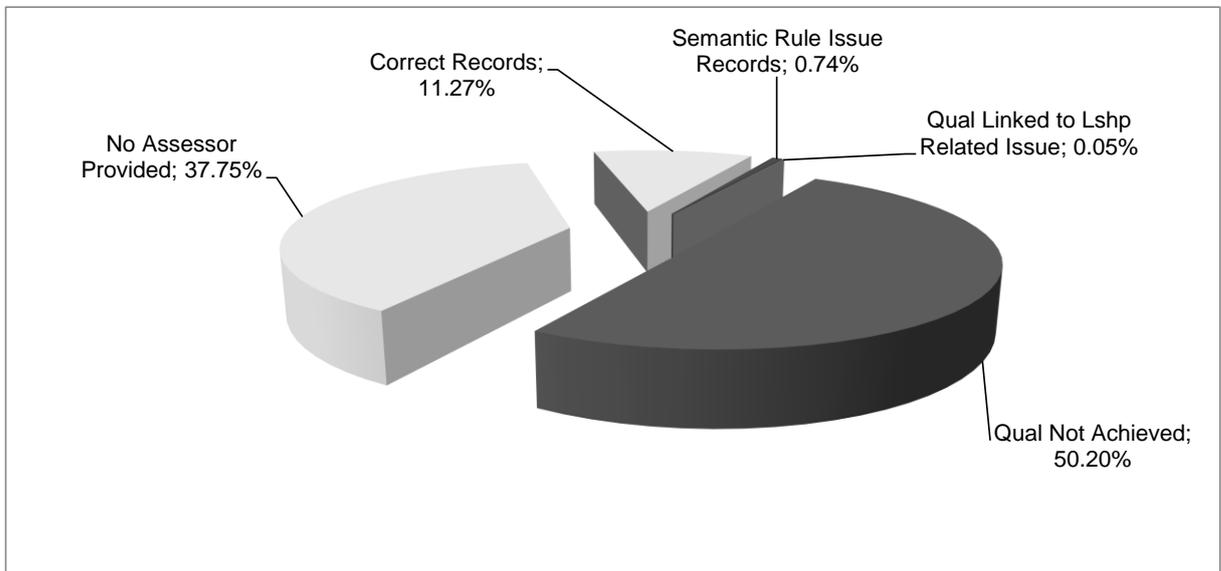


Figure 4.6.2.1 % records according to the semantic business rule that requires that the assessor must be registered at the time of the achievement of the qualification

The total percentage of records that infringe on this semantic business rule is very low, namely 0.74%. The records that infringe on this semantic business rule are comprised of three categories:

- Qual Achieved After Assessor Registration (51.11%)

This category indicates that the qualification enrolment was achieved and assessed by an assessor after the assessor’s registration expired.

Of the 29 discrete ETQEs in the dataset, 8 ETQEs are linked to this category. Of these records, 98.38% were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 75 qualifications are linked to this category. Of these 75 qualifications, 10 qualifications contribute to 84.64% of records in this category.

Of the 2336 discrete assessors in the dataset, 100 assessors are linked to this category. Of these 100 assessors, 10 assessors contribute to 83.05% of the records. Most notably, assessor identifier 3018255 contributes to 62.14% of the records found in this category. Further, although 26 of the 100 assessors contribute 3.48% of the records; the records for these assessors represent 100% of the qualification enrolment records submitted to the NLRD for the assessor.

- No Registration (29.52%)

This category indicates that the assessor that conducted the assessment of the achieved qualification has never had an active registration. This category is of greatest concern to SAQA.

Of the 29 discrete ETQEs in the dataset, 6 ETQEs are linked to this category. Of these records, 96.92% were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 72 qualifications are linked to this category. Of these 72 qualifications, 10 qualifications contribute to 62.24% of records in this category.

Of the 2336 discrete assessors in the dataset, 194 assessors are linked to this category. Of these 194 assessors, 10 assessors contribute to 36.98% of the records. Most notably, 177 of the 194 assessors contribute 61.37% of the records; the records for these assessors represent 100% of the qualification enrolment records submitted to the NLRD for the assessor.

As already noted, this category indicates that the assessor has never had an active registration. As a result, assessors in this category cannot be reported on in any of the

other categories that form part of this research. However, another category that these assessors can exist in is the ‘No Registration (Qual Linked to Lshp)’ category that is excluded from this research.

The reader should therefore note that the above statement “...177 of the 194 assessors contribute 61.37% of the records; the records for these assessors represent 100% of the qualification enrolment records submitted to the NLRD for the assessor” must further be interpreted to mean that for the remaining 17 assessors, 100% of the records for the specific assessor fall into the categories ‘No Registration’ or ‘No Registration (Qual Linked to Lshp)’.

Unlike any of the other categories for this semantic business rule, it is unlikely that qualification enrolment records that appear in the ‘No Registration’ category do so as a result of data capturing issues related to the enrolment record. Rather the data capturing or data quality issues reside in the assessor record. As a result, the effort of the analysis of this category focuses on the assessor records.

Table 4.6.2.2 provides an overview of the records found in this category grouped by submitting ETQE.

Table 4.6.2.2 ‘No Registration’ records by submitting ETQE identifier, count of qualification identifier, count of assessors, % qualification enrolment records in the category and records in this category as a % of the records submitted by the ETQE

ETQE Identifier	Count of QUAL Identifier	Count of Assessor Identifier	% of Rule	% of Records Submitted
1034	3	4	2.21%	1.08%
1102	1	1	0.04%	0.01%
1103	2	2	0.77%	0.02%
1115	46	253	67.31%	3.42%
1116	13	37	8.34%	0.65%
1126	7	22	21.33%	0.27%
<b>Grand Total</b>	<b>72</b>	<b>319</b>	<b>100.00%</b>	<b>5.45%</b>

Analysis in Table 4.6.2.2 shows the following notable trends:

- ETQE identifier 1115 has the highest incidence of the number of assessors in this category (253 of which 143 are unique assessors). Further analysis shows that the

143 assessors represent 54.58% of the overall number of assessors that this ETQE references in qualification enrolment records.

- ETQE identifier 1115 also has the highest percentage of records in this category (67.31%).
  - ETQE identifier 1115 has the highest percentage of qualification enrolment records in this category, when compared to the total number of records submitted to the NLRD by the ETQE (3.42%).
- Qual Achieved Before Assessor Registration (19.38%)  
This category indicates that the qualification enrolment was achieved and assessed by an assessor that was not yet registered.

Of the 29 discrete ETQEs in the dataset, 8 ETQEs are linked to this category. Of these records, 86.35% were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 60 qualifications are linked to this category. Of these 60 qualifications, 10 qualifications contribute to 85.58% of records in this category.

Of the 2336 discrete assessors in the dataset, 102 assessors are linked to this category. Of these 102 assessors, 10 assessors contribute to 56.61% of the records. Most notably, although 32 of the 102 assessors contribute 25.68% of the records; the records for these assessors represent 100% of the qualification enrolment records submitted to the NLRD for the assessor.

Overall the results for this semantic business rule indicate very few issues exist in regard to whether the assessor was registered at the time of the achievement of the qualification. In the case of the ‘Qual Achieved After Assessor Registration’ and ‘Qual Achieved Before Assessor Registration’ categories; the incidence of infringements against this rule is so low and/or so diverse as to suggest that the issues arise as a result of data capturing problems rather than systemic issues with assessor registrations. Although the incidence of records in the ‘No Registration’ category is also very low, the analysis seems to indicate the possibility that there are systemic issues in regard to assessor registrations for ETQE identifier 1115.

### 4.6.3 Unit Standard enrolments

As defined in Appendix G.2, the indicator ASOR\_IND denotes whether the assessor was registered at the time of the achievement of the unit standard. The manner in which the categories in this indicator is derived is detailed in Appendix G.3.8. An overview of the derived categories, with ASOR\_IND\_DESC shown as Description and the frequency of these values as a percentage, is presented in Table 4.6.3.1:

Table 4.6.3.1 Assessor registration categories for unit standard enrolments

Description	% Records
No Assessor Provided	52.61%
No Registration	0.88%
No Registration (UStd Linked to Lshp)	0.00%
Not Achieved	33.70%
OK	11.34%
Pre First Submission UStd Achieved After Assessor Registration	0.04%
Pre First Submission UStd Achieved After Assessor Registration (UStd Linked to Lshp)	0.00%
Pre First Submission UStd Achieved Before Assessor Registration	0.84%
Pre First Submission UStd Achieved Before Assessor Registration (UStd Linked to Lshp)	0.01%
UStd Achieved After Assessor Registration	0.31%
UStd Achieved After Assessor Registration (UStd Linked to Lshp)	0.00%
UStd Achieved Before Assessor Registration	0.25%
UStd Achieved Before Assessor Registration (UStd Linked to Lshp)	0.00%
<b>Total</b>	<b>100.00%</b>

An overview of what the text in these categories denotes is as follows:

- ‘Ustd Achieved After Assessor Registration’ denotes a record where the unit standard was achieved after the assessor’s active registration time period,
- ‘Ustd Achieved Before Assessor Registration’ denotes a record where the unit standard was achieved before the assessor’s active registration time period,
- ‘No Assessor Provided’ denotes a record where the unit standard has been achieved but no assessor information was provided with the unit standard enrolment record,
- ‘No Registration’ denotes a record where the assessor has never had an active registration,
- ‘Not Achieved’ denotes a record where the unit standard has not been achieved,
- ‘OK’ indicates that the assessor was registered at the time of the achievement of the unit standard,

- ‘Pre First Submission’ indicates a record where the learner enrolled on the unit standard prior to the first full data submission from the ETQE to the NLRD (Section 3.8.3.1), and
- ‘(Ustd Linked to Lshp)’ indicates that the unit standard is linked to a learnership and as a result the problem may be related to the issue described in Section 3.8.3.7.

Any record with a category that ends with the text ‘(Ustd Linked to Lshp)’ may have infringed on this semantic business rule as a result of the contextual issues described in Section 3.8.3.7. It was decided, in consultation with the Director of the NLRD, that the further analysis of these types of records will not form part of this research. The percentage of records for this specific issue is however illustrated for all of the semantic business rules in order to provide an understanding of how the specific issue impacts the data in the NLRD.

Any record with a category of ‘No Assessor Provided’ is a unit standard enrolment that has been achieved but the details of an assessor have not been provided on the unit standard enrolment record. The provision of assessor information against an achieved unit standard is optional and as a result these records are considered correct for the purposes of this research.

Any record with a category of ‘Not Achieved’ is a unit standard enrolment that has not been achieved. Assessor information may only be provided against a unit standard enrolment record if the unit standard enrolment has been achieved. For the purposes of this research these records are considered correct.

Any record with a category that starts with the text ‘Pre First Submission’ is a unit standard enrolment record with an enrolment date that precedes the date on which the submitting ETQE made its first full data submission to the NLRD. As a result the history of the assessor’s active registration time period may not be complete (Section 3.8.3.1) resulting in the possibility that the data in the NLRD in this regard would be incomplete. It was decided in consultation with the Director of the NLRD that these records will be considered as correct for the purposes of this research.

The only categories of records that are considered for this research have a description of ‘Ustd Achieved After Assessor Registration’, ‘Ustd Achieved Before Assessor Registration’

or 'No Registration'. Figure 4.6.3.1 presents an overview of the percentage of records that infringe on the semantic business rule that requires that the assessor must be registered at the time of the achievement of the unit standard. For illustrative purposes the figure also includes the categories 'Not Achieved' and 'No Assessor Provided'.

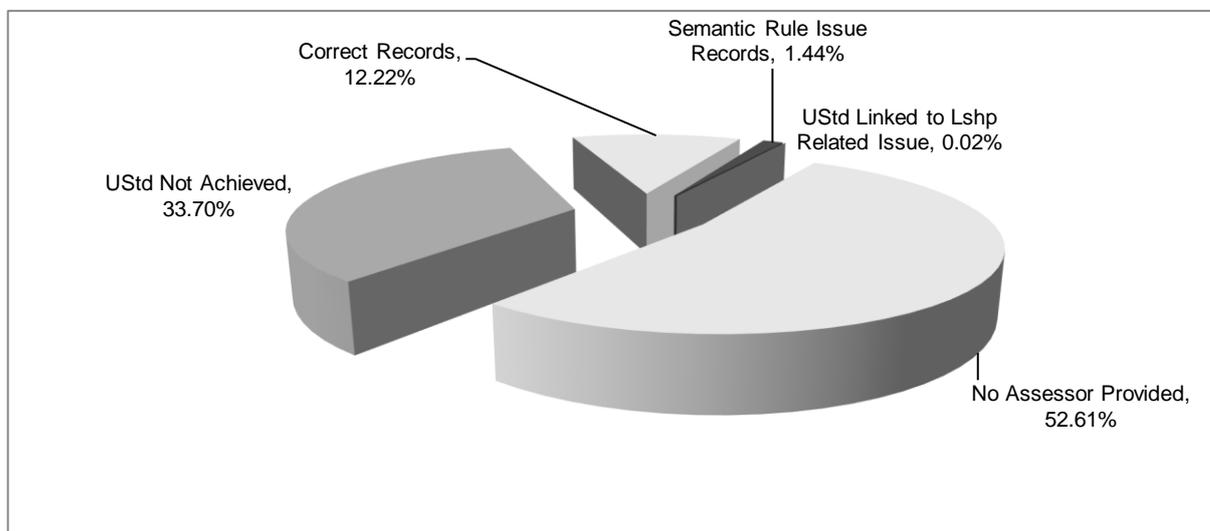


Figure 4.6.3.1 % records according to the semantic business rule that requires that the assessor must be registered at the time of the achievement of the unit standard

The total percentage of records that infringe on this semantic business rule is very low, namely 1.44%. The records that infringe on this semantic business rule are comprised of three categories:

- No Registration (60.71%)

This category indicates that the assessor that conducted the assessment of the achieved unit standard has never had an active registration. This category is of greatest concern to SAQA.

Of the 29 discrete ETQEs in the dataset, 12 ETQEs are linked to this category. Of these records, 98.69% were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 794 are linked to this category. Of these 794 unit standards, 10 unit standards contribute to 18.64% of records in this category.

Of the 9178 discrete assessors in the dataset, 285 assessors are linked to this category. Of these 285 assessors, 10 assessors contribute to 90.45% of the records. Most notably, assessor identifier 3013505 contributes to 83.57% of the records found in this category. Further, 246 of the 285 assessors contribute 12.08% of the records; the records for these assessors represent 100% of the unit standard enrolment records submitted to the NLRD for the assessor.

As already noted, this category indicates that the assessor has never had an active registration. As a result, assessors in this category cannot be reported on in any of the other categories that form part of this research. However, another category that these assessors can exist in is the ‘No Registration (Ustd Linked to Lshp)’ category that is excluded from this research.

The reader should therefore note that the above statement “...246 of the 285 assessors contribute 12.08% of the records; the records for these assessors represent 100% of the unit standard enrolment records submitted to the NLRD for the assessor” must further be interpreted to mean that for the remaining 39 assessors, 100% of the records for the specific assessor fall into the categories ‘No Registration’ or ‘No Registration (Ustd Linked to Lshp)’.

Unlike any of the other categories for this semantic business rule, it is unlikely that unit standard enrolment records that appear in the ‘No Registration’ category do so as a result of data capturing issues related to the enrolment record. Rather the data capturing or data quality issues reside in the assessor record. As a result, the effort of the analysis of this category focuses on the assessor records.

Table 4.6.3.2 provides an overview of the records found in this category grouped by submitting ETQE.

Table 4.6.3.2 ‘No Registration’ records by submitting ETQE identifier, count of unit standard identifier, count of assessors, % unit standard enrolment records in the category and records in this category as a % of the records submitted by the ETQE

ETQE Identifier	Count of USTD Identifier	Count of Assessor Identifier	% of Rule	% of Records Submitted
1102	79	12	0.25%	0.11%
1105	1	1	0.00%	0.00%
1107	5	2	0.00%	0.00%
1109	26	7	0.25%	0.09%
1111	149	20	0.99%	0.07%
1114	3	2	0.01%	0.00%
1115	3433	179	9.19%	8.32%
1116	139	10	84.96%	37.20%
1118	29	4	0.01%	0.02%
1120	43	29	0.02%	0.01%
1123	36	1	0.13%	0.02%
1126	472	18	4.19%	0.30%
<b>Grand Total</b>	<b>4415</b>	<b>285</b>	<b>100.00%</b>	<b>46.12%</b>

Analysis in Table 4.6.3.2 shows the following notable trends:

- ETQE identifier 1115 has the highest incidence of the number of assessors in this category (179 assessors).
- ETQE identifier 1116 has the highest percentage of records in this category (84.96%).
- ETQE identifier 1116 also has the highest percentage of unit standard enrolment records in this category, when compared to the total number of records submitted to the NLRD by the ETQE (37.20%).
- Ustd Achieved After Assessor Registration (21.77%)

This category indicates that the unit standard enrolment was achieved and assessed by an assessor after the assessor’s registration expired.

Of the 29 discrete ETQEs in the dataset, 18 ETQEs are linked to this category. Of these records, 57.99% were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 1534 are linked to this category. Of these 1534 unit standards, 10 unit standards contribute to 19.16% of records in this category. Most notably, although 5 of the 1534 unit standards contribute less than 0.01% of the records; the records for these unit standards represent 100% of the unit standard enrolment records submitted to the NLRD for the unit standard.

Of the 9178 discrete assessors in the dataset, 543 assessors are linked to this category. Of these 543 assessors, 10 assessors contribute to 37.11% of the records. Most notably, although 82 of the 100 assessors contribute 0.05% of the records; the records for these assessors represent 100% of the unit standard enrolment records submitted to the NLRD for the assessor.

- Ustd Achieved Before Assessor Registration (17.52%)

This category indicates that the unit standard enrolment was achieved and assessed by an assessor that was not yet registered.

Of the 29 discrete ETQEs in the dataset, 17 ETQEs are linked to this category. Of these records, 80.70% were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 1169 are linked to this category. Of these 1169 unit standards, 10 unit standards contribute to 18.56% of records in this category. Most notably, although 2 of the 1169 unit standards contribute less than 0.01% of the records; the records for these unit standards represent 100% of the unit standard enrolment records submitted to the NLRD for the unit standard.

Of the 9178 discrete assessors in the dataset, 625 assessors are linked to this category. Of these 625 assessors, 10 assessors contribute to 37.34% of the records. Most notably, although 100 of the 625 assessors contribute 0.09% of the records; the records for these assessors represent 100% of the unit standard enrolment records submitted to the NLRD for the assessor.

Overall the results for this semantic business rule indicate very few issues exist in regard to whether the assessor was registered at the time of the achievement of the unit standard. In the case of the ‘Ustd Achieved After Assessor Registration’ and ‘Ustd Achieved Before Assessor Registration’ categories; the incidence of infringements against this rule is so low and/or so diverse as to suggest that the issues arise as a result of data capturing problems rather than systemic issues with assessor registrations. Although the incidence of records in the ‘No Registration’ category is also very low, the analysis seems to indicate the

possibility that there are systemic issues in regard to assessor registrations for ETQE identifier 1116.

#### **4.6.4 Conclusion**

This section focuses on the analysis of learner enrolment records in relation to whether the assessor was registered at the time of the learner's completion of the learnership or achievement of the qualification/unit standard. The section therefore focuses on the nominal data value ASOR\_IND which contains a value denoting the record's compliance in regard to whether the assessor was registered at the time of the completion of the learnership or achievement of the qualification/unit standard.

Overall the results for this semantic business rule indicate very few issues exist in regard to whether the assessor was registered at the time of the learner's completion of the learnership or achievement of the qualification/unit standard. The analysis highlights that the ETQEs that most frequently infringe on this semantic business rule can be described as follows:

- ETQE Identifier 1116 has the highest rate of infringements for both learnership and unit standard enrolments, and
- ETQE Identifier 1115 shows pronounced infringements for qualification enrolments.

Specific recommendations in regard to data records that are not compliant to this semantic business rule are provided in Appendix P.5.1 for learnership enrolments, Appendix P.5.2 for qualification enrolments and Appendix P.5.3 for unit standard enrolments.

#### **4.7 Assessor registration to assess the qualification or unit standard**

This section presents the results of the analysis of learner enrolment records in relation to whether the assessor was registered to assess the qualification/unit standard at the time of the learner's achievement of the qualification/unit standard. The section therefore focuses on the nominal data value ASOR\_REGSTR\_IND which contains a value denoting the record's compliance in this regard.

This section presents the results of the analysis of this data field for qualification enrolment records and unit standard enrolment records.

#### 4.7.1 Qualification enrolments

As defined in Appendix E.2, the indicator ASOR\_REGSTR\_IND denotes whether the assessor was registered to assess the qualification at the time of the achievement of the qualification. The manner in which the categories in this indicator is derived is detailed in Appendix E.3.9. An overview of the derived categories, with ASOR\_REGSTR\_IND\_DESC shown as Description and the frequency of these values as a percentage, is presented in Table 4.7.1.1:

Table 4.7.1.1 Assessor registration to assess the qualification categories

Description	% Records
No Assessor Provided	37.75%
No Registration	1.35%
No Registration (Qual Linked to Lshp)	0.09%
Not Achieved	50.20%
OK	9.12%
Pre First Submission Qual Achieved After Assessor Registration	0.12%
Pre First Submission Qual Achieved After Assessor Registration (Qual Linked to Lshp)	0.01%
Pre First Submission Qual Achieved Before Assessor Registration	0.62%
Pre First Submission Qual Achieved Before Assessor Registration (Qual Linked to Lshp)	0.01%
Qual Achieved After Assessor Registration	0.28%
Qual Achieved After Assessor Registration (Qual Linked to Lshp)	0.00%
Qual Achieved Before Assessor Registration	0.44%
Qual Achieved Before Assessor Registration (Qual Linked to Lshp)	0.00%
<b>Total</b>	<b>100.00%</b>

An overview of what the text in these categories denotes is as follows:

- ‘Qual Achieved After Assessor Registration’ denotes a record where the qualification was achieved after the assessor’s active registration to assess the qualification time period,
- ‘Qual Achieved Before Assessor Registration’ denotes a record where the qualification was achieved before the assessor’s active registration to assess the qualification time period,
- ‘No Assessor Provided’ denotes a record where the qualification has been achieved but no assessor information was provided with the qualification enrolment record,
- ‘No Registration’ denotes a record where the assessor has never had an active registration to assess the qualification,
- ‘Not Achieved’ denotes a record where the qualification has not been achieved,
- ‘OK’ indicates that the assessor was registered to assess the qualification at the time of the achievement of the qualification, and

- ‘Pre First Submission’ indicates a record where the learner enrolled on the qualification prior to the first full data submission from the ETQE to the NLRD (Section 3.8.3.1), and
- ‘(Qual Linked to Lshp)’ indicates that the qualification is linked to a learnership and as a result the problem may be related to the issue described in Section 3.8.3.7.

Any record with a category that ends with the text ‘(Qual Linked to Lshp)’ may have infringed on this semantic business rule as a result of the contextual issues described in Section 3.8.3.7. It was decided, in consultation with the Director of the NLRD, that the further analysis of these types of records will not form part of this research. The percentage of records for this specific issue is however illustrated for all of the semantic business rules in order to provide an understanding of how the specific issue impacts the data in the NLRD.

Any record with a category of ‘No Assessor Provided’ is a qualification enrolment that has been achieved but the details of an assessor have not been provided on the qualification enrolment record. The provision of assessor information against an achieved qualification is optional and as a result these records are considered correct for the purposes of this research.

Any record with a category of ‘Not Achieved’ is a qualification enrolment that has not been achieved. Assessor information may only be provided against a qualification enrolment record if the qualification enrolment has been achieved. For the purposes of this research these records are considered correct.

Any record with a category that starts with the text ‘Pre First Submission’ is a qualification enrolment record with an enrolment date that precedes the date on which the submitting ETQE made its first full data submission to the NLRD. As a result the history of the assessor’s active registration to assess the qualification time period may not be complete (Section 3.8.3.1) resulting in the possibility that the data in the NLRD in this regard would be incomplete. It was decided in consultation with the Director of the NLRD that these records will be considered as correct for the purposes of this research.

The only categories of records that are considered for this research have a description of ‘Qual Achieved After Assessor Registration’, ‘Qual Achieved Before Assessor Registration’

or 'No Registration'. Figure 4.7.1.1 presents an overview of the percentage of records that infringe on the semantic business rule that requires that the assessor must be registered to assess the qualification at the time of the achievement of the qualification. For illustrative purposes the figure also includes the categories 'Not Achieved' and 'No Assessor Provided'.

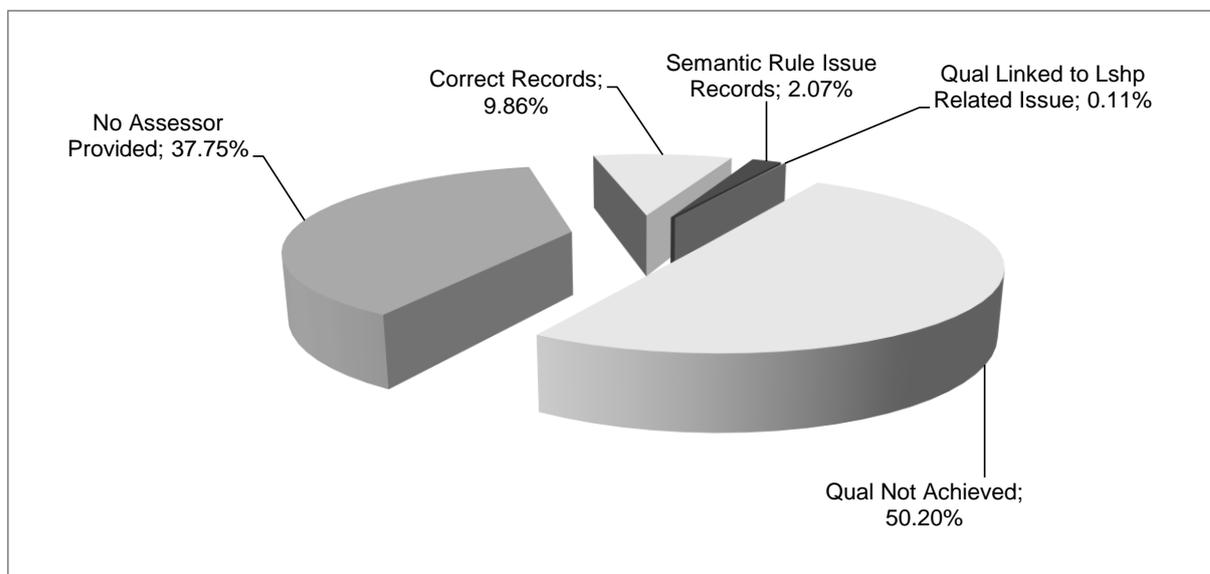


Figure 4.7.1.1 % records according to the semantic business rule that requires that the assessor must be registered to assess the qualification at the time of the achievement of the qualification

The total percentage of records that infringe on this semantic business rule is low, namely 2.07%. The records that infringe on this semantic business rule are comprised of three categories:

- No Registration (65.31%)

This category indicates that the assessor that conducted the assessment of the achieved qualification has never had an active registration to assess the qualification. This category is of greatest concern to SAQA.

Of the 29 discrete ETQEs in the dataset, 13 ETQEs are linked to this category. Of these records, 88.00% were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 106 qualifications are linked to this category. Of these 106 qualifications, 10 qualifications contribute to 88.12% of records in this category.

Of the 2336 discrete assessors in the dataset, 328 assessors are linked to this category. Of these 328 assessors, 10 assessors contribute to 61.39% of the records. Most notably, 257 of the 328 assessors contribute 54.87% of the records; the records for these assessors represent 100% of the qualification enrolment records submitted to the NLRD for the assessor.

As already noted, this category indicates that the assessor has never had an active registration to assess the qualification. As a result, assessors in this category cannot be reported on in any of the other categories that form part of this research. However, another category that these assessors can exist in is the ‘No Registration (Qual Linked to Lshp)’ category that is excluded from this research.

The reader should therefore note that the above statement “...257 of the 328 assessors contribute 54.87% of the records; the records for these assessors represent 100% of the qualification enrolment records submitted to the NLRD for the assessor” must further be interpreted to mean that for the remaining 71 assessors, 100% of the records for the specific assessor fall into the categories ‘No Registration’ or ‘No Registration (Qual Linked to Lshp)’.

Unlike any of the other categories for this semantic business rule, it is unlikely that qualification enrolment records that appear in the ‘No Registration’ category do so as a result of data capturing issues related to the enrolment record. Rather the data capturing or data quality issues reside in the assessor record. As a result, the analysis of this category focuses on the assessor records.

Table 4.7.1.2 provides an overview of the records found in this category grouped by submitting ETQE.

Table 4.7.1.2 ‘No Registration’ records by submitting ETQE identifier, count of qualification identifier, count of assessors, % qualification enrolment records in the category and records in this category as a % of the records submitted by the ETQE

ETQE Identifier	Count of QUAL Identifier	Count of Assessor Identifier	% of Rule	% of Records Submitted
1034	3	19	3.18%	9.06%
1075	1	3	0.15%	0.30%
1102	2	2	0.02%	0.02%
1103	8	14	2.27%	0.42%
1105	4	80	69.54%	10.20%
1106	12	27	14.35%	1.76%
1107	1	1	0.07%	0.06%
1113	5	5	1.03%	0.65%
1114	7	41	1.09%	0.32%
1115	38	115	3.48%	1.03%
1116	14	50	2.27%	1.04%
1120	3	7	0.68%	0.52%
1126	8	28	1.88%	0.14%
<b>Grand Total</b>	<b>106</b>	<b>392</b>	<b>100.00%</b>	<b>25.52%</b>

Analysis in Table 4.7.1.2 shows the following notable trends:

- ETQE identifier 1115 has the highest incidence of the number of assessors in this category (115 of which 76 are unique assessors). Further analysis shows that the 76 assessors represent 29.01% of the overall number of assessors that this ETQE references in qualification enrolment records.
- ETQE identifier 1105 has the highest percentage of records in this category (69.54%). In addition, this ETQE has the second highest incidence of the number of assessors in this category (all of which are unique assessors). Further analysis shows that the 80 assessors represent 82.47% of the overall number of assessors that this ETQE references in qualification enrolment records.
- ETQE identifier 1105 also has the highest percentage of qualification enrolment records in this category, when compared to the total number of records submitted to the NLRD by the ETQE (10.20%).
- Qual Achieved Before Assessor Registration (21.17%)  
This category indicates that the qualification enrolment was achieved and assessed by an assessor that was not yet registered to assess the qualification.

Of the 29 discrete ETQEs in the dataset, 10 ETQEs are linked to this category. Of these records, 77.60% were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 102 qualifications are linked to this category. Of these 102 qualifications, 10 qualifications contribute to 68.73% of records in this category.

Of the 2336 discrete assessors in the dataset, 248 assessors are linked to this category. Of these 248 assessors, 10 assessors contribute to 38.06% of the records. Most notably, although 58 of the 248 assessors contribute 18.89% of the records; the records for these assessors represent 100% of the qualification enrolment records submitted to the NLRD for the assessor.

- Qual Achieved After Assessor Registration (13.52%)

This category indicates that the qualification enrolment was achieved and assessed by an assessor after the assessor's registration to assess the qualification expired.

Of the 29 discrete ETQEs in the dataset, 9 ETQEs are linked to this category. Of these records, 87.29% were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 76 qualifications are linked to this category. Of these 76 qualifications, 10 qualifications contribute to 70.26% of records in this category.

Of the 2336 discrete assessors in the dataset, 112 assessors are linked to this category. Of these 112 assessors, 10 assessors contribute to 72.26% of the records. Further, although 34 of the 112 assessors contribute 12.92% of the records; the records for these assessors represent 100% of the qualification enrolment records submitted to the NLRD for the assessor.

Overall the results for this semantic business rule indicate few issues exist in regard to whether the assessor was registered to assess the qualification at the time of the achievement of the qualification. In the case of the 'No Registration' category the analysis seems to indicate that there are systemic issues in regard to the registration of assessors to

assess qualifications for ETQE identifiers 1105 and 1115. The ‘Qual Achieved Before Assessor Registration‘ and ‘Qual Achieved After Assessor Registration‘ incidences of infringements against this rule is so low and/or so diverse as to suggest that the issues arise as a result of data capturing problems rather than systemic issues with assessor registrations. Overall the ETQE identifiers with the highest infringements for this business rule are 1115, 1106 and 1113.

#### 4.7.2 Unit Standard enrolments

As defined in Appendix G.2, the indicator ASOR\_REGSTR\_IND denotes whether the assessor was registered to assess the unit standard at the time of the achievement of the unit standard. The manner in which the categories in this indicator is derived is detailed in Appendix G.3.9. An overview of the derived categories, with ASOR\_REGSTR\_IND\_DESC shown as Description and the frequency of these values as a percentage, is presented in Table 4.7.2.1:

Table 4.7.2.1 Assessor registration to assess the unit standard categories

Description	% Records
No Assessor Provided	52.58%
No Registration	1.96%
No Registration (UStd Linked to Lshp)	0.00%
Not Achieved	33.78%
OK	9.18%
Pre First Submission UStd Achieved After Assessor Registration	0.09%
Pre First Submission UStd Achieved After Assessor Registration (UStd Linked to Lshp)	0.00%
Pre First Submission UStd Achieved Before Assessor Registration	0.77%
Pre First Submission UStd Achieved Before Assessor Registration (UStd Linked to Lshp)	0.01%
UStd Achieved After Assessor Registration	0.86%
UStd Achieved After Assessor Registration (UStd Linked to Lshp)	0.01%
UStd Achieved Before Assessor Registration	0.77%
UStd Achieved Before Assessor Registration (UStd Linked to Lshp)	0.00%
<b>Total</b>	<b>100.00%</b>

An overview of what the text in these categories denotes is as follows:

- ‘UStd Achieved After Assessor Registration’ denotes a record where the unit standard was achieved after the assessor’s active registration to assess the unit standard time period,
- ‘UStd Achieved Before Assessor Registration’ denotes a record where the unit standard was achieved before the assessor’s active registration to assess the unit standard time period,

- ‘No Assessor Provided’ denotes a record where the unit standard has been achieved but no assessor information was provided with the unit standard enrolment record,
- ‘No Registration’ denotes a record where the assessor has never had an active registration to assess the unit standard,
- ‘Not Achieved’ denotes a record where the unit standard has not been achieved,
- ‘OK’ indicates that the assessor was registered to assess the unit standard at the time of the achievement of the unit standard, and
- ‘Pre First Submission’ indicates a record where the learner enrolled on the unit standard prior to the first full data submission from the ETQE to the NLRD (Section 3.8.3.1), and
- ‘(Ustd Linked to Lshp)’ indicates that the unit standard is linked to a learnership and as a result the problem may be related to the issue described in Section 3.8.3.7.

Any record with a category that ends with the text ‘(Ustd Linked to Lshp)’ may have infringed on this semantic business rule as a result of the contextual issues described in Section 3.8.3.7. It was decided, in consultation with the Director of the NLRD, that the further analysis of these types of records will not form part of this research. The percentage of records for this specific issue is however illustrated for all of the semantic business rules in order to provide an understanding of how the specific issue impacts the data in the NLRD.

Any record with a category of ‘No Assessor Provided’ is a unit standard enrolment that has been achieved but the details of an assessor have not been provided on the unit standard enrolment record. The provision of assessor information against an achieved unit standard is optional and as a result these records are considered correct for the purposes of this research.

Any record with a category of ‘Not Achieved’ is a unit standard enrolment that has not been achieved. Assessor information may only be provided against a unit standard enrolment record if the unit standard enrolment has been achieved. For the purposes of this research these records are considered correct.

Any record with a category that starts with the text ‘Pre First Submission’ is a unit standard enrolment record with an enrolment date that precedes the date on which the submitting

ETQE made its first full data submission to the NLRD. As a result the history of the assessor’s active registration to assess the unit standard time period may not be complete (Section 3.8.3.1) resulting in the possibility that the data in the NLRD in this regard would be incomplete. It was decided in consultation with the Director of the NLRD that these records will be considered as correct for the purposes of this research.

The only categories of records that are considered for this research have a description of ‘Ustd Achieved After Assessor Registration’, ‘Ustd Achieved Before Assessor Registration’ or ‘No Registration’. Figure 4.7.2.1 presents an overview of the percentage of records that infringe on the semantic business rule that requires that the assessor must be registered to assess the unit standard at the time of the achievement of the unit standard. For illustrative purposes the figure also includes the categories ‘Not Achieved’ and ‘No Assessor Provided’.

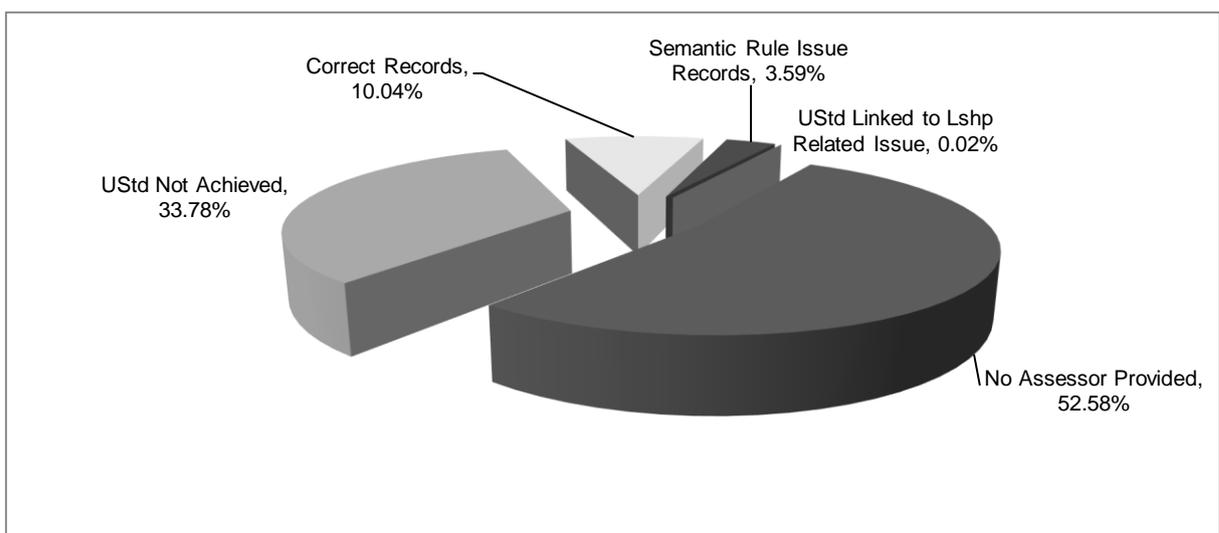


Figure 4.7.2.1 % records according to the semantic business rule that requires that the assessor must be registered to assess the unit standard at the time of the achievement of the unit standard

The total percentage of records that infringe on this semantic business rule is low, namely 3.59%. The records that infringe on this semantic business rule are comprised of three categories:

- No Registration (54.55%)

This category indicates that the assessor that conducted the assessment of the achieved unit standard has never had an active registration to assess the unit standard. This category is of greatest concern to SAQA.

Of the 29 discrete ETQEs in the dataset, 20 ETQEs are linked to this category. Of these records, 74.59% were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 2498 are linked to this category. Of these 2498 unit standards, 10 unit standards contribute to 13.72% of records in this category.

Of the 9178 discrete assessors in the dataset, 1451 assessors are linked to this category. Of these 1451 assessors, 10 assessors contribute to 52.77% of the records. Most notably, 562 of the 1451 assessors contribute 14.31% of the records; the records for these assessors represent 100% of the unit standard enrolment records submitted to the NLRD for the assessor.

As already noted, this category indicates that the assessor has never had an active registration to assess the unit standard. As a result, assessors in this category cannot be reported on in any of the other categories that form part of this research. However, another category that these assessors can exist in is the ‘No Registration (Ustd Linked to Lshp)’ category that is excluded from this research.

The reader should therefore note that the above statement “...562 of the 1451 assessors contribute 14.31% of the records; the records for these assessors represent 100% of the unit standard enrolment records submitted to the NLRD for the assessor” must further be interpreted to mean that for the remaining 889 assessors, 100% of the records for the specific assessor fall into the categories ‘No Registration’ or ‘No Registration (Ustd Linked to Lshp)’.

Unlike any of the other categories for this semantic business rule, it is unlikely that unit standard enrolment records that appear in the ‘No Registration’ category do so as a result of data capturing issues related to the enrolment record. Rather the data

capturing or data quality issues reside in the assessor record. As a result, the analysis of this category focuses on the assessor records.

Table 4.7.2.2 provides an overview of the records found in this category grouped by submitting ETQE.

Table 4.7.2.2 ‘No Registration’ records by submitting ETQE identifier, count of unit standard identifier, count of assessors, % unit standard enrolment records in the category and records in this category as a % of the records submitted by the ETQE

ETQE Identifier	Count of USTD Identifier	Count of Assessor Identifier	% of Rule	% of Records Submitted
1075	211	29	0.39%	1.26%
1102	11	4	0.03%	0.03%
1103	62	11	0.04%	0.03%
1105	206	34	6.82%	1.16%
1106	3599	398	26.92%	8.25%
1107	1012	179	2.82%	3.58%
1108	17	3	0.04%	2.35%
1109	609	44	2.52%	2.02%
1110	12	8	0.04%	0.05%
1111	874	222	1.78%	0.27%
1112	349	12	0.42%	0.40%
1113	126	15	0.60%	1.27%
1114	61	5	0.20%	0.05%
1115	7812	271	7.80%	16.04%
1116	277	14	36.83%	36.66%
1118	24	2	0.64%	2.89%
1120	104	21	0.14%	0.16%
1123	108	8	0.01%	0.00%
1125	1366	82	1.59%	2.44%
1126	2132	89	10.38%	1.68%
<b>Grand Total</b>	<b>18972</b>	<b>1451</b>	<b>100.00%</b>	<b>80.61%</b>

Analysis in Table 4.7.2.2 shows the following notable trends:

- ETQE identifier 1106 has the highest incidence of the number of assessors in this category (398 assessors).
- ETQE identifier 1116 has the highest percentage of records in this category (36.83%).
- ETQE identifier 1116 also has the highest percentage of unit standard enrolment records in this category, when compared to the total number of records submitted to the NLRD by the ETQE (36.66%).
- Ustd Achieved After Assessor Registration (24.05%)

This category indicates that the unit standard enrolment was achieved and assessed by an assessor after the assessor's registration to assess the unit standard expired.

Of the 29 discrete ETQEs in the dataset, 19 ETQEs are linked to this category. Of these records, 78.81% were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 1767 are linked to this category. Of these 1767 unit standards, 10 unit standards contribute to 59.35% of records in this category. Further, although 1 of the 1767 unit standards contribute less than 0.01% of the records; the records for this unit standard represent 100% of the unit standard enrolment records submitted to the NLRD for the unit standard.

Of the 9178 discrete assessors in the dataset, 1481 assessors are linked to this category. Of these 1481 assessors, 10 assessors contribute to 16.38% of the records. Notably, although 155 of the 1481 assessors contribute 0.03% of the records; the records for these assessors represent 100% of the unit standard enrolment records submitted to the NLRD for the assessor.

- Ustd Achieved Before Assessor Registration (21.40%)

This category indicates that the unit standard enrolment was achieved and assessed by an assessor that was not yet registered to assess the unit standard.

Of the 29 discrete ETQEs in the dataset, 16 ETQEs are linked to this category. Of these records, 80.71% were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 1677 are linked to this category. Of these 1677 unit standards, 10 unit standards contribute to 16.20% of records in this category. Further, although 3 of the 1677 unit standards contribute less than 0.01% of the records; the records for this unit standard represent 100% of the unit standard enrolment records submitted to the NLRD for the unit standard.

Of the 9178 discrete assessors in the dataset, 1432 assessors are linked to this category. Of these 1432 assessors, 10 assessors contribute to 35.15% of the records. Most notably, although 138 of the 1432 assessors contribute 0.04% of the records; the records for these

assessors represent 100% of the unit standard enrolment records submitted to the NLRD for the assessor.

Overall the results for this semantic business rule indicate few issues exist in regard to whether the assessor was registered to assess the unit standard at the time of the achievement of the unit standard. In the case of the 'No Registration' category the analysis seems to indicate that there are systemic issues in regard to the registration of assessors to assess unit standards for ETQE identifiers 1106 and 1116. The 'Ustd Achieved Before Assessor Registration' and 'Ustd Achieved After Assessor Registration' incidences of infringements against this rule is so low and/or so diverse as to suggest that the issues arise as a result of data capturing problems rather than systemic issues with assessor registrations. Overall the ETQE identifiers with the highest infringements for this business rule are 1116, 1106 and 1115.

#### **4.7.3 Conclusion**

This section focuses on the analysis of learner enrolment records in relation to whether the assessor was registered to assess the qualification/unit standard at the time of the learner's achievement of the qualification/unit standard. The section therefore focuses on the nominal data value ASOR\_REGSTR\_IND which contains a value denoting the record's compliance in this regard.

Overall the results for this semantic business rule indicate very few issues exist in regard to whether the assessor was registered to assess the qualification/unit standard at the time of the learner's achievement of the qualification/unit standard. The analysis highlights that the ETQEs that most frequently infringe on this semantic business rule can be described as follows:

- ETQE Identifier 1115 and 1116 both have the highest rate of infringements for both qualification and unit standard enrolments in regard to assessors that are not registered to assess the qualification/unit standard, and
- ETQE Identifier 1106 shows pronounced infringements for unit standard enrolments.

Specific recommendations in regard to data records that are not compliant to this semantic business rule are provided in Appendix P.6.1 for qualification enrolments and Appendix P.6.2 for unit standard enrolments.

#### 4.8 Correlation between learnerships and their associated qualifications

This section presents the results of the analysis of learnership enrolment records in relation to whether, when a learner has completed a learnership, a corresponding qualification achievement record has been submitted to the NLRD. The reader should note that this specific semantic business rule is only applicable to learnership enrolment records.

The section therefore focuses on the indicator QENROL\_IND which, as defined in Appendix C.2, denotes whether when a learner has completed a learnership, a corresponding qualification achievement record has been submitted to the NLRD. The manner in which the categories in this indicator is derived is detailed in Appendix C.3.7. An overview of the derived categories, with QENROL\_IND\_DESC shown as Description and the frequency of these values as a percentage, is presented in Table 4.8.1:

Table 4.8.1 A corresponding qualification achievement record has been submitted for completed learnership categories

Description	% Records
Both Lshp Not Completed and Qual Not Achieved	45.23%
Lshp Completed After Qual	0.04%
Lshp Completed After Qual (Derived)	0.26%
Lshp Completed Before Qual	0.54%
Lshp Completed Before Qual (Derived)	0.58%
Lshp Completed, Qual Enrolled	0.82%
Lshp Completed, Qual Enrolled (Derived)	0.77%
Lshp Enrolled, Qual Achieved	0.84%
Lshp Enrolled, Qual Achieved (Derived)	3.52%
No Qual Enrolment	9.29%
OK	38.11%
<b>Total</b>	<b>100.00%</b>

An overview of what the text in these categories denotes is as follows:

- ‘Both Lshp Not Completed and Qual Not Achieved’ denotes a record where both the learnership has not been completed and the associated qualification has not been achieved,
- ‘Lshp Completed After Qual’ denotes a record where the learnership was completed 12 months or more after the achievement of the qualification,
- ‘Lshp Completed After Qual (Derived)’ denotes a record where the learnership was completed 12 months or more after the achievement of the qualification and the

learnership identifier of the learnership enrolment record differs from the learnership identifier on the qualification enrolment record,

- ‘Lshp Completed Before Qual’ denotes a record where the learnership was completed 12 months or more before the achievement of the qualification,
- ‘Lshp Completed Before Qual (Derived)’ denotes a record where the learnership was completed 12 months or more before the achievement of the qualification and the learnership identifier of the learnership enrolment record differs from the learnership identifier on the qualification enrolment record,
- ‘Lshp Completed, Qual Enrolled’ denotes a record where the learnership has been completed and the associated qualification has not been achieved,
- ‘Lshp Completed, Qual Enrolled (Derived)’ denotes a record where the learnership has been completed and the associated qualification has not been achieved and the learnership identifier of the learnership enrolment record differs from the learnership identifier on the qualification enrolment record,
- ‘Lshp Enrolled, Qual Achieved’ denotes a record where the learnership has not been completed and the associated qualification has been achieved,
- ‘Lshp Enrolled, Qual Achieved (Derived)’ denotes a record where the learnership has not been completed and the associated qualification has been achieved and the learnership identifier of the learnership enrolment record differs from the learnership identifier on the qualification enrolment record,
- ‘No Qual Enrolment’ denotes a record where no associated qualification enrolment record could be found, and
- ‘OK’ indicates that the learnership was completed and the associated qualification was achieved within 12 months of each other.

All of the categories, with the exception of the ‘OK’ and ‘Both Lshp Not Completed and Qual Not Achieved’ categories are considered for this research. The ‘No Qual Enrolment’ category is of greatest concern to SAQA. The category distinguishes itself from the remaining categories in that it does not denote a possible mismatch in information between the learnership enrolment record and its associated qualification enrolment, rather the category denotes missing data (regardless of whether the learnership has been completed). As a consequence the analysis of this semantic business rule has addressed these records as a discrete point of concern. Figure 4.8.1 presents an overview of the percentage of records that

infringe on the semantic business rule that requires that the intrinsic relationship between the completion of a learnership and achievement of its related qualification has been upheld.

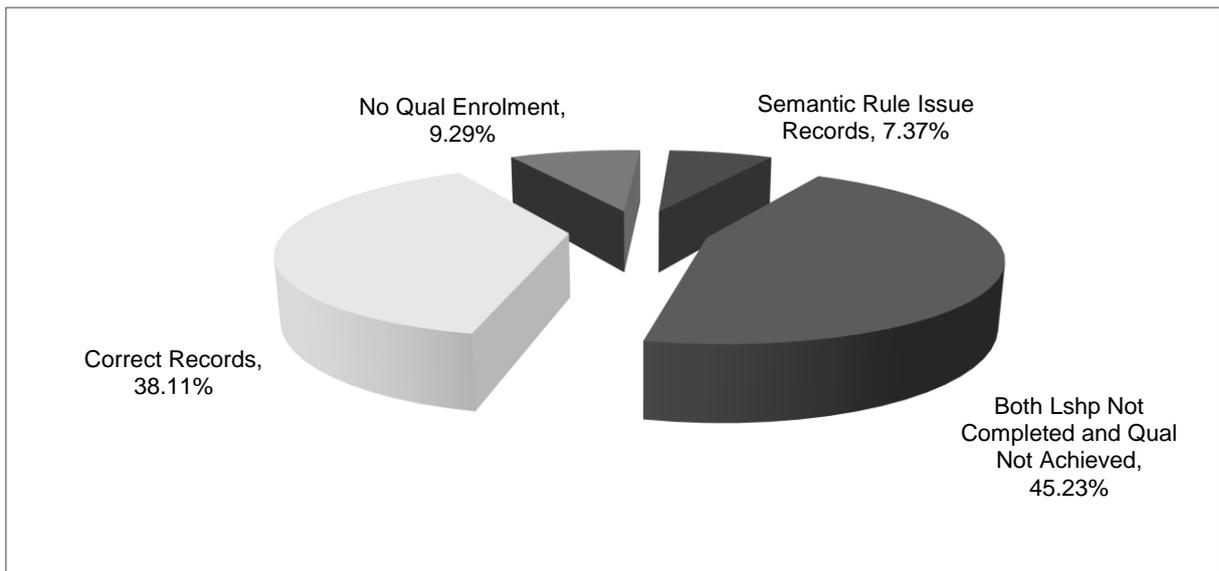


Figure 4.8.1 % records according to the semantic business rule that requires that the intrinsic relationship between the completion of a learnership and achievement of its related qualification has been upheld

The percentage of records that infringe on this semantic business rule is 7.37%. However 9.29% of the learnership enrolment records do not have an associated qualification enrolment record with which to determine whether the record complies with the semantic business rule. Therefore, for the purposes of this research, the total percentage of records that infringe on this semantic business rule is considered to be 16.66%. As a result the records that infringe on this semantic business rule are comprised of 9 categories. Figure 4.8.2 provides an overview of the percentage of records found in each of these categories:

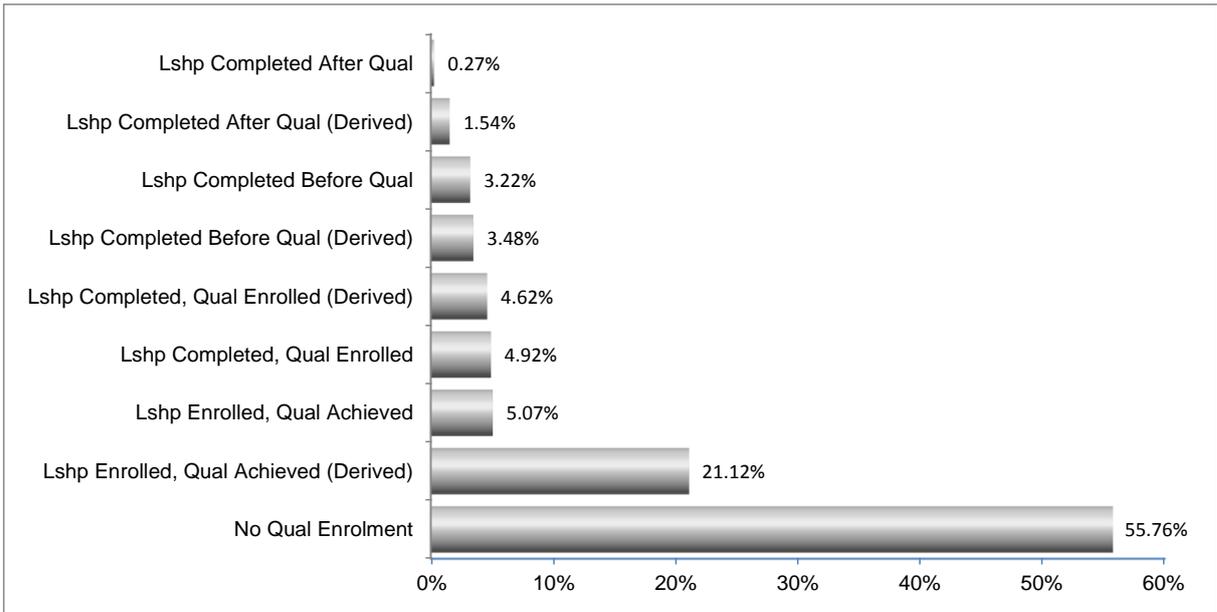


Figure 4.8.2 % records that infringe the semantic business rule that requires that the intrinsic relationship between the completion of a learnership and achievement of its related qualification has been upheld

The scope and volume of records that infringe on this semantic business rule require a detailed review of each of these categories. The following sections present the results of these reviews.

#### ***4.8.1 No Qual Enrolment***

This category indicates that an associated qualification enrolment record does not exist for the learnership enrolment record. This category contains 55.76% of all of the records that infringe on this semantic business rule and is of greatest concern to SAQA.

Of the 27 discrete ETQEs in the dataset, all 27 ETQEs are linked to this category. Of these records, 64.78% were submitted to the NLRD by 3 ETQEs.

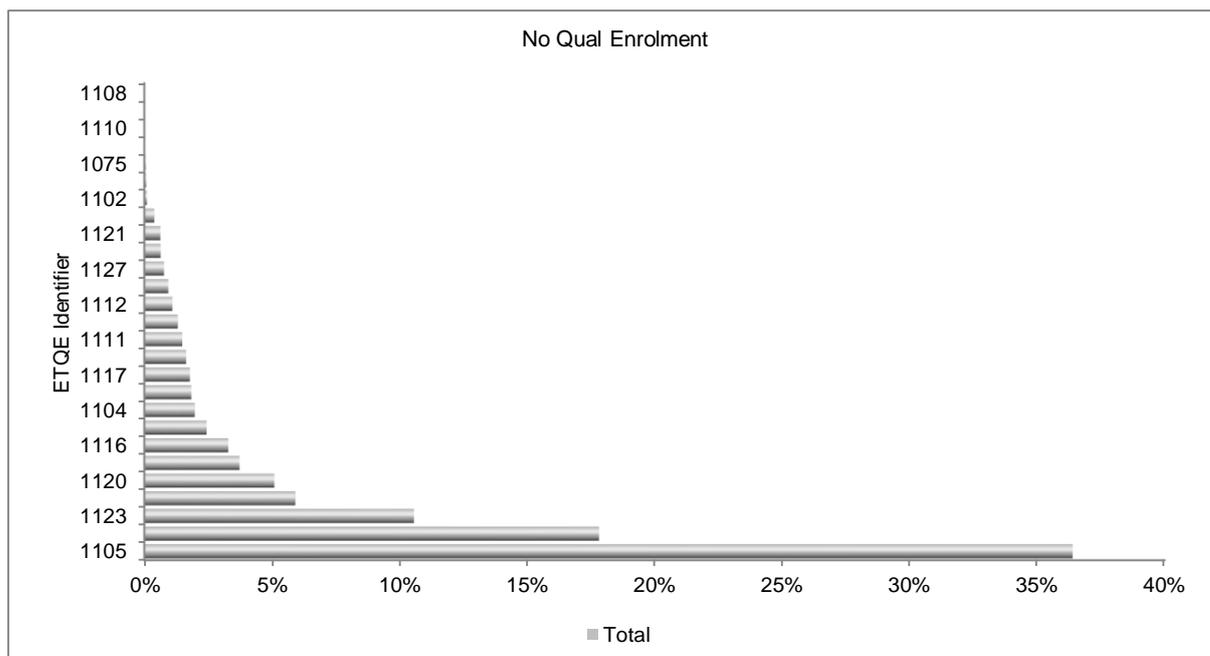


Figure 4.8.1.1 % records by ETQE where an associated qualification enrolment record does not exist for the learnership enrolment record

Of the 814 discrete learnerships in the dataset, 312 are linked to this category. Of these 312 learnerships, 10 contribute to 75.70% of records in this category. Most notably, although 59 of the 312 learnerships constitute 49.08% of the records; the records for these learnerships represent 100% of the enrolment records submitted to the NLRD for the learnerships.

The volume of records found in this category exceeded 9% of the total learnership enrolment records that form part of the research. The implementation of data mining techniques, beyond exploratory data mining techniques, is needed to provide SAQA with both a succinct description of these records and to attempt to identify any data records that exist in this category as a result of possible incorrect data capturing at the source of the data.

The clustering data mining technique, as described in Appendix I.3, is applied to this data set in an effort to achieve both these aims.

The data mining effort results in the development of 8 data clusters (a technical description of each cluster is provided in Appendix N.1) that have the following most dominant characteristics:

1. Cluster 1

This cluster describes more than 35% of the records. All of the records that are in this cluster are linked to one of two providers. This cluster predominantly describes enrolments against learnership identifier 1554 (further analysis shows that all of the records for this learnership are found in this category). Further, the majority of the records in this cluster have been submitted to the NLRD by ETQE identifier 1105.

2. Cluster 2

The cluster describes nearly 19% of the records as belonging to three learnerships (learnership identifier 884, 880 and 878), which have been submitted to the NLRD by ETQE identifier 1126. Further analysis shows that for learnership identifiers 884 and 878 these records represent more than 90% of the learnership enrolment records for these learnerships.

3. Cluster 3

This cluster describes slightly more than 12% of the records. The cluster is diverse in that it describes records submitted by 8 ETQEs, covering 35 learnerships offered by 93 different providers.

4. Cluster 4

The cluster describes nearly 12% of the records as having been submitted to the NLRD by ETQE identifier 1123. The cluster constitutes 2 learnerships (learnership identifier 483 and 474) offered by 21 providers. Further analysis shows that the records for learnership identifier 474 represent nearly 100% of the learnership enrolment records for this learnership.

5. Cluster 5

This cluster describes slightly more than 6.5% of the records as belonging to learnership identifier 364, as offered by 3 providers, and submitted to the NLRD by ETQE identifier 1107. Further analysis shows that these records represent slightly more than 60% of the learnership enrolment records for this learnership.

6. Cluster 6

The cluster describes nearly 6% of the records found in this category as having been submitted to the NLRD by ETQE identifier 1115. These records comprise learnership enrolment records for 6 learnerships, which encompasses 20% of the learnerships that the ETQE implemented. Further, these records are linked to 15 providers, which encompass more than 25% of the providers that the ETQE references in learnership enrolment records.

7. Cluster 7

This cluster describes slightly more than 5.5% of the records. The cluster is diverse in that it describes records submitted by 9 ETQEs, covering 28 learnerships offered by 69 different providers.

#### 8. Cluster 8

The cluster describes nearly 4% of the records found in this category as having been submitted to the NLRD by ETQE identifier 1120. These records constitute learnership enrolment records for 17 learnerships, which encompasses 27% of the learnerships that the ETQE implemented. Further, these records are linked to 5 providers, which encompass more than 16% of the providers that the ETQE references in learnership enrolment records.

The most notable clusters that are generated for this category are clusters 1, 2, 4, 5, 6 and 8. Clusters 1, 2, 4 and 5 seem to describe specific problems with the implementation of specific learnerships whereas Clusters 6 and 8 seem to describe systemic problems arising at the level of the ETQE.

The clustering algorithm does not generate any clusters with a population of less than 1%. A review of the probabilities allocated to each record in a cluster does however allow for the identification of records that are low enough to be considered anomalous as per the criteria defined in Appendix I.3. These records constitute 2.39% of the records found in this category, and possibly exist in this category as a result of data capturing problems at the source of the data

#### ***4.8.2 Lshp Enrolled, Qual Achieved (Derived)***

This category indicates that the learnership enrolment record has a completion status of enrolled whilst its associated qualification enrolment record has an enrolment status of achieved. The linkage between the learnership enrolment record and the qualification enrolment record for these records has not been clearly defined on the qualification enrolment record (i.e. the LEARNERSHIP\_ID on the qualification enrolment record is either NULL or has a value other than the LEARNERSHIP\_ID of the learnership enrolment record). As a result, these learnership enrolment records have a derived association to their qualification enrolment records. This category contains 21.12% of all of the records that infringe on this semantic business rule. Of these records 76.70% are linked to a qualification enrolment record where the learnership identifier on the qualification

enrolment record is NULL. The remaining 23.30% of these records are linked to qualification enrolment records that have a learnership identifier other than the learnership identifier of the learnership enrolment record.

Of the 27 discrete ETQEs in the dataset, 23 ETQEs are linked to this category. Of these records, 56.72% were submitted to the NLRD by 3 ETQEs.

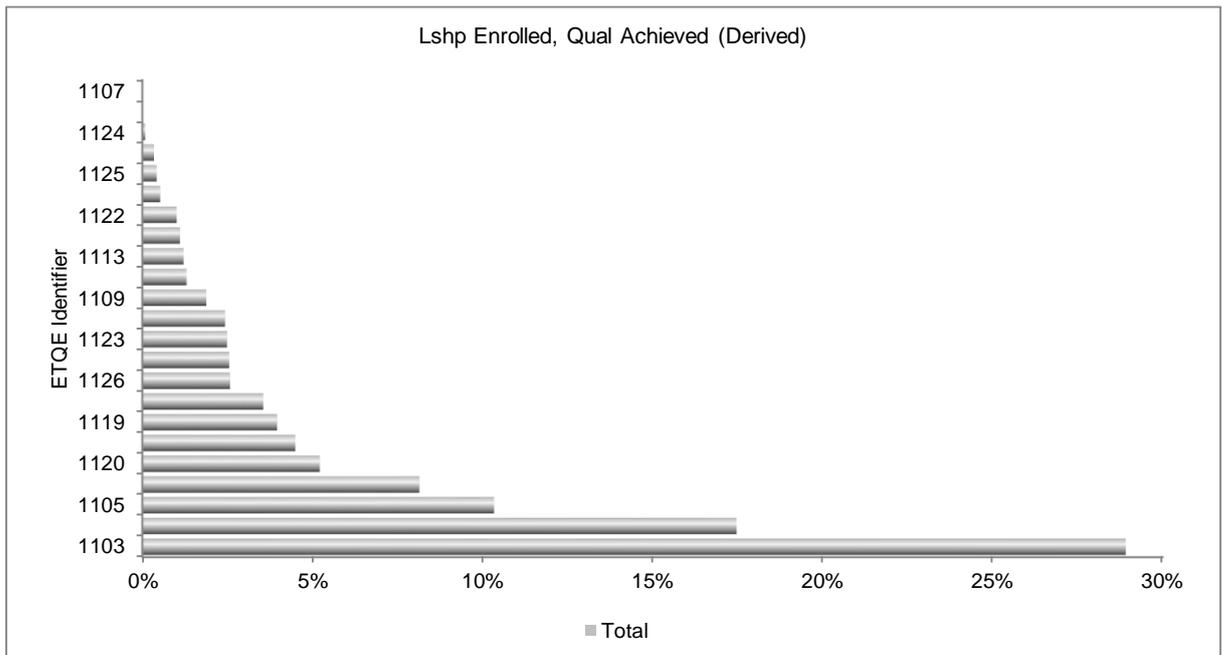


Figure 4.8.2.1 % records by ETQE where the learnership enrolment record has a completion status of enrolled whilst it's associated qualification enrolment record has an enrolment status of achieved

Of the 814 discrete learnerships in the dataset, 262 are linked to this category. Of these 262 learnerships, 10 contribute to 48.58% of records in this category. Most notably, although 5 of the 262 learnerships constitute only 0.26% of the records; the records for these learnerships represent 100% of the enrolment records submitted to the NLRD for the learnerships.

The volume of records found in this category exceeded 3% of the total learnership enrolment records that form part of the research. The implementation of data mining techniques, beyond exploratory data mining techniques, is needed to provide SAQA with

both a succinct description of these records and to attempt to identify any data records that exist in this category as a result of possible incorrect data capturing at the source of the data.

The clustering data mining technique, as described in Appendix I.3, is applied to this data set in an effort to achieve both these aims.

The data mining effort results in the development of 8 data clusters (a technical description of each cluster is provided in Appendix N.2) that have the following most dominant characteristics:

1. Cluster 1

This cluster describes more than 33% of the records. The cluster is diverse in that it describes 22 learnerships that are implemented by 7 ETQEs. The learnerships predominantly share the characteristic of having an NQF Level of Level 4.

2. Cluster 2

The cluster describes nearly 22% of the records as belonging to ETQE identifier 1103. These records are linked to 9 different learnerships which constitute approximately 6% of the learnerships implemented by this ETQE.

3. Cluster 3

This cluster describes slightly more than 17% of the records. The cluster is relatively diverse in that it contains records submitted to the NLRD by 7 different ETQEs and contains records belonging to 17 different learnerships, of which more than half have an NQF Level of Level 3.

4. Cluster 4

The cluster describes slightly more than 10% of the records as having been submitted to the NLRD by ETQE identifier 1105. The records in this cluster are linked to 2 learnerships which constitute nearly 12% of the learnerships implemented by this ETQE.

5. Cluster 5

This cluster describes nearly 7.5% of the records. The cluster is diverse as it contains learnership enrolment records linked to 16 different learnerships as implemented by 5 different ETQEs. The majority of these learnerships have a NQF Level of Level 2.

6. Cluster 6

The cluster predominantly describes 3.75% of the records as having an NQF Level of Level 5. The cluster contains records submitted to the NLRD by 5 ETQEs linked to 9 different learnerships.

7. Cluster 7

This cluster describes slightly more than 3.5% of the records as having been submitted to the NLRD by ETQE identifier 1111. The cluster contains 15 different learnerships which constitutes nearly 17% of the learnerships implemented by this ETQE.

8. Cluster 8

The cluster describes slightly more than 2.5% of the records as having a NQF Level of Level 1. The majority of the records in this cluster belong to three learnerships and have been submitted to the NLRD by ETQE identifier 1115 and 1113.

The most notable clusters that are generated for this category are clusters 2, 4 and 7. Each of these clusters seem to describe specific problems with the implementation of specific learnerships by their implementing ETQEs.

The clustering algorithm does not generate any clusters with a population of less than 1%. A review of the probabilities allocated to each record in a cluster does however allow for the identification of records that are low enough to be considered anomalous as per the criteria defined in Appendix I.3. These records constitute 1.32% of the records found in this category, and possibly exist in this category as a result of data capturing problems at the source of the data.

#### ***4.8.3 Lshp Enrolled, Qual Achieved***

This category indicates that the learnership enrolment record has a completion status of enrolled whilst its associated qualification enrolment record has an enrolment status of achieved. This category contains 5.07% of all of the records that infringe on this semantic business rule.

Of the 27 discrete ETQEs in the dataset, 13 ETQEs are linked to this category. Of these records, 95.71% were submitted to the NLRD by 3 ETQEs.

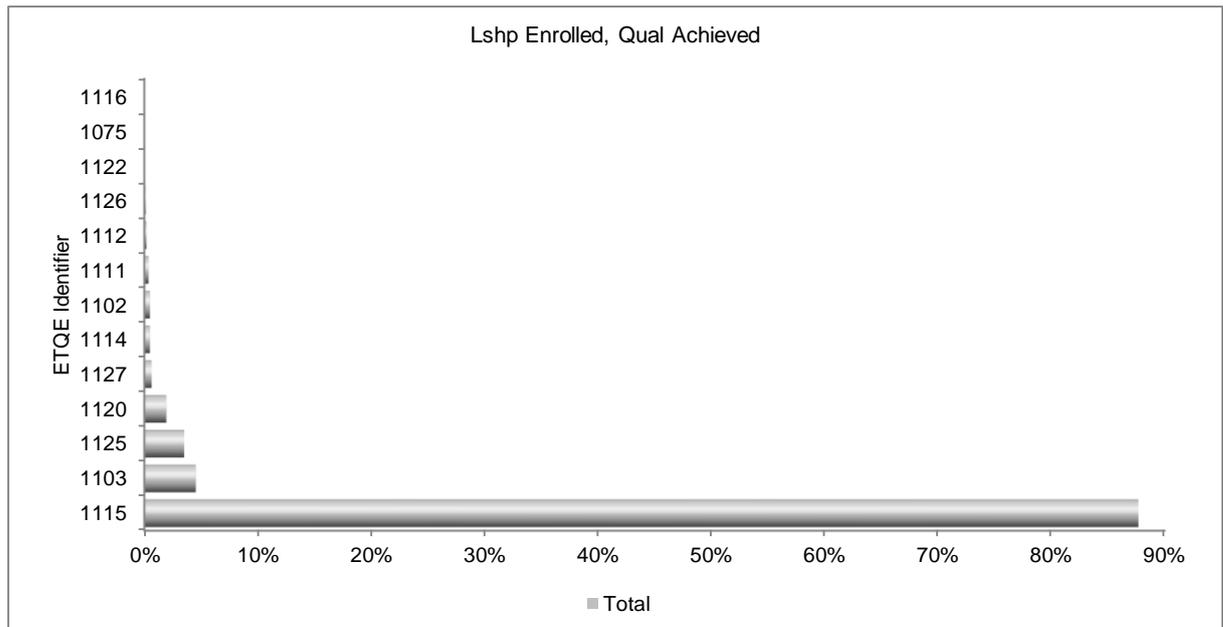


Figure 4.8.3.1 % records by ETQE where the learnership enrolment record has a completion status of enrolled whilst its associated qualification enrolment record has an enrolment status of achieved

Of the 814 discrete learnerships in the dataset, 89 are linked to this category. Of these 89 learnerships, 10 contribute to 74.02% of records in this category.

The most notable characteristic of this category is that nearly 88% (87.67%) of these records were submitted to the NLRD by ETQE identifier 1115. These records constitute more than 20% (21.55%) of the learnership enrolment records submitted to the NLRD by this ETQE. Further, nearly 94% (93.57%) of the learnership enrolments for this ETQE, found in this category, have a *START\_DATE\_IND* range from 70 to 120 (in other words the vast majority of these learnership enrolments started between May 2006 and May 2010). This trend peaked over the *START\_DATE\_IND* range from 94 to 106 (i.e. May 2008 to May 2009).

#### 4.8.4 *Lshp Completed, Qual Enrolled*

This category indicates that the learnership enrolment record has a completion status of completed whilst its associated qualification enrolment record has an enrolment status of not achieved. This category contains 4.92% of all of the records that infringe on this semantic business rule.

Of the 27 discrete ETQEs in the dataset, 10 ETQEs are linked to this category. Of these records, 95.49% were submitted to the NLRD by 3 ETQEs.

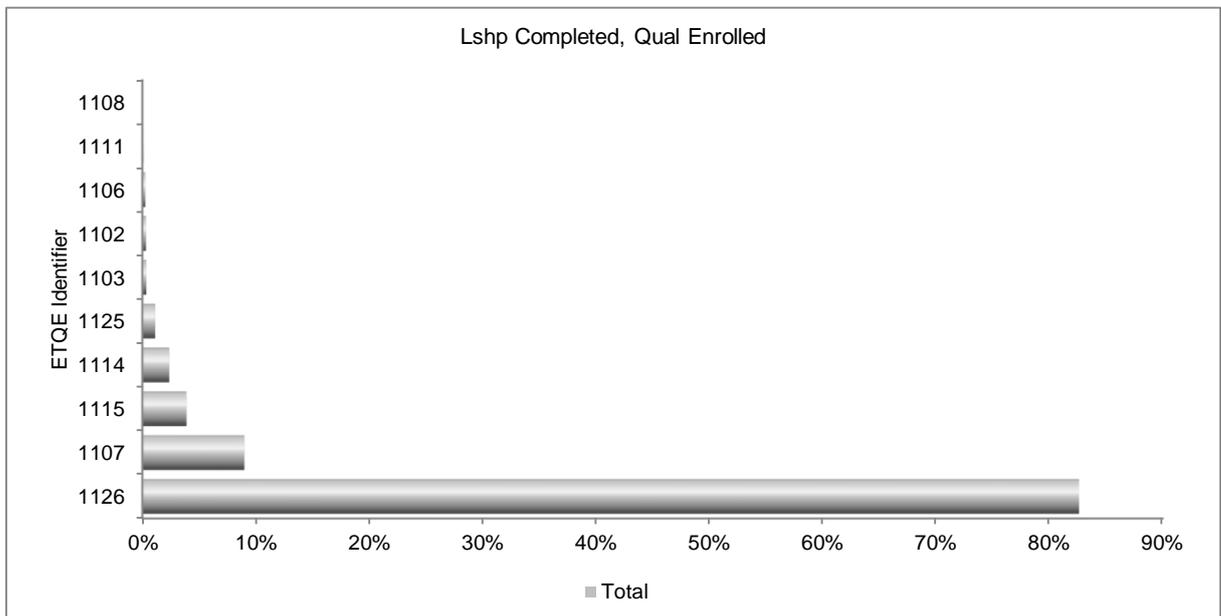


Figure 4.8.4.1 % records by ETQE where the learnership enrolment record has a completion status of completed whilst its associated qualification enrolment record has an enrolment status of not achieved

Of the 814 discrete learnerships in the dataset, 74 are linked to this category. Of these 74 learnerships, 10 contribute to 71.23% of records in this category.

The most notable characteristic of this category is that nearly 83% (82.61%) of these records were submitted to the NLRD by ETQE identifier 1126. These records constitute slightly more than 5% (5.08%) of the learnership enrolment records submitted to the NLRD by this ETQE. Further, nearly 90% (89.99%) of the learnership enrolments for this ETQE, found in this category, have an END\_DATE\_IND range from 85 to 141 (in other words the vast majority of these learnership enrolments were completed between August 2007 and February 2012). This trend peaked over the END\_DATE\_IND range from 107 to 130 (i.e. May 2009 to April 2011).

#### 4.8.5 *Lshp Completed, Qual Enrolled (Derived)*

This category indicates that the learnership enrolment record has a completion status of completed whilst its associated qualification enrolment record has an enrolment status of

not achieved. The linkage between the learnership enrolment record and the qualification enrolment record for these records has not been clearly defined on the qualification enrolment record (i.e. the LEARNERSHIP\_ID on the qualification enrolment record is either NULL or has a value other than the LEARNERSHIP\_ID of the learnership enrolment record). As a result, these learnership enrolment records have a derived association to their qualification enrolment records. This category contains 4.62% of all of the records that infringe on this semantic business rule.

Of these records, 96.72% are linked to a qualification enrolment record where the learnership identifier on the qualification enrolment record is NULL. The remaining 3.28% of these records are linked to qualification enrolment records that have a learnership identifier other than the learnership identifier of the learnership enrolment record.

Of the 27 discrete ETQEs in the dataset, 22 ETQEs are linked to this category. Of these records, 63.93% were submitted to the NLRD by 3 ETQEs.

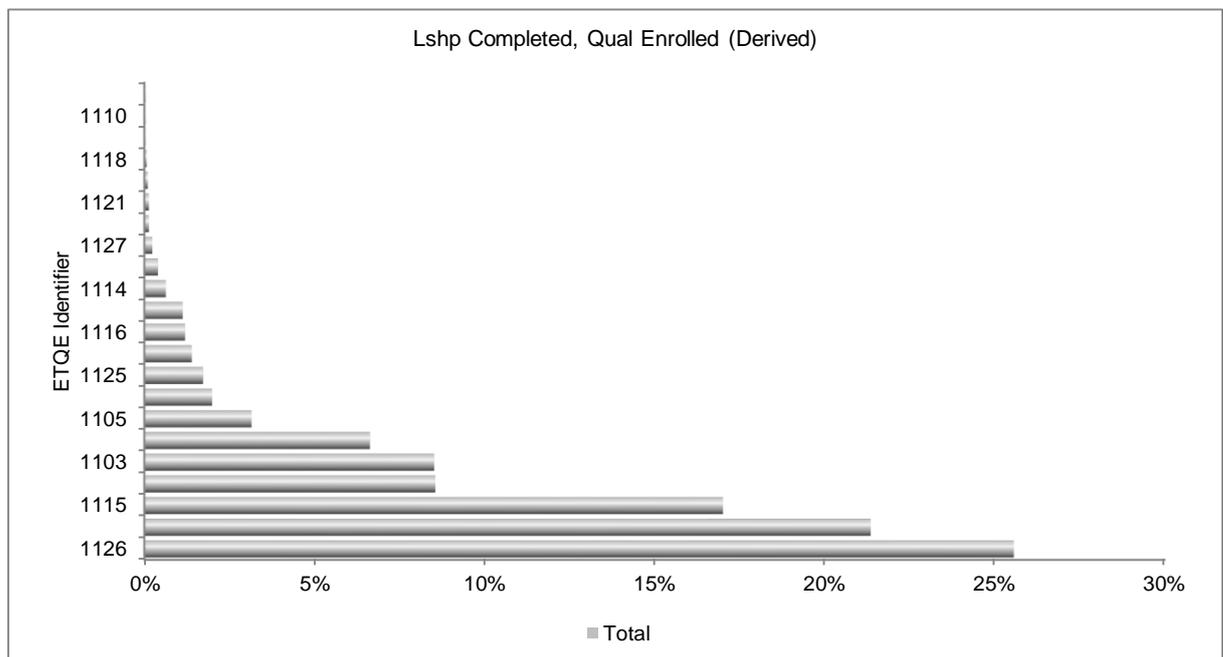


Figure 4.8.5.1 % records by ETQE where the learnership enrolment record has a completion status of completed whilst its derived associated qualification enrolment record has an enrolment status of not achieved

Of the 814 discrete learnerships in the dataset, 115 are linked to this category. Of these 115 learnerships, 10 contribute to 65.75% of records in this category.

The category is relatively diverse with no particularly distinctive characteristics with which to describe its records.

#### ***4.8.6 Lshp Completed Before Qual (Derived)***

This category indicates that the learnership enrolment record was completed more than a year prior to the achievement of the associated qualification enrolment record. The linkage between the learnership enrolment record and the qualification enrolment record for these records has not been clearly defined on the qualification enrolment record (i.e. the LEARNERSHIP\_ID on the qualification enrolment record is either NULL or has a value other than the LEARNERSHIP\_ID of the learnership enrolment record). As a result, these learnership enrolment records have a derived association to their qualification enrolment records. This category contains 3.48% of all of the records that infringe on this semantic business rule.

Of these records 83.93% are linked to a qualification enrolment record where the learnership identifier on the qualification enrolment record is NULL. The remaining 16.07% of these records are linked to qualification enrolment records that have a learnership identifier other than the learnership identifier of the learnership enrolment record.

Of the 27 discrete ETQEs in the dataset, 21 ETQEs are linked to this category. Of these records, 86.05% were submitted to the NLRD by 3 ETQEs.

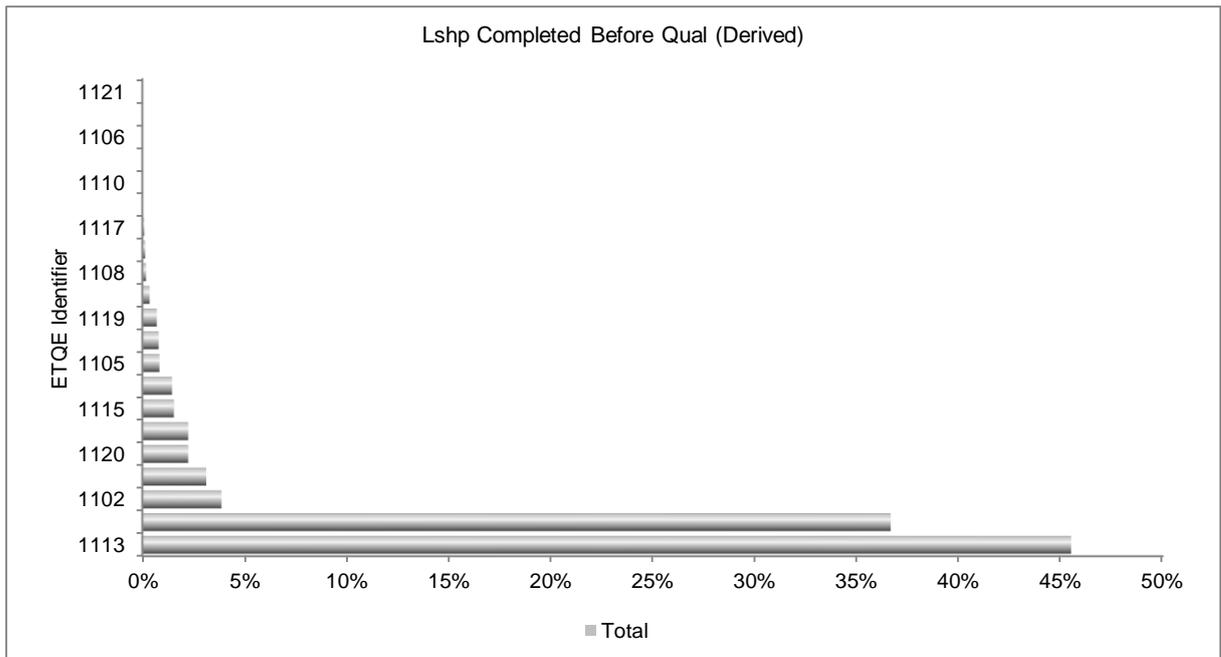


Figure 4.8.6.1 % records by ETQE where the learnership enrolment record was completed more than a year prior to the achievement of the associated qualification enrolment record

Of the 814 discrete learnerships in the dataset, 139 are linked to this category. Of these 139 learnerships, 10 contribute to 41.81% of records in this category.

Slightly more than 97% (97.08%) of the records that fall into this category have a learnership completion date between 1 and 4 years prior to the achievement of the qualification.

The records in this category are relatively diverse with no particularly distinctive characteristics with which to describe its records.

#### ***4.8.7 Lshp Completed Before Qual***

This category indicates that the learnership enrolment record was completed more than a year prior to the achievement of the associated qualification enrolment record. This category contains 3.22% of all of the records that infringe on this semantic business rule.

Of the 27 discrete ETQEs in the dataset, 8 ETQEs are linked to this category. Of these records, 97.86% were submitted to the NLRD by 3 ETQEs.

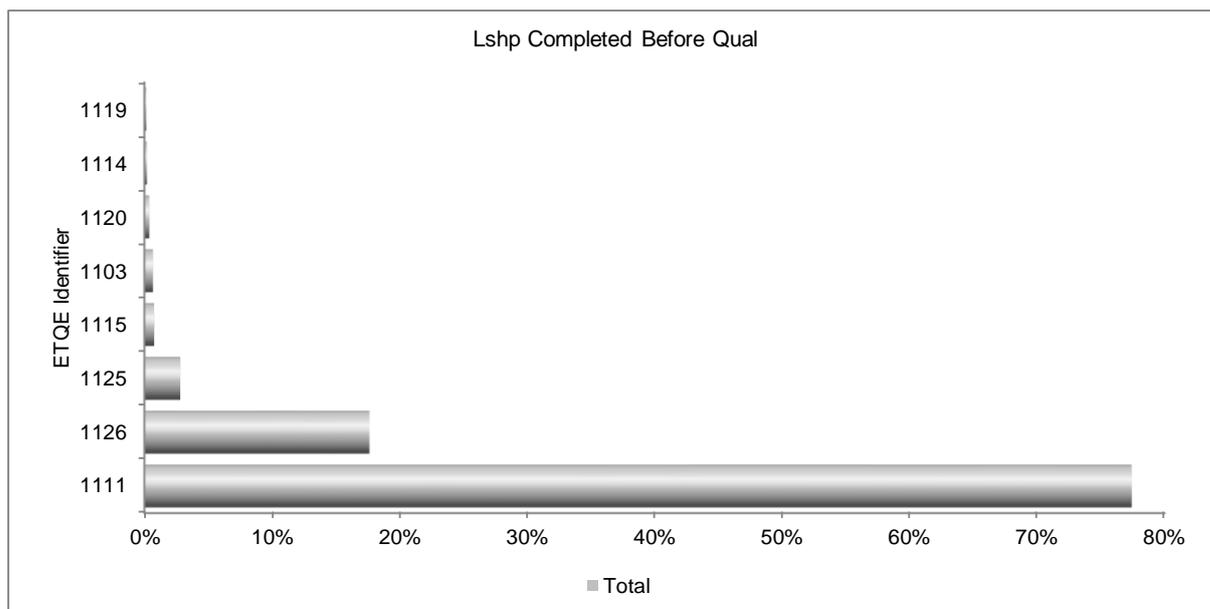


Figure 4.8.7.1 % records by ETQE where the learnership enrolment record was completed more than a year prior to the achievement of the derived associated qualification enrolment record

Of the 814 discrete learnerships in the dataset, 65 are linked to this category. Of these 65 learnerships, 10 contribute to 62.96% of records in this category.

The most notable characteristic of this category is that more than 77% (77.41%) of these records were submitted to the NLRD by ETQE identifier 1111. These records constitute more than 4% (4.32%) of the learnership enrolment records submitted to the NLRD by this ETQE. Further, nearly 99% (98.77%) of the learnership enrolments for this ETQE, found in this category, have an END\_DATE\_IND value of 125 (in other words the vast majority of these learnership enrolments were completed in December 2010).

#### 4.8.8 *Lshp Completed After Qual (Derived)*

This category indicates that the learnership enrolment record was completed more than a year after the achievement of the associated qualification enrolment record. The linkage between the learnership enrolment record and the qualification enrolment record for these records has not been clearly defined on the qualification enrolment record (i.e. the LEARNERSHIP\_ID on the qualification enrolment record is either NULL or has a value other than the LEARNERSHIP\_ID of the learnership enrolment record). As a result, these

enrolment records have a derived association to their qualification enrolment records. This category contains 1.54% of all of the records that infringe on this semantic business rule.

Of these records, 96.73% are linked to a qualification enrolment record where the learnership identifier on the qualification enrolment record is NULL. The remaining 3.27% of these records are linked to qualification enrolment records that have a learnership identifier other than the learnership identifier of the learnership enrolment record.

Of the 27 discrete ETQEs in the dataset, 15 ETQEs are linked to this category. Of these records, 81.85% were submitted to the NLRD by 3 ETQEs.

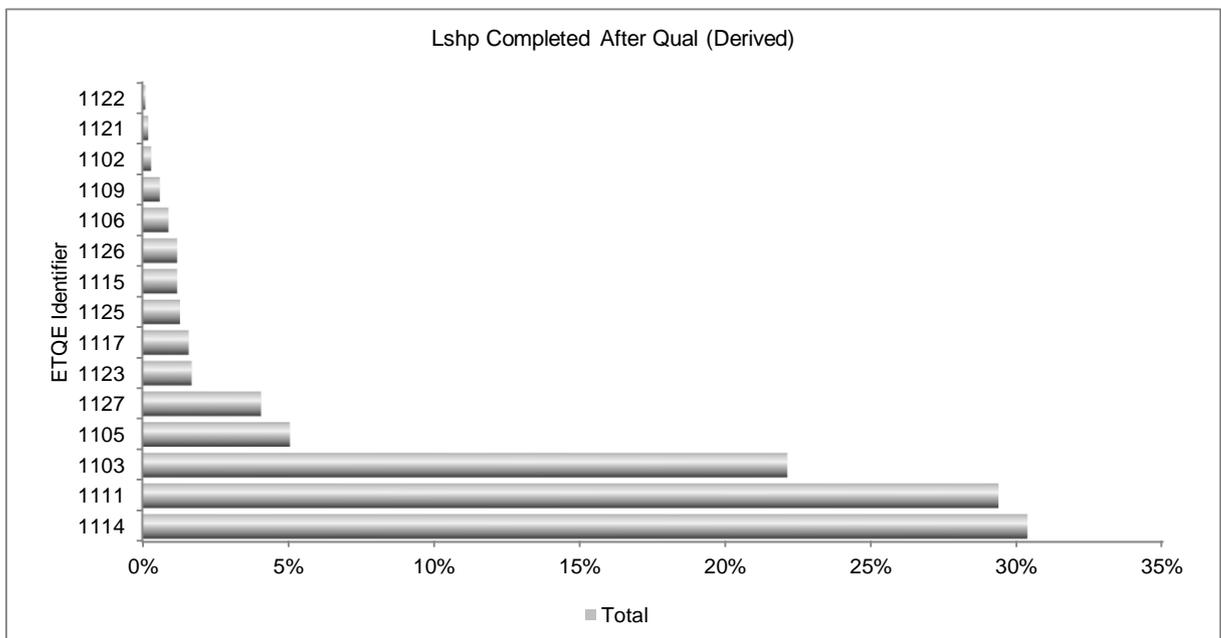


Figure 4.8.8.1 % records by ETQE where the learnership enrolment record was completed more than a year after the achievement of the derived associated qualification enrolment record

Of the 814 discrete learnerships in the dataset, 72 are linked to this category. Of these 72 learnerships, 10 contribute to 66.27% of records in this category.

The records in this category are relatively diverse with no particularly distinctive characteristics with which to describe its records.

#### 4.8.9 Lshp Completed After Qual

This category indicates that the learnership enrolment record was completed more than a year after the achievement of the associated qualification enrolment record. This category contains 0.27% of all of the records that infringe on this semantic business rule.

Of the 27 discrete ETQEs in the dataset, 7 ETQEs are linked to this category. Of these records, 92.53% were submitted to the NLRD by 3 ETQEs.

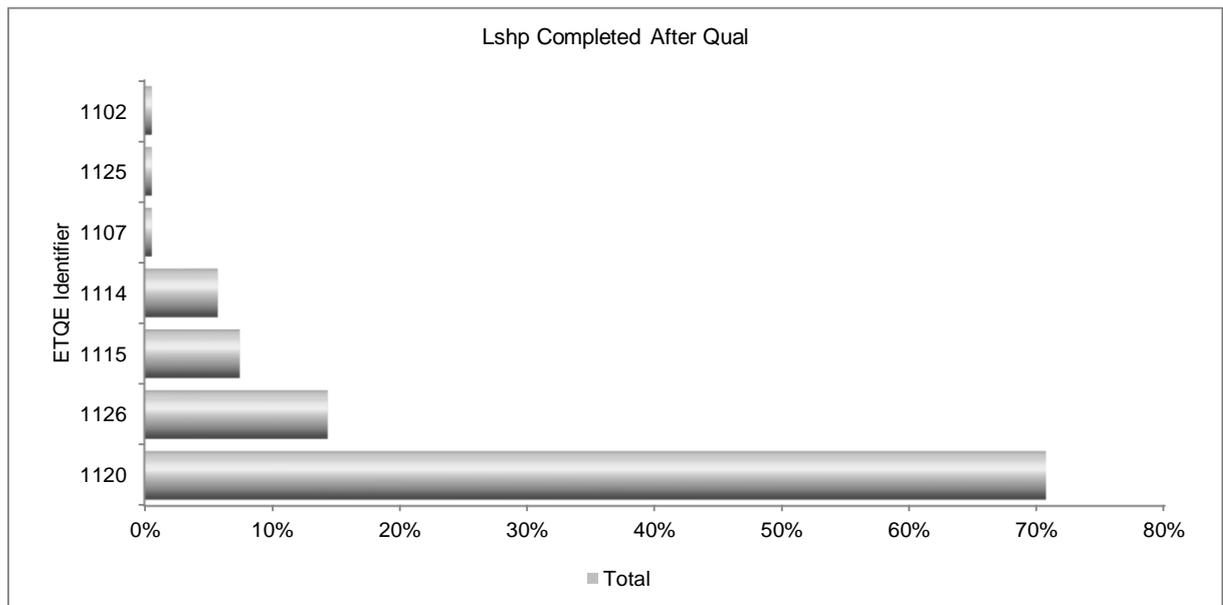


Figure 4.8.9.1 % records by ETQE where the learnership enrolment record was completed more than a year after the achievement of the associated qualification enrolment record

Of the 814 discrete learnerships in the dataset, 23 are linked to this category. Of these 23 learnerships, 10 contribute to 92.53% of records in this category.

The most notable characteristic of this category is that nearly 71% (70.69%) of these records were submitted to the NLRD by ETQE identifier 1120. These records however constitute less than 1% (0.80%) of the learnership enrolment records submitted to the NLRD by this ETQE. The records for this ETQE are shared by one of two learnerships, with 95.12% of the records belonging to learnership identifier 81 and the remaining 4.88% belonging to learnership identifier 65. Further, slightly more than 86% (86.18%) of the learnership enrolments for this ETQE found in this category, have an END\_DATE\_IND

value of 134 (in other words the vast majority of these learnership enrolments were completed in September 2011).

#### ***4.8.10 Summary of semantic infringements by ETQE***

The preceding sections provide the results of records that infringe on this semantic business rule from the granular perspective of the learnership enrolment record in relation to the complete dataset. This approach supports the determination of patterns within the data that point to systemic and anomalous problems within the overall dataset, which in turn lends itself to assessing the quality of the data in the data set.

The approach however, ignores the diverse nature of ETQEs, and in particular the volume of the records that each ETQE submits to the NLRD. The final step in the analysis of this semantic business rule provides an overview of the percentage of records, calculated as a percentage of the number of records submitted by the ETQE, which infringe on this semantic business rule.

The results are presented as the percentage of records submitted by the ETQE that do not have an associated qualification enrolment record (No Qual Enrolment), the percentage of records submitted by the ETQE that fall into a category that describes a semantic business rule issue ("Lshp Enrolled, Qual Achieved (Derived)", "Lshp Enrolled, Qual Achieved", "Lshp Completed, Qual Enrolled", "Lshp Completed, Qual Enrolled (Derived)", "Lshp Completed Before Qual (Derived)", "Lshp Completed Before Qual", "Lshp Completed After Qual (Derived)" and "Lshp Completed After Qual") and the sum of the percentage of these two broader categories (see Table 4.8.10.1):

Table 4.8.10.1 % of records submitted by an ETQE that do not have an associated qualification enrolment record, % of records submitted by an ETQE that have a category that describes a semantic business rule issue, and the sum percentage of both

ETQE Identifier	% No Qual Enrolment	% Semantic Rule Issue	Total
1123	83.72%	13.47%	97.19%
1104	71.40%	0.10%	71.50%
1121	51.12%	17.71%	68.83%
1115	16.00%	35.19%	51.19%
1105	40.27%	4.84%	45.11%
1124	26.67%	12.22%	38.89%
1117	7.50%	28.06%	35.57%
1113	7.57%	15.29%	22.86%
1116	18.60%	3.06%	21.66%
1126	12.41%	8.12%	20.52%
1107	14.14%	5.11%	19.25%
1120	11.99%	6.25%	18.24%
1103	1.05%	10.18%	11.23%
1111	1.43%	8.81%	10.24%
1102	0.37%	9.62%	9.99%
1127	2.59%	7.03%	9.62%
1122	4.50%	2.77%	7.27%
1125	1.74%	4.14%	5.88%
1119	0.11%	5.52%	5.63%
1109	4.22%	1.27%	5.49%
1114	1.55%	1.89%	3.45%
1106	1.60%	1.81%	3.41%
1112	1.93%	0.40%	2.33%
1110	1.25%	0.54%	1.79%
1108	0.07%	1.58%	1.65%
1075	1.34%	0.07%	1.41%
1118	0.88%	0.35%	1.23%

The results clearly illustrate that the infringement of this semantic business rule could be considered systemic at a number of the ETQEs.

#### **4.8.11 Conclusion**

In order to better understand the results of the analysis of this semantic business rule the context of learnerships and their implementation needs to be elaborated on.

The Skills Development Act, 1998, defines a learnership to mean a structured learning component with practical work experience of a specific nature and duration, registered with the Director General of Labour, which would lead to the achievement of a qualification

registered by SAQA (Ministry in the Office of the President, Skills Development Act, Act 97 of 1998, p. 20).

The definition omits the practical aspects of the implementation of learnerships. The most fundamental aspect of a learnership is a contractual agreement (learnership agreement) between a learner, a provider and an employer in regard to the structured learning component that is to be undertaken. The provider's role is to offer structured learning, whereas the employer's role is to provide the practical work experience. Further, learnerships are generally funded by an ETQE and as a result the learnership agreement has a finite time period.

A learnership is intrinsically linked to a qualification i.e. a learner cannot enrol on a learnership without enrolling on a qualification and a learner cannot complete a learnership without having achieved a qualification. A learnership agreement can however be entered into before the learner enrolls on the learnership and its associated qualification. Although the learnership is intrinsically linked to a qualification, the ETQE that implements the learnership may be different from the ETQE that has been accredited to quality assure the qualification. In other words, in this instance, the learnership enrolment record would be maintained by one ETQE whereas the qualification enrolment record would be maintained by another ETQE.

The funding of learnership agreements and the manner in which the ETQE administers the disbursement of funded learnerships may in some instances create extreme lag times between the start of a learnership agreement and the actual enrolment of the learner, if ever, on the learnership and its associated qualification. In most instances the ETQE learnership disbursement process requires the provision of the detailed learner information prior to the release of funding. These learner details may be collected as part of a tendering process which may take many months, and in some instances years, to be completed. The employer and provider will not proceed with the process of enrolling the learner on the learnership and its associated qualification until initial funding has been received from the ETQE. Often a learner that was initially included in a specific tender is no longer employed by the employer once the tender has been awarded and the initial funding has been disbursed.

Further, a learnership agreement may expire during the time period that the learner is enrolled on the learnership and its associated qualification (i.e. the learner cannot complete the qualification enrolment within the time period stipulated on the agreement and continues the qualification without further practical work experience and funding).

Although learnerships were initially implemented in 1999, the submission of learnership enrolment records to the NLRD was only introduced into the Specifications for Load Files for the National Learners' Records Database in August 2007. It can be reasonably assumed that the operational information system at the ETQE would, up until that point in time, only have focused on the collection of data in regard to the learnership agreement aspect of the learnership enrolment. In other words the operational system would have focused on capturing the fundamental details of the learner, the provider and the employer, and the start and end date of the agreement. It can further be reasonably assumed that operational information system processes would then be developed around these data elements to ensure the accurate disbursement of funds for the learnership enrolment in accordance with the learnership agreement.

The introduction of the submission of learnership enrolment records to the NLRD in August 2007 should have resulted in a change in the operational information system at the ETQE. The learnership agreement record should have been modified to include conceptual data aspects such as the date of enrolment on the learnership (in addition to the start date of the learnership agreement), the completion date of the learnership (in addition to the end date of the learnership agreement) and additional data aspects that would allow for the differentiation between a learnership agreement having expired and the learnership enrolment having been completed. Further, the operational system would have needed to be changed to ensure tight coupling between the learnership enrolment record and the associated qualification enrolment record in order to ensure consistency between the two types of data records.

The final conclusions drawn from the results of the analysis of this semantic business rule must take the context of the learnerships and the implementation of learnerships into consideration.

The 'No Qual Enrolment' category records may exist as a result of the following types of issues:

- The relationship between the learnership and its associated qualification, which is communicated to the NLRD by the ETQE, may have been incorrectly defined by the ETQE.
- The qualification enrolment record may be maintained by an ETQE other than the ETQE that is maintaining the learnership enrolment record. In this instance the ETQE that is maintaining the learnership is submitting the learnership enrolment record to the NLRD whereas the ETQE that is maintaining the qualification enrolment record has not submitted the qualification enrolment record to the NLRD.
- The ETQE prematurely captured the learnership enrolment record on their operational information system in order to initiate the funding process. Once the funding was received by the employer/provider the learner may no longer have been employed at the employer and as a result the learner never enrolled on the associated qualification (or for that matter the learnership).
- The operational information system at the ETQE does not capture the details of the associated qualification enrolment record and as a result the ETQE cannot submit a qualification enrolment record to the NLRD.

The 'Lshp Enrolled, Qual Achieved (Derived)', 'Lshp Enrolled, Qual Achieved', 'Lshp Completed After Qual (Derived)' and 'Lshp Completed After Qual' categories suggest that the learnership is in fact completed but a lack of coupling between the learnership enrolment record and the qualification enrolment record results in the status of the learnership enrolment record not being updated when the qualification enrolment record's status is changed to achieved. Of great concern, should this be the case, is that the same faulty coupling may result in the further funding, by the ETQE, of a learnership that has already been successfully completed.

The 'Lshp Completed, Qual Enrolled', 'Lshp Completed, Qual Enrolled (Derived)', 'Lshp Completed Before Qual (Derived)' and 'Lshp Completed Before Qual' categories suggest that the completion status on the learnership enrolment record is reflecting the expiry of the learnership agreement, rather than the completion of the learnership enrolment. Alternately these categories could be the result of an operational information system, which only allows

the user to update the status of the learnership enrolment record, which lacks the correct coupling between the learnership enrolment record and the qualification enrolment record.

All of the categories that contain the text '(Derived)' in them show that the operational information systems at the respective ETQEs are lacking the required coupling between learnership enrolment records and qualification enrolment records.

Overall the analysis of learnership enrolment records, in regard to whether the intrinsic relationship between the completion of a learnership and achievement of its related qualification has been upheld, highlights the possibility of a number of systemic issues.

The cluster analyses that are conducted on two of the nine categories ('No Qual Enrolment' and 'Lshp Enrolled, Qual Achieved (Derived)') is able to provide a clear description of the data in these categories. The cluster analyses are also able to identify records that may exist in these categories as a result of incorrect data capturing on the learnership enrolment record.

Exploratory data mining is able to provide a clear description of the data in four of the remaining seven categories. The only categories that are too diverse to provide a clear description are the 'Lshp Completed, Qual Enrolled (Derived)', 'Lshp Completed Before Qual (Derived)' and 'Lshp Completed After Qual (Derived)' categories.

Finally, the summary of semantic infringements by ETQE, which shows the percentage of infringements of this semantic business rule calculated as a percentage of the number of learnership enrolment records submitted to the NLRD by the ETQE, shows clear trends of a systemic nature at some ETQEs.

Specific recommendations in regard to data records that are not compliant to this semantic business rule are provided in Appendix P.7.

#### **4.9 Qualification/Unit Standard registration**

This section presents the results of the analysis of learner enrolment records in relation to whether qualification/unit standard was registered for the duration of the learner's active enrolment on the qualification/unit standard. The section therefore focuses on the nominal

data value QUAL\_REGSTR\_IND and USTD\_REGSTR\_IND which contains a value denoting the record's compliance in regard to whether the qualification/unit standard was registered for the duration of the learner's active enrolment on the qualification/unit standard.

This section presents the results of the analysis of these data fields for qualification enrolment records and unit standard enrolment records.

#### 4.9.1 Qualification enrolments

This section presents the results of the analysis of qualification enrolment records in relation to whether the qualification was registered for the duration of the learner's active enrolment on the qualification.

The section therefore focuses on the indicator QUAL\_REGSTR\_IND which, as defined in Appendix E.2, denotes whether the qualification was registered for the duration of the learner's active enrolment on the qualification. The manner in which the categories in this indicator is derived is detailed in Appendix E.3.10. An overview of the derived categories, with QUAL\_REGSTR\_IND \_DESC shown as Description and the frequency of these values as a percentage, is presented in Table 4.9.2.1:

Table 4.9.2.1 Qualification was registered for the duration of the learner's active enrolment on the qualification categories

Description	% Records
OK	93.70%
Start After, End After	0.63%
Start After, End After (Qual Linked to Lshp)	1.19%
Start After, End After Predicted	0.00%
Start After, End After Predicted (Qual Linked to Lshp)	0.01%
Start After, End During	0.52%
Start After, End During (Qual Linked to Lshp)	0.78%
Start After, End During Predicted (Qual Linked to Lshp)	0.00%
Start Before, End After (Qual Linked to Lshp)	0.00%
Start Before, End Before	0.02%
Start Before, End Before (Qual Linked to Lshp)	0.69%
Start Before, End During	0.17%
Start Before, End During (Qual Linked to Lshp)	1.75%
Start During, End After	0.06%
Start During, End After (Qual Linked to Lshp)	0.46%
<b>Total</b>	<b>100.00%</b>

An overview of what the text in these categories denotes is as follows:

- ‘OK’ indicates that the qualification was registered for the duration of the learner’s active enrolment on the qualification,
- ‘Start After’ indicates that the active time period of the qualification enrolment record started after the qualification’s active registration time period,
- ‘Start Before’ indicates that the active time period of the qualification enrolment record started before the qualification’s active registration time period,
- ‘Start During’ indicates that the active time period of the qualification enrolment record started during the qualification’s active registration time period,
- ‘End After’ indicates that the active time period of the qualification enrolment record ended after the qualification’s active registration time period,
- ‘End Before’ indicates that the active time period of the qualification enrolment record ended before the qualification’s active registration time period,
- ‘End During’ indicates that the active time period of the qualification enrolment record ended during the qualification’s active registration time period,
- ‘Predicted’ indicates a current qualification enrolment record that has not yet been achieved and the expected active enrolment on the qualification has not yet expired, and
- ‘(Qual Linked to Lshp)’ indicates that the qualification is linked to a learnership and as a result the problem may be related to the issue described in Section 3.8.3.7.

Any record with a category that ends with the text ‘(Qual Linked to Lshp)’ may have infringed on this semantic business rule as a result of the contextual issues described in Section 3.8.3.7. It was decided, in consultation with the Director of the NLRD, that the further analysis of these types of records will not form part of this research. The percentage of records for this specific issue is however illustrated for all of the semantic business rules in order to provide an understanding of how the specific issue impacts the data in the NLRD.

Any record with a category that ends with the text ‘Predicted’ is a current qualification enrolment. The data of the qualification enrolment record or the data in the Qualification table for these types of records may change before the qualification enrolment record’s active time period expires. As a result of this uncertainty it was decided in consultation with the Director of the NLRD that these types of records will be assumed as correct for the purposes of this research.

As a result the ‘Start After, End After’, ‘Start After, End During’, ‘Start Before, End Before’, ‘Start Before, End During’ and ‘Start During, End After’ categories are considered for this research. Figure 4.9.2.1 presents an overview of the percentage of records that infringe on the semantic business rule that requires that the qualification must be registered for the duration of the learner’s active enrolment on the qualification.

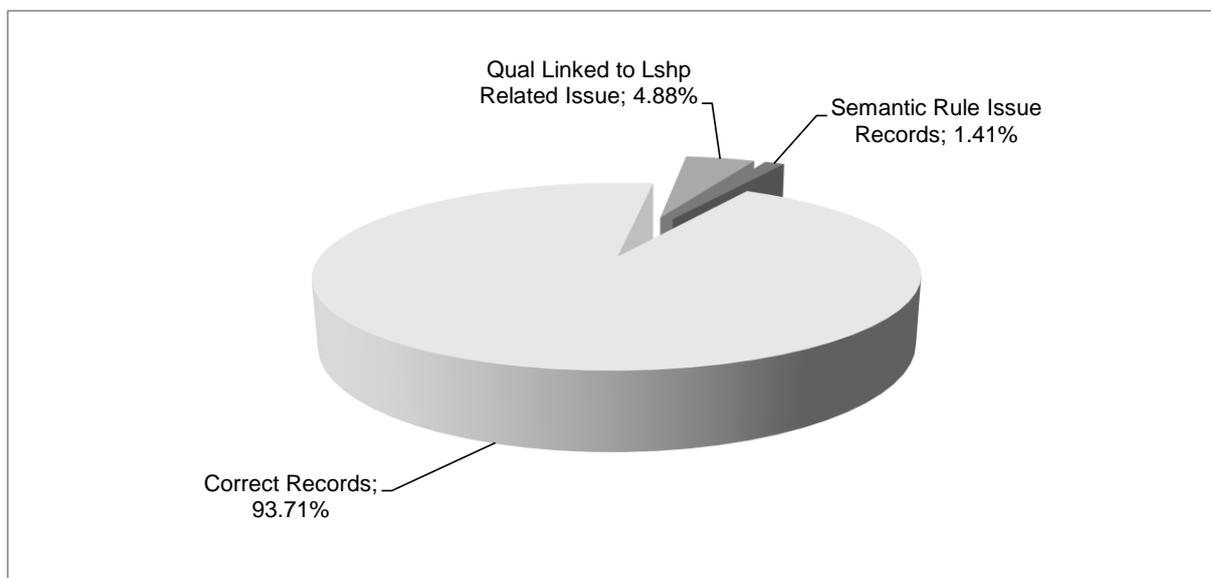


Figure 4.9.2.1 % records according to the semantic business rule that requires that the qualification must be registered for the duration of the learner’s active enrolment on the qualification

The total percentage of records that infringe on this semantic business rule is very low, namely 1.41%. The reader should note that a further 4.88% of the records are identified as having infringed on this semantic business rule, but are excluded from further analysis because these records may have infringed on this semantic business rule as a result of the contextual issues described in Section 3.8.3.7. The records that infringe on this semantic business rule are comprised of 5 categories:

- Start After, End After (45.10%)

This category indicates that the qualification enrolment started after and ended after the qualification’s registration.

Of the 29 discrete ETQEs in the dataset, 4 ETQE's are linked to this category. The majority of these records (99.62%) belong to a single ETQE (ETQE Identifier 1103) and include 13 different qualifications. These records constitute nearly 9% (8.52%) of the records submitted to the NLRD by this ETQE. In this particular incidence it would seem that the ETQE is not aware that these qualifications are no longer registered.

The remaining records (0.38%) found in this category are shared by three different ETQEs across 3 different qualifications. Although these records have a low incidence, these ETQEs may also not be aware that these qualifications are no longer registered.

- Start After, End During (37.26%)

This category indicates that the qualification enrolment started after the last date of enrolment for the qualification and ended before the last date of achievement for the qualification.

Of the 29 discrete ETQEs in the dataset, 5 ETQE's are linked to this category. The majority of these records (96.58%) belong to a single ETQE (ETQE Identifier 1103) and include 14 different qualifications. These records constitute nearly 7% (6.82%) of the records submitted to the NLRD by this ETQE. Further, this is the same ETQE to which the majority of the 'Start After, End After' records, described above, belong to. Additionally 13 of the 14 qualifications also appear in the 'Start After, End After' records for this ETQE.

As is observed in the 'Start After, End After' analysis, in this particular incidence it would seem that the ETQE is not aware that these qualifications are no longer registered.

- Start Before, End During (12.44%)

This category indicates that the qualification enrolment started before the first date on which the qualification was registered and ended during the qualification's active registration period.

Of the 29 discrete ETQEs in the dataset, 16 ETQE's are linked to this category. Of these records, 82.50% were submitted to the NLRD by 3 ETQE's.

Of the 861 discrete qualifications in the dataset, 26 qualifications are linked to this category. Of these 26 qualifications, 10 qualifications contribute to 96.80% of the records. Most notably, although 2 of the 26 qualifications only constitute 1.20% of the records; the records for the qualifications represent 100% of the qualification enrolment records submitted to the NLRD for the qualifications.

The majority of these qualification enrolments (81.88%) have a qualification enrolment start date that precedes the qualification registration start date by no more than one year. This suggests that the learners were enrolled on the qualification whilst the qualification was in the process of being registered. As a result these records are less likely to be in this category as a result of incorrect data. The remaining 18.12% of these records, which have a start date that precedes the qualification registration start date by more than one year, may however be in this category as a result of incorrect data.

- Start During, End After (3.99%)

This category indicates that the qualification enrolment started during the qualification's active registration period and ended after last date of achievement for the qualification.

Of the 29 discrete ETQEs in the dataset, 5 ETQE's are linked to this category. Of these records, 87.52% were submitted to the NLRD by 3 ETQE's.

Of the 861 discrete qualifications in the dataset, 16 qualifications are linked to this category. Of these 16 qualifications, 10 qualifications contribute to 96.43% of the records.

All of these qualification enrolment records have an enrolment status of achieved. The majority of these qualification achievements (66.79%) have a qualification enrolment end date that succeeds the qualification registration end date by less than one year. This suggests that some difficulty was experienced in finalizing the qualification achievements prior to the last date of achievement for the qualification. As a result these records are less likely to be in this category as a result of incorrect data. The

remaining 33.21% of these records, which have an end date that succeeds the qualification's last date of achievement by more than one year, may however be in this category as a result of incorrect data.

- Start Before, End Before (1.20%)

This category indicates that the qualification enrolment started before the qualification's active registration period and ended before last date of achievement for the qualification.

Of the 29 discrete ETQEs in the dataset, 9 ETQE's are linked to this category. Of these records, 91.12% were submitted to the NLRD by 3 ETQE's.

Of the 861 discrete qualifications in the dataset, 12 qualifications are linked to this category. Of these 12 qualifications, 10 qualifications contribute to 98.82% of the records. Most notably, although 2 of the 12 qualifications only constitute 5.92% of the records; the records for the qualifications represent 100% of the qualification enrolment records submitted to the NLRD.

The majority of these records (96.53%) have a qualification start date that precedes the qualification registration start date by more than one year. This suggests that the records that exist in this category are as a result of incorrect data.

Overall the results for this semantic business rule indicate very few issues exist in regard to whether the qualification was registered for the duration of the learner's active enrolment on the qualification. The two largest categories of infringements ('Start After, End After' and 'Start After, End During') show that a systemic issue exists in regard to 13 qualifications that are quality assured by one ETQE. The majority of the records that exist in the 'Start Before, End During' and 'Start During, End After' categories do so within both acceptable and understandable tolerances to the aspect of the semantic business rule constraints. The remaining records either are of such a low volume and/or are so diverse as to suggest that the issues arise as a result of data capturing problems rather than systemic issues with qualification registrations

#### 4.9.2 Unit Standard enrolments

This section presents the results of the analysis of unit standard enrolment records in relation to whether the unit standard was registered for the duration of the learner’s active enrolment on the unit standard.

The section therefore focuses on the indicator USTD\_REGSTR\_IND which, as defined in Appendix G.2, denotes whether the unit standard was registered for the duration of the learner’s active enrolment on the unit standard. The manner in which the categories in this indicator is derived is detailed in Appendix G.3.10. An overview of the derived categories, with USTD\_REGSTR\_IND\_DESC shown as Description and the frequency of these values as a percentage, is presented in Table 4.9.2.1:

Table 4.9.2.1 Unit standard was registered for the duration of the learner’s active enrolment on the unit standard categories

Description	% Records
No Registration (UStd Linked to Lshp)	0.06%
OK	93.45%
Start After, End After	0.82%
Start After, End After (UStd Linked to Lshp)	0.04%
Start After, End During	1.42%
Start After, End During (UStd Linked to Lshp)	0.20%
Start Before, End After	0.00%
Start Before, End After (UStd Linked to Lshp)	0.00%
Start Before, End Before	1.45%
Start Before, End Before (UStd Linked to Lshp)	0.13%
Start Before, End During	1.02%
Start Before, End During (UStd Linked to Lshp)	0.12%
Start During, End After	1.15%
Start During, End After (UStd Linked to Lshp)	0.13%
<b>Total</b>	<b>100.00%</b>

An overview of what the text in these categories denotes is as follows:

- ‘OK’ indicates that the unit standard was registered for the duration of the learner’s active enrolment on the unit standard,
- ‘Start After’ indicates that the active time period of the unit standard enrolment record started after the unit standard’s active registration time period,
- ‘Start Before’ indicates that the active time period of the unit standard enrolment record started before the unit standard’s active registration time period,

- ‘Start During’ indicates that the active time period of the unit standard enrolment record started during the unit standard’s active registration time period,
- ‘End After’ indicates that the active time period of the unit standard enrolment record ended after the unit standard’s active registration time period,
- ‘End Before’ indicates that the active time period of the unit standard enrolment record ended before the unit standard’s active registration time period,
- ‘End During’ indicates that the active time period of the unit standard enrolment record ended during the unit standard’s active registration time period,
- ‘Predicted’ indicates a current unit standard enrolment record that has not yet been achieved and the expected active enrolment on the unit standard has not yet expired, and
- ‘(UStd Linked to Lshp)’ indicates that the unit standard is linked to a learnership and as a result the problem may be related to the issue described in Section 3.8.3.7.

Any record with a category that ends with the text ‘(UStd Linked to Lshp)’ may have infringed on this semantic business rule as a result of the contextual issues described in Section 3.8.3.7. It was decided, in consultation with the Director of the NLRD, that the further analysis of these types of records will not form part of this research. The percentage of records for this specific issue is however illustrated for all of the semantic business rules in order to provide an understanding of how the specific issue impacts the data in the NLRD.

Any record with a category that ends with the text ‘Predicted’ is a current unit standard enrolment. The data of the unit standard enrolment record or the data in the Unit standard table for these types of records may change before the unit standard enrolment record’s active time period expires. As a result of this uncertainty it was decided in consultation with the Director of the NLRD that these types of records will be assumed as correct for the purposes of this research.

As a result the ‘Start After, End After’, ‘Start After, End During’, ‘Start Before, End After’, ‘Start Before, End Before’, ‘Start Before, End During’ and ‘Start During, End After’ categories are considered for this research. Figure 4.9.2.1 presents an overview of the percentage of records that infringe on the semantic business rule that requires that the unit

standard must be registered for the duration of the learner’s active enrolment on the unit standard.

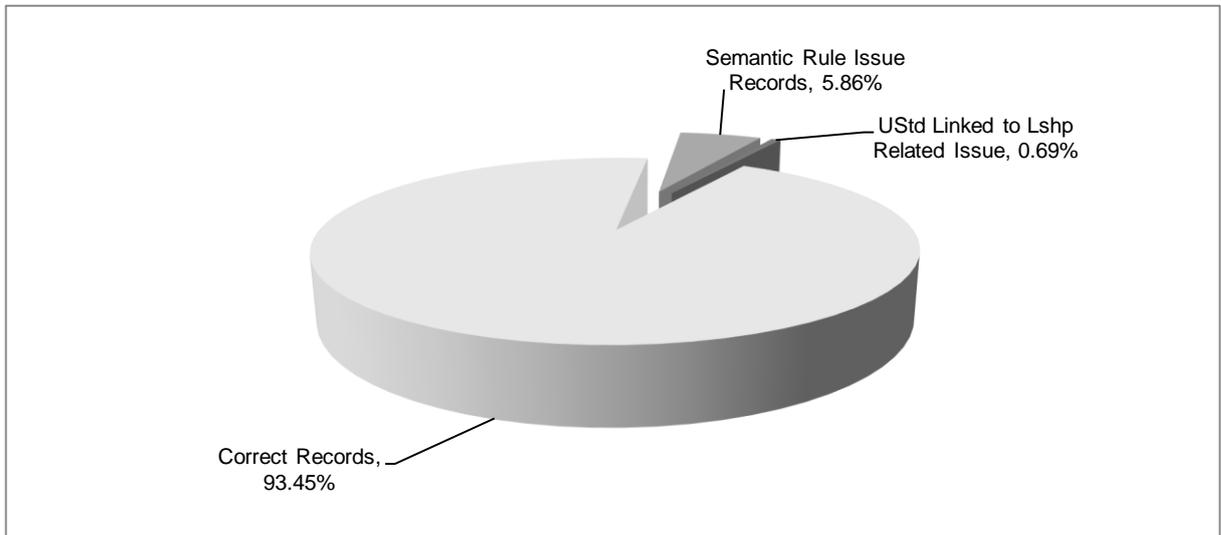


Figure 4.9.2.1 % records according to the semantic business rule that requires that the unit standard must be registered for the duration of the learner’s active enrolment on the unit standard

The total percentage of records that infringe on this semantic business rule is statistically significant, namely 5.86%. The reader should note that a further 0.69% of the records are identified as having infringed on this semantic business rule, but are excluded from further analysis because these records may have infringed on this semantic business rule as a result of the contextual issues described in Section 3.8.3.7. The records that infringe on this semantic business rule are comprised of 6 categories:

- Start Before, End Before (24.70%)

This category indicates that the unit standard enrolment started before the unit standard’s active registration period and ended before last date of achievement for the unit standard.

Of the 29 discrete ETQEs in the dataset, 28 ETQE’s are linked to this category. Of these records, 79.87% were submitted to the NLRD by 3 ETQE’s.

Of the 9124 discrete unit standards in the dataset, 2799 are linked to this category. Of these 2799 unit standards, 10 unit standards contribute to 24.40% of the records. Most

notably, although 7 of the 2799 unit standards constitute less than 0.01% of the records; the records for the unit standards represent 100% of the unit standard enrolment records submitted to the NLRD.

The majority of these unit standard enrolments (59.51%) have a unit standard enrolment start date that precedes the unit standard registration start date by no more than one year. This suggests that the learners were enrolled on the unit standard whilst this was in the process of being registered. As a result these records are less likely to be in this category as a result of incorrect data. The remaining 40.49% of these records, which have a start date that precedes the unit standard registration start date by more than one year, may however be in this category as a result of incorrect data.

- Start After, End During (24.31%)

This category indicates that the unit standard enrolment started after the last date of enrolment for the unit standard and ended before the last date of achievement for the unit standard.

Of the 29 discrete ETQEs in the dataset, 27 ETQE's are linked to this category. Of these records, 60.93% were submitted to the NLRD by 3 ETQE's.

Of the 9124 discrete unit standards in the dataset, 1726 are linked to this category. Of these 1726 unit standards, 10 unit standards contribute to 15.95% of the records. Most notably, although 13 of the 1726 unit standards constitute less than 0.01% of the records; the records for the unit standards represent 100% of the unit standard enrolment records submitted to the NLRD.

- Start During, End After (19.62%)

This category indicates that the unit standard enrolment started during the unit standard's active registration period and ended after last date of achievement for the unit standard.

Of the 29 discrete ETQEs in the dataset, 20 ETQE's are linked to this category. Of these records, 85.72% were submitted to the NLRD by 3 ETQE's.

Of the 9124 discrete unit standards in the dataset, 1135 are linked to this category. Of these 1135 unit standards, 10 unit standards contribute to 37.29% of the records.

- Start Before, End During (17.35%)

This category indicates that the unit standard enrolment started before the first date on which the unit standard was registered and ended during the unit standard's active registration period.

Of the 29 discrete ETQEs in the dataset, 27 ETQE's are linked to this category. Of these records, 71.52% were submitted to the NLRD by 3 ETQE's.

Of the 9124 discrete unit standards in the dataset, 2997 are linked to this category. Of these 2997 unit standards, 10 unit standards contribute to 7.02% of the records.

- Start After, End After (13.98%)

This category indicates that the unit standard enrolment started after and ended after the unit standard's registration.

Of the 29 discrete ETQEs in the dataset, 26 ETQE's are linked to this category. Of these records, 65.49% were submitted to the NLRD by 3 ETQE's.

Of the 9124 discrete unit standards in the dataset, 1382 are linked to this category. Of these 1382 unit standards, 10 unit standards contribute to 25.49% of the records. Most notably, although 6 of the 1382 unit standards constitute less than 0.01% of the records; the records for the unit standards represent 100% of the unit standard enrolment records submitted to the NLRD.

- Start Before, End After (0.04%)

This category indicates that the unit standard enrolment started before the unit standard's active registration period and ended after last date of achievement for the unit standard.

Of the 29 discrete ETQEs in the dataset, 5 ETQE's are linked to this category. Of these records, 96.26% were submitted to the NLRD by 3 ETQE's.

Of the 9124 discrete unit standards in the dataset, 91 are linked to this category. Of these 91 unit standards, 10 unit standards contribute to 63.20% of the records.

The majority of these records (23.52%) have a unit standard start date that precedes the unit standard registration start date by more than one year. This suggests that the records that exist in this category are as a result of incorrect data.

Overall the results for this semantic business rule indicate that significant issues exist in regard to whether the unit standard was registered for the duration of the learner's active enrolment on the unit standard. On review of the results at a unit standard level the records either are of such a low volume and/or are so diverse as to suggest that the issues arise as a result of data capturing problems rather than systemic issues with unit standard registrations

#### **4.9.3 Conclusion**

This section focuses on the analysis of learner enrolment records in relation to whether the qualification/unit standard was registered for the duration of the learner's active enrolment on the qualification/unit standard. The section therefore focuses on the nominal data value `QUAL_REGSTR_IND` and `USTD_REGSTR_IND` which contains a value denoting the record's compliance in regard to whether the qualification/unit standard was registered for the duration of the learner's active enrolment on the qualification/unit standard.

Overall the results for this semantic business rule indicate very few issues exist in regard to whether the qualification/unit standard was registered at the time of the learner's enrolment on the qualification/unit standard. The analysis highlights the following:

- ETQE Identifier 1103 does not seem to be aware that fourteen of the qualifications that it is quality assuring are no longer registered, and
- the low volume and/or diversity in qualifications/unit standards suggest that the issues arise as a result of data capturing problems rather than systemic issues with qualification registrations

Specific recommendations in regard to data records that are not compliant to this semantic business rule are provided in Appendix P.8.1 for qualification enrolments and Appendix P.8.2 for unit standard enrolments.

#### **4.10 Unit Standard based qualification achievements**

This section presents the results of the analysis of qualification enrolment records in relation to whether, in the case where the learner has achieved the qualification, and the qualification is a unit standards based qualification, the learner has achieved the correct number and mix of credits for the qualification. The reader should note that this specific semantic business rule is only applicable to qualification enrolment records.

The section therefore focuses on the indicator `UNIT_STD_MIX_IND` which, as defined in Appendix E.2, denotes whether the learner achieved:

- the minimum required number of credits for the qualification, on or before the achievement of the qualification, based on the achievement of unit standards related to the qualification, and
- the correct range of credits for the qualification, achieved on or before the achievement of the qualification, based on the achievement of unit standards that have been defined as core, fundamental and elective unit standards for the qualification.

The manner in which the categories in this indicator is derived is detailed in Appendix E.3.11. An overview of the derived categories, with `UNIT_STD_MIX_IND_DESC` shown as Description and the frequency of these values as a percentage, is presented in Table 4.10.1:

Table 4.10.1 Learner has achieved the correct number and mix of credits for the qualification categories

Description	% Records
Insufficient Credits Achieved, Core Credits OK, Fundamental Credits OK, Insufficient Elective Credits	0.87%
Insufficient Credits Achieved, Core Credits OK, Fundamental Credits OK, Insufficient Elective Credits (Qual Linked to Lshp)	0.36%
Insufficient Credits Achieved, Core Credits OK, Insufficient Fundamental Credits, Elective Credits OK	1.20%
Insufficient Credits Achieved, Core Credits OK, Insufficient Fundamental Credits, Elective Credits OK (Qual Linked to Lshp)	0.04%
Insufficient Credits Achieved, Core Credits OK, Insufficient Fundamental Credits, Insufficient Elective Credits	0.49%
Insufficient Credits Achieved, Core Credits OK, Insufficient Fundamental Credits, Insufficient Elective Credits (Qual Linked to Lshp)	0.01%
Insufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK	2.99%
Insufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK (Qual Linked to Lshp)	0.07%
Insufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Insufficient Elective Credits	1.03%
Insufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Insufficient Elective Credits (Qual Linked to Lshp)	0.02%
Insufficient Credits Achieved, Insufficient Core Credits, Insufficient Fundamental Credits, Elective Credits OK	2.37%
Insufficient Credits Achieved, Insufficient Core Credits, Insufficient Fundamental Credits, Elective Credits OK (Qual Linked to Lshp)	0.07%
Insufficient Credits Achieved, Insufficient Core Credits, Insufficient Fundamental Credits, Insufficient Elective Credits	0.87%
Insufficient Credits Achieved, Insufficient Core Credits, Insufficient Fundamental Credits, Insufficient Elective Credits (Qual Linked to Lshp)	0.09%
No Unit Standards Achieved	6.57%
No Unit Standards Achieved (Qual Linked to Lshp)	0.30%
Not Achieved	41.59%
Not Unit Standard Based	15.75%
Sufficient Credits Achieved	22.36%
Sufficient Credits Achieved, Core Credits OK, Fundamental Credits OK, Insufficient Elective Credits	0.00%
Sufficient Credits Achieved, Core Credits OK, Fundamental Credits OK, Insufficient Elective Credits (Qual Linked to Lshp)	0.13%
Sufficient Credits Achieved, Core Credits OK, Insufficient Fundamental Credits, Elective Credits OK	1.56%
Sufficient Credits Achieved, Core Credits OK, Insufficient Fundamental Credits, Elective Credits OK (Qual Linked to Lshp)	0.03%
Sufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK	0.83%
Sufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK (Qual Linked to Lshp)	0.02%
Sufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Insufficient Elective Credits (Qual Linked to Lshp)	0.00%
Sufficient Credits Achieved, Insufficient Core Credits, Insufficient Fundamental Credits, Elective Credits OK	0.53%
Sufficient Credits Achieved, Insufficient Core Credits, Insufficient Fundamental Credits, Elective Credits OK (Qual Linked to Lshp)	0.06%
<b>Total</b>	<b>100.00%</b>

An overview of what the text in these categories denotes is as follows:

- ‘Insufficient Credits Achieved’ indicates that the minimum number of credits for the qualification were not achieved on or before the achievement of the qualification,
- ‘Insufficient [Type of Credits] Credits’ indicates that the minimum number of a specific type of credit for the qualification was not achieved on or before the achievement of the qualification, where [Type of Credits] is ‘Core’, ‘Fundamental’ or ‘Elective’,
- ‘[Type of Credits] Credits OK’, indicates that the minimum number of a specific type of credit for the qualification was achieved on or before the achievement of the qualification, where [Type of Credits] is ‘Core’, ‘Fundamental’ or ‘Elective’,
- ‘No Unit Standards Achieved’ indicates that no credits for the qualification were achieved on or before the achievement of the qualification,
- ‘Not Achieved’ indicates that the qualification has not been achieved,
- ‘Not Unit Standard Based’ indicates that the qualification is not a unit standard based qualification,
- ‘Sufficient Credits Achieved’ indicates that the minimum number of credits for the qualification were achieved on or before the achievement of the qualification, and

- '(Qual Linked to Lshp)' indicates that the qualification is linked to a learnership and as a result the problem may be related to the issue described in Section 3.8.3.7.

Any record with a category that ends with the text '(Qual Linked to Lshp)' may have infringed on this semantic business rule as a result of the contextual issues described in Section 3.8.3.7. It was decided, in consultation with the Director of the NLRD, that the further analysis of these types of records will not form part of this research. The percentage of records for this specific issue is however illustrated for all of the semantic business rules in order to provide an understanding of how the specific issue impacts the data in the NLRD.

As a result the 'Insufficient Credits Achieved, Core Credits OK, Fundamental Credits OK, Insufficient Elective Credits', 'Insufficient Credits Achieved, Core Credits OK, Insufficient Fundamental Credits, Elective Credits OK', 'Insufficient Credits Achieved, Core Credits OK, Insufficient Fundamental Credits, Insufficient Elective Credits', 'Insufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK', 'Insufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Insufficient Elective Credits', 'Insufficient Credits Achieved, Insufficient Core Credits, Insufficient Fundamental Credits, Elective Credits OK', 'Insufficient Credits Achieved, Insufficient Core Credits, Insufficient Fundamental Credits, Insufficient Elective Credits', 'No Unit Standards Achieved', 'Sufficient Credits Achieved, Core Credits OK, Fundamental Credits OK, Insufficient Elective Credits', 'Sufficient Credits Achieved, Core Credits OK, Insufficient Fundamental Credits, Elective Credits OK', 'Sufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK' and 'Sufficient Credits Achieved, Insufficient Core Credits, Insufficient Fundamental Credits, Elective Credits OK' categories are considered for this research. Figure 4.10.1 presents an overview of the percentage of records that infringe on the semantic business rule that requires that in the case where the learner has achieved the qualification, and the qualification is a unit standards based qualification, the learner has achieved the correct number and mix of credits for the qualification.

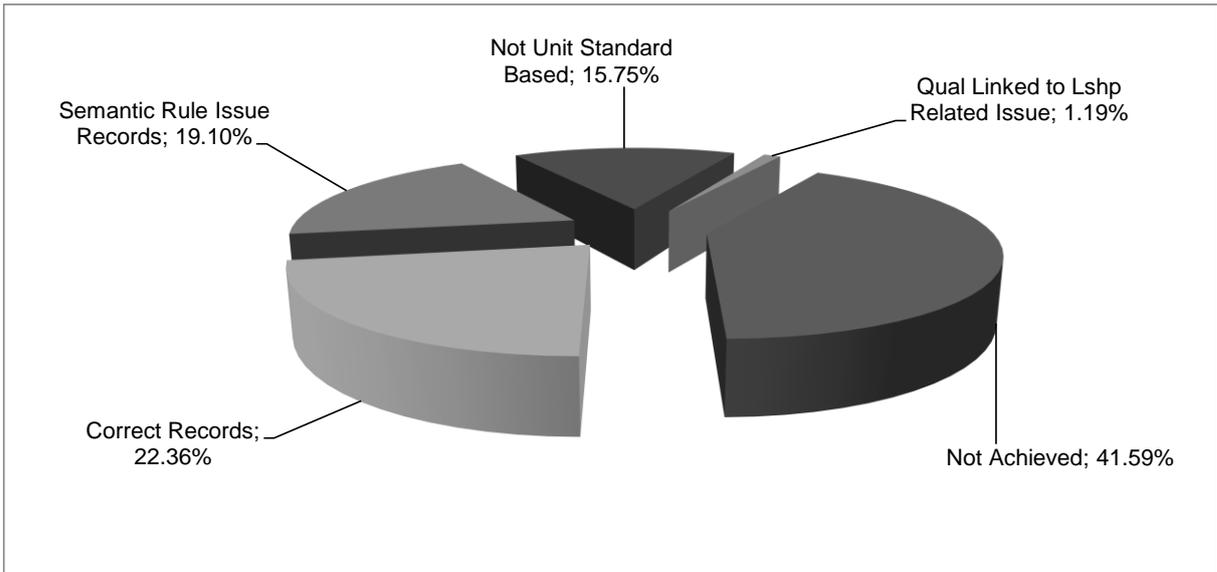


Figure 4.10.1 % records according to the semantic business rule that requires that in the case where the learner has achieved the qualification, and the qualification is a unit standards based qualification, the learner has achieved the correct number and mix of credits for the qualification

The total percentage of records that infringe on this semantic business rule is high, namely 19.10%. The reader should note that a further 1.19% of the records are identified as having infringed on this semantic business rule, but are excluded from further analysis because these records may have infringed on this semantic business rule as a result of the contextual issues described in Section 3.8.3.7. Figure 4.10.2 provides an overview of the percentage of records found in each of these categories:

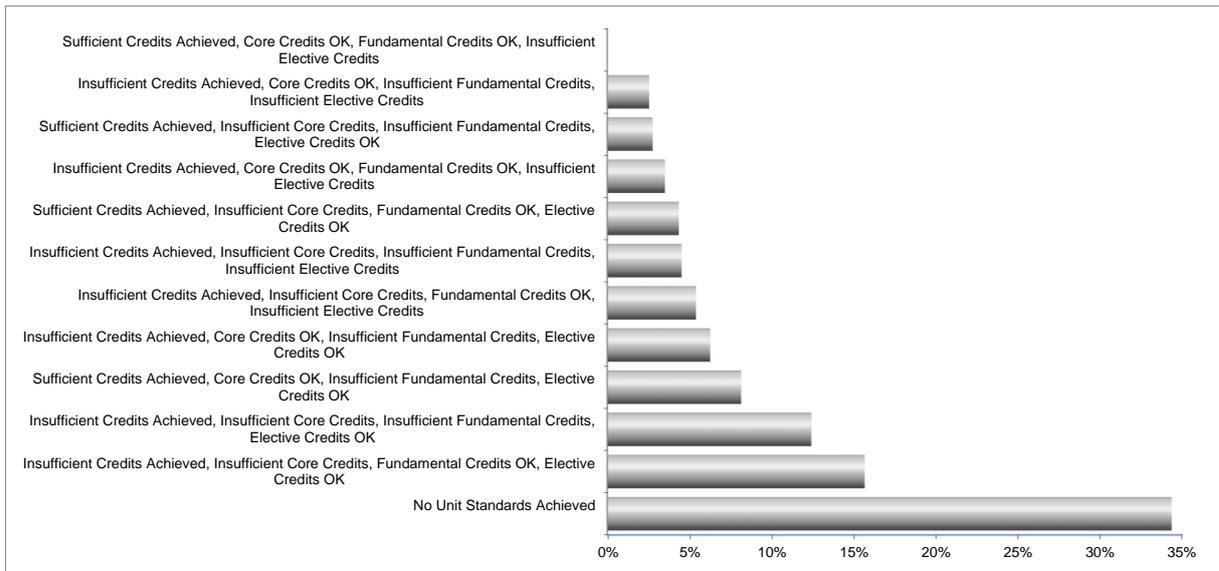


Figure 4.10.2 % records that infringe the semantic business rule that requires that in the case where the learner has achieved the qualification, and the qualification is a unit standards based qualification, the learner has achieved the correct number and mix of credits for the qualification

An analysis of the results shown in Figure 4.10.2 shows that the categories that describe records that infringe on this semantic business rule can be grouped into three groups:

- **Insufficient Unit Standard Credits Achieved (50.36%)**  
This group contains all categories that indicate that the learner has not achieved the correct number of credits for the qualification.
- **No Unit Standard Credits Achieved (34.37%)**  
This group contains the category ‘No Unit Standards Achieved’
- **Incorrect Mix of Unit Standard Credits Achieved (15.28%)**  
This group contains all categories that indicate that the learner has achieved the correct number of credits for the qualification, but the learner has failed to achieve the correct mix of credits for the qualification.

The following sections provide a more in-depth analysis of these three groups.

#### ***4.10.1 Insufficient Unit Standard Credits Achieved***

This group includes the following categories that indicate that the learner has not achieved the correct number of credits for the qualification:

- **Insufficient Credits Achieved, Core Credits OK, Fundamental Credits OK, Insufficient Elective Credits**

- Insufficient Credits Achieved, Core Credits OK, Insufficient Fundamental Credits, Elective Credits OK
- Insufficient Credits Achieved, Core Credits OK, Insufficient Fundamental Credits, Insufficient Elective Credits
- Insufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK
- Insufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Insufficient Elective Credits
- Insufficient Credits Achieved, Insufficient Core Credits, Insufficient Fundamental Credits, Elective Credits OK
- Insufficient Credits Achieved, Insufficient Core Credits, Insufficient Fundamental Credits, Insufficient Elective Credits

Of the 29 discrete ETQEs in the dataset, 26 ETQEs are linked to this group. Of these records, 51.04% were submitted to the NLRD by 3 ETQEs.

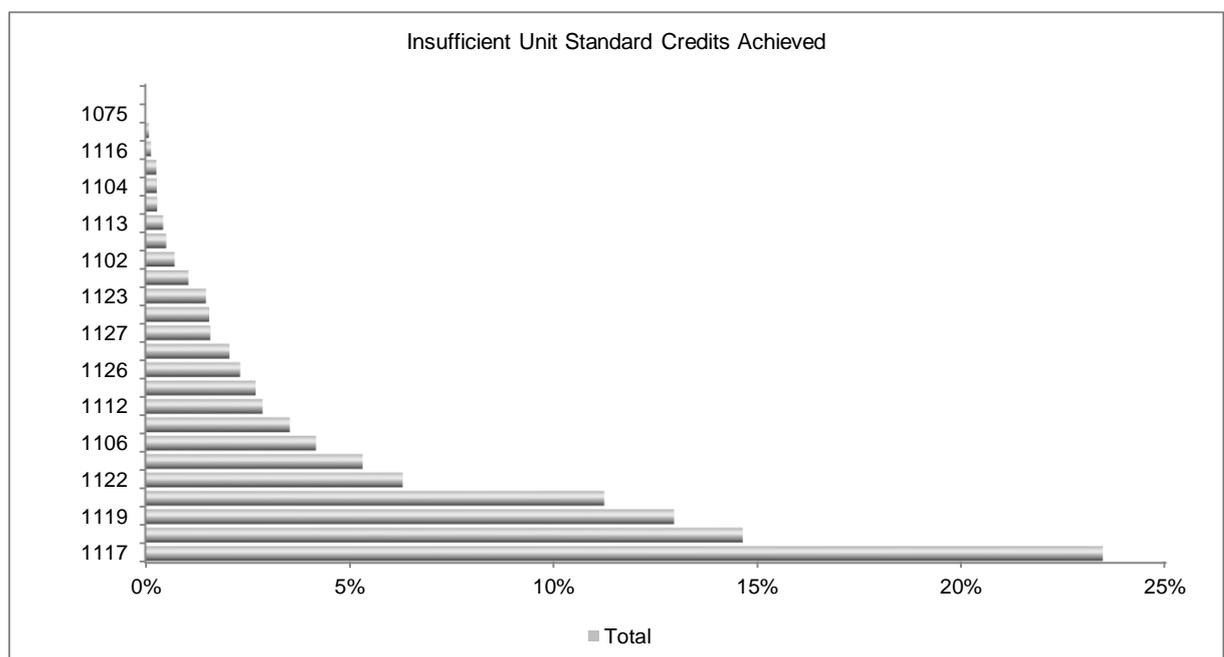


Figure 4.10.1.1 % records by ETQE where the learner has not achieved the correct number of credits for the qualification

Of the 861 discrete qualifications in the dataset, 390 qualifications are linked to this group. Of these 390 qualifications, 10 qualifications contribute to 48.66% of records in this

category. Most notably, 38 of the 390 qualifications contribute 4.04% of the records in this group; the records for these qualifications represent 100% of the achieved qualification enrolment records submitted to the NLRD for the qualification.

As indicated in the previous section this group contains 50.36% of the qualification enrolment records that infringe on this semantic business rule. The implementation of data mining techniques, beyond exploratory data mining techniques, is needed to provide SAQA with both a succinct description of these records and to attempt to identify any data records that exist in this category as a result of possible incorrect data capturing at the source of the data.

The clustering data mining technique, as described in Appendix I.3, is applied to this data set in an effort to achieve both these aims.

The data mining effort results in the development of 8 data clusters (a technical description of each cluster is provided in Appendix O.1) that have the following most dominant characteristics:

1. Cluster 1

This cluster describes slightly more than 20% of the records in this group. The cluster is relatively diverse and describes 71 qualifications that were offered by 95 providers and assessed by 90 assessors. The qualifications predominantly have subfield descriptions of 'Manufacturing and Assembly' and 'Fabrication and Extraction' and were submitted to the NLRD by ETQE identifiers 1111, 1107 and 1103.

2. Cluster 2

The cluster describes nearly 19% of the records in this group and is diverse. None of the qualification enrolment records in this cluster are linked to an assessor. The cluster describes 59 qualifications offered by 137 providers. The records were submitted to the NLRD by 10 ETQEs.

3. Cluster 3

This cluster describes slightly more than 13% of the records in this group. The cluster describes enrolments against 3 qualifications (qualification identifiers 24214, 22507 and 20513), all with a subfield description of 'Safety in Society'. These qualifications were offered by 18 providers and were submitted to the NLRD by ETQE identifier 1105.

4. Cluster 4

The cluster describes nearly 13% of the records in this group as having been submitted to the NLRD by ETQE identifier 1117. The cluster constitutes 1 qualification (qualification identifier 49623) offered by 26 providers. Further analysis shows that these records for qualification identifier 49623 represent nearly 76% of the qualification enrolment records for this qualification.

5. Cluster 5

This cluster describes nearly 12% of the records in this group. The cluster is relatively diverse in that it describes 38 qualifications as offered by 105 providers. The qualifications predominantly have a NQF level description of Level 4 and were submitted to the NLRD by 11 ETQEs.

6. Cluster 6

The cluster describes more than 10% of the records found in this group as having been submitted to the NLRD by ETQE identifier 1119. These records comprise qualification enrolment records for 9 qualifications as offered by 31 providers. None of these enrolments are linked to a learnership enrolment. All of the qualifications have a subfield description of 'Hospitality, Tourism, Travel, Gaming and Leisure' and constitute 33% of the qualifications that the ETQE has submitted qualification enrolments against to the NLRD.

7. Cluster 7

This cluster describes slightly more than 7.5% of the records in this group. The cluster is relatively diverse in that it describes 12 qualifications as offered by 46 providers. None of these enrolments are linked to a learnership enrolment. The subfield description for the majority of these qualifications is 'Safety in Society', 'Early Childhood Development' and 'Adult Learning'. The records were submitted to the NLRD by ETQE identifiers 1106 and 1105.

8. Cluster 8

The cluster describes nearly 5.5% of the records found in this group as having been submitted to the NLRD by ETQE identifier 1117. These records constitute qualification enrolment records for 4 qualifications, all with a subfield description of 'Promotive Health and Developmental Services' or 'Curative Health'. Further, the majority of these records have a NQF level description of Level 1.

The most notable clusters that are generated for this group are clusters 1, 3, 4, 6, 7 and 8. Clusters 3, 4 and 8 seem to describe specific problems with the implementation of specific

qualifications whereas Clusters 1, 6 and 7 seem to describe systemic problems arising at the level of the ETQE.

The clustering algorithm does not generate any clusters with a population of less than 1%. A review of the probabilities allocated to each record in a cluster does however allow for the identification of records that are low enough to be considered anomalous as per the criteria defined in Appendix I.3. These records constitute 1.25% of the records found in this group, and possibly exist in this group as a result of data capturing problems at the source of the data.

#### ***4.10.2 No Unit Standard Credits Achieved***

This group contains the category ‘No Unit Standards Achieved’ and contains records where:

- the learner has achieved the qualification,
- the qualification is a unit standards based qualification,
- the learner has not achieved any credits for the qualification.

This group of records is of greatest concern to SAQA.

Of the 29 discrete ETQEs in the dataset, 26 ETQEs are linked to this group. Of these records, 50.38% were submitted to the NLRD by 3 ETQEs.

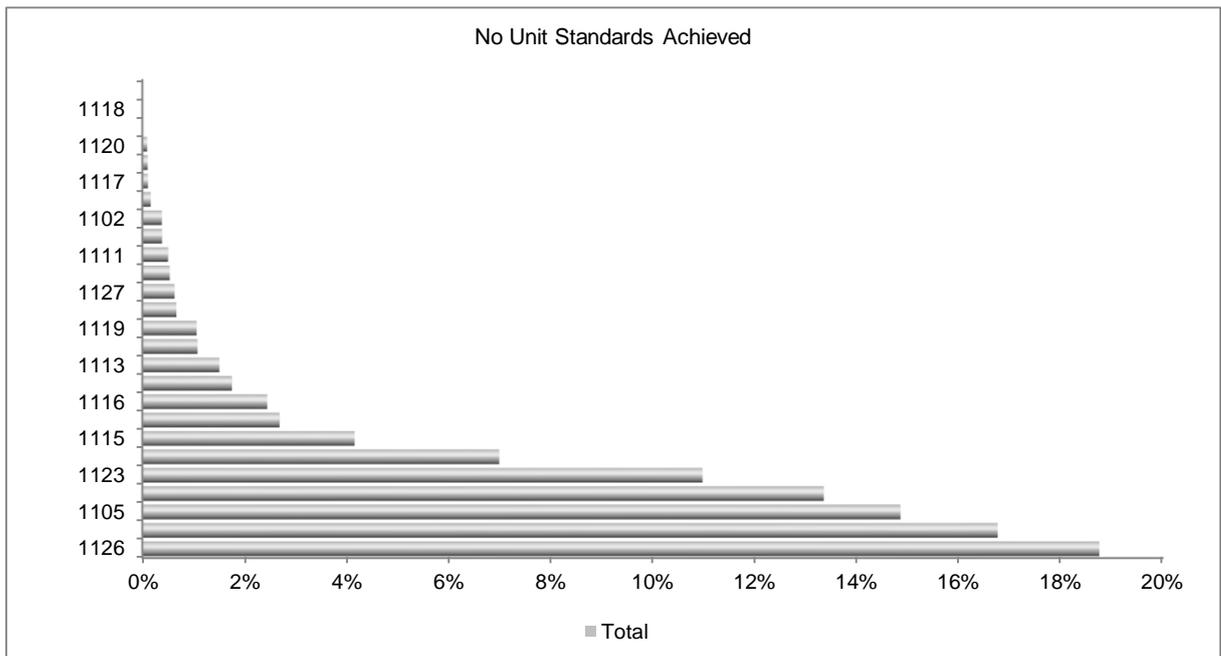


Figure 4.10.2.1 % records by ETQE where the learner has achieved the qualification, the qualification is a unit standards based qualification and the learner has not achieved any credits for the qualification

Of the 861 discrete qualifications in the dataset, 367 qualifications are linked to this group. Of these 367 qualifications, 10 qualifications contribute to 55.84% of records in this category. Most notably, 40 of the 367 qualifications contribute 7.98% of the records in this group; the records for these qualifications represent 100% of the achieved qualification enrolment records submitted to the NLRD for the qualification.

As indicated in the previous section this group contains 34.37% of the qualification enrolment records that infringe on this semantic business rule. The implementation of data mining techniques, beyond exploratory data mining techniques, is needed to provide SAQA with both a succinct description of these records and to attempt to identify any data records that exist in this category as a result of possible incorrect data capturing at the source of the data.

The clustering data mining technique, as described in Appendix I.3, is applied to this data set in an effort to achieve both these aims.

The data mining effort results in the development of 8 data clusters (a technical description of each cluster is provided in Appendix O.2) that have the following most dominant characteristics:

1. Cluster 1

This cluster describes slightly more than 32% of the records in this group. The cluster is relatively diverse and describes 35 qualifications that were offered by 94 providers. The qualifications predominantly have subfield descriptions of 'Manufacturing and Assembly' and 'Information Technology and Computer Sciences' and were submitted to the NLRD by 5 different ETQEs.

2. Cluster 2

The cluster describes slightly more than 30% of the records in this group as belonging to 7 qualifications that have a subfield description of 'Safety in Society' or 'Early Childhood Development'. These qualifications were offered by 32 providers and the enrolment records were submitted to the NLRD by ETQE identifiers 1105 and 1106.

3. Cluster 3

This cluster describes slightly more than 10.5% of the records in this group. The cluster describes enrolments against 4 qualifications (qualification identifiers 73286, 72027, 23671 and 21810), all with a subfield description of 'Marketing' or 'Generic Management'. These qualifications were offered by 21 providers and were submitted to the NLRD by ETQE identifier 1126.

4. Cluster 4

The cluster describes more than 8% of the records in this group and is diverse in that these records were submitted to the NLRD by 9 ETQEs. The cluster describes enrolments against 23 qualifications offered by 53 providers.

5. Cluster 5

This cluster describes 7.5% of the records in this group. The cluster is relatively diverse in that it describes 24 qualifications as offered by 51 providers. These enrolment records were submitted to the NLRD by ETQE identifiers 1126 and 1108.

6. Cluster 6

The cluster describes more than 4.5% of the records found in this group as having been submitted to the NLRD by ETQE identifier 1108. These records comprise qualification enrolment records for 9 qualifications as offered by 22 providers. The majority of these qualifications have a NQF Level description of Level 4.

7. Cluster 7

This cluster describes nearly 4% of the records in this group. The cluster is relatively diverse in that it describes 10 qualifications as offered by 43 providers. The records were submitted to the NLRD by ETQE identifiers 1112, 1111 and 1108.

#### 8. Cluster 8

The cluster describes slightly more than 2.5% of the records found in this group as having been submitted to the NLRD by ETQE identifier 1126. These records constitute qualification enrolment records for 4 qualifications, all with a subfield description of 'Consumer Services' or 'Cleaning, Domestic, Hiring, Property and Rescue Services'. Further, the majority of these records have a NQF level description of Level 4.

The most notable clusters that are generated for this group are clusters 2, 3, 5, 6, 7 and 8. Clusters 2, 6, 7, 8 seem to describe specific problems with the implementation of specific qualifications whereas Clusters 3 and 5 seem to describe systemic problems arising at the level of the ETQE. The recurrence of ETQE identifiers 1108 and 1126 in more than one of the clusters may however point specifically at a systemic problem related to these ETQEs.

The clustering algorithm does not generate any clusters with a population of less than 1%. A review of the probabilities allocated to each record in a cluster does however allow for the identification of records that are low enough to be considered anomalous as per the criteria defined in Appendix I.3. These records constitute 1.76% of the records found in this group, and possibly exist in this group as a result of data capturing problems at the source of the data.

#### ***4.10.3 Incorrect Mix of Unit Standard Credits Achieved***

This group includes the following categories that indicate that even though the learner has achieved the correct number of credits for the qualification, the number of credits derived from core, fundamental or elective unit standards is incorrect:

- Sufficient Credits Achieved, Core Credits OK, Fundamental Credits OK, Insufficient Elective Credits
- Sufficient Credits Achieved, Core Credits OK, Insufficient Fundamental Credits, Elective Credits OK
- Sufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK

- Sufficient Credits Achieved, Insufficient Core Credits, Insufficient Fundamental Credits, Elective Credits OK

Of the 29 discrete ETQEs in the dataset, 24 ETQEs are linked to this group. Of these records, 41.85% were submitted to the NLRD by 3 ETQEs.

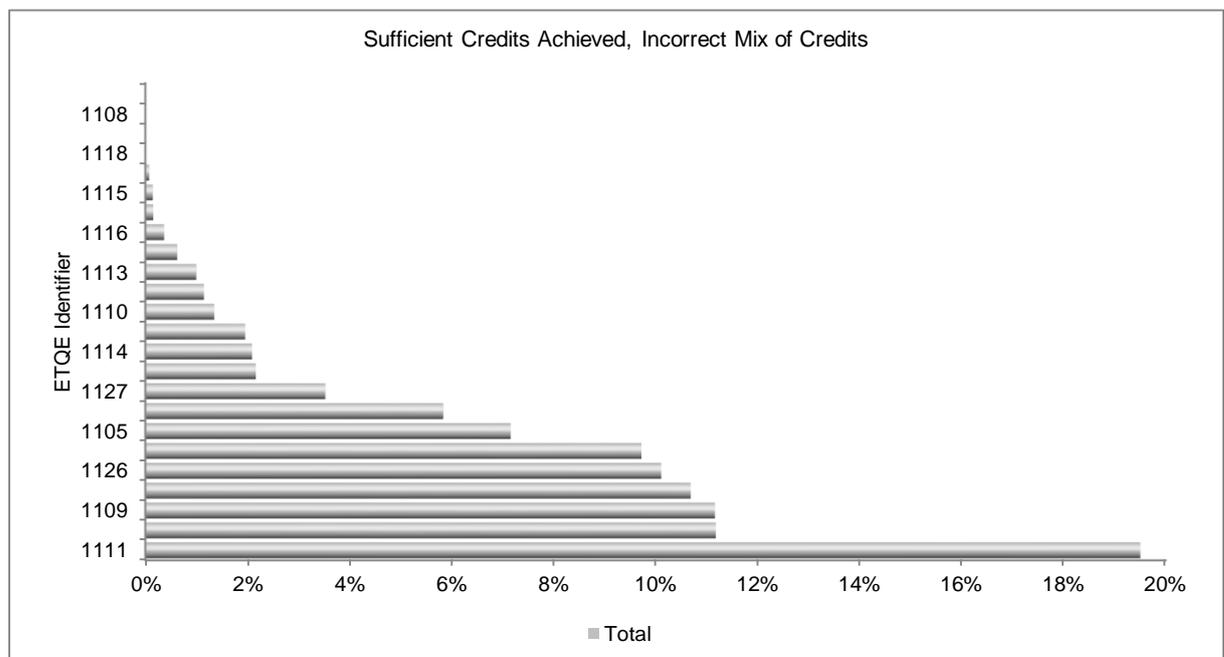


Figure 4.10.3.1 % records by ETQE where even though the learner has achieved the correct number of credits for the qualification, the number of credits derived from core, fundamental or elective unit standards is incorrect

Of the 861 discrete qualifications in the dataset, 247 qualifications are linked to this group. Of these 247 qualifications, 10 qualifications contribute to 46.20% of records in this category. Most notably, 7 of the 247 qualifications contribute 0.82% of the records in this group; the records for these qualifications represent 100% of the achieved qualification enrolment records submitted to the NLRD for the qualification.

As indicated in the previous section this group contains 15.28% of the qualification enrolment records that infringe on this semantic business rule. The implementation of data mining techniques, beyond exploratory data mining techniques, is needed to provide SAQA with both a succinct description of these records and to attempt to identify any data records

that exist in this category as a result of possible incorrect data capturing at the source of the data.

The clustering data mining technique, as described in Appendix I.3, is applied to this data set in an effort to achieve both these aims.

The data mining effort results in the development of 8 data clusters (a technical description of each cluster is provided in Appendix O.3) that have the following most dominant characteristics:

1. Cluster 1

This cluster describes nearly 39% of the records in this group. The cluster is relatively diverse in that it describes 25 qualifications offered by 61 providers. The majority of the qualifications in this group have a field description of 'Manufacturing, Engineering and Technology'. The qualification enrolment records were submitted to the NLRD by 4 ETQEs (ETQE identifiers 1112, 1111, 1107 and 1103).

2. Cluster 2

The cluster describes nearly 15.5% of the records in this group as belonging to 7 qualifications. These qualification enrolment records were submitted to the NLRD by three ETQEs (ETQE identifiers 1126, 1123 and 1106). The majority of these records have a USTD\_MIX\_IND description of 'Sufficient Credits Achieved, Core Credits OK, Insufficient Fundamental Credits, Elective Credits OK'.

3. Cluster 3

This cluster describes slightly more than 15% of the records in this group. The cluster describes 7 qualifications as offered by 35 providers. The qualification enrolment records were submitted to the NLRD by ETQE 1112 and 1109.

4. Cluster 4

The cluster describes slightly more than 8% of the records in this group as belonging to 5 qualifications. These qualification enrolment records were submitted to the NLRD by four ETQEs (ETQE identifiers 1114, 1106, 1105 and 1104). The majority of these records have a USTD\_MIX\_IND description of 'Sufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK'.

5. Cluster 5

This cluster describes nearly 7.5% of the records in this group as belonging to one qualification (qualification identifier 78981) as offered by 6 providers. All of the

qualification enrolment records were submitted to the NLRD by ETQE identifier 1123 and the USTD\_MIX\_IND description of these enrolment records is 'Sufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK'.

6. Cluster 6

The cluster describes slightly more than 6% of the records in this group as belonging to 15 qualifications. These qualification enrolment records were submitted to the NLRD by five ETQEs (ETQE identifiers 1127, 1126, 1120, 1116 and 1110). The majority of these qualifications have a field description of 'Business, Commerce and Management Studies'.

7. Cluster 7

This cluster describes slightly more than 5% of the records in this group as having been submitted to the NLRD by ETQE identifier 1112. The cluster contains 6 different qualifications all with a subfield description of 'Primary Agriculture'. The majority of these records have a USTD\_MIX\_IND description of 'Sufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK'.

8. Cluster 8

The cluster describes nearly 4% of the records in this group as having been submitted to the NLRD by ETQE 1113 and 1105. These records belong to 6 qualifications that have a field description of 'Law, Military Science and Security'. These majority of these qualification enrolment records have a USTD\_MIX\_IND description of 'Sufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK'.

The most notable clusters that are generated for this group are clusters 3, 5, 7 and 8. All of these clusters seem to describe specific problems with the implementation of specific qualifications. The recurrence of ETQE identifiers 1105, 1106, 1112, 1123 and 1126 in more than one of the clusters may however point specifically at a systemic problem related to these ETQEs.

The clustering algorithm does not generate any clusters with a population of less than 1%. A review of the probabilities allocated to each record in a cluster does however allow for the identification of records that are low enough to be considered anomalous as per the criteria defined in Appendix I.3. These records constitute 3.73% of the records found in this

group, and possibly exist in this group as a result of data capturing problems at the source of the data.

#### ***4.10.4 Summary of semantic infringements by ETQE***

The preceding sections provide the results of records that infringe on this semantic business rule from the granular perspective of the qualification enrolment record in relation to the complete dataset. This approach supports the determination of patterns within the data that point to systemic and anomalous problems within the overall dataset, which in turn lends itself to assessing the quality of the data in the data set.

The approach however ignores the diverse nature of ETQEs, and in particular the volume of the records that each ETQE submits to the NLRD. The final step in the analysis of this semantic business rule provides an overview of the percentage of records, calculated as a percentage of the number of records submitted by the ETQE, which infringe on this semantic business rule.

The results are presented as the percentage of records submitted by the ETQE that fall into a category that describes a semantic business rule issue (see Table 4.10.4.1):

Table 4.10.4.1 % of records submitted by an ETQE that in the case where the learner has achieved the qualification, and the qualification is a unit standards based qualification, the learner has achieved the correct number and mix of credits for the qualification

ETQE Identifier	% Semantic Rule Issue
1108	67.59%
1119	61.17%
1117	56.33%
1111	49.00%
1103	48.04%
1122	44.43%
1123	38.48%
1107	38.19%
1112	33.91%
1105	32.62%
1120	28.61%
1104	28.16%
1121	28.00%
1125	26.58%
1109	21.52%
1115	17.63%
1127	14.97%
1110	14.76%
1075	14.69%
1106	14.36%
1113	12.06%
1118	11.17%
1126	9.27%
1116	8.89%
1102	7.85%
1114	2.00%
1124	0.26%

The results clearly illustrate that the infringement of this semantic business rule could be considered systemic at a number of the ETQEs.

#### **4.10.5 Conclusion**

The analysis of qualification enrolment records in regard to whether, in the case where the learner has achieved the qualification, and the qualification is a unit standards based qualification the learner has achieved the correct number and mix of credits for the qualification, highlights the possibility of systemic issues in regard to unit standard based qualification achievements.

The cluster analysis for the 'Insufficient Unit Standard Credits Achieved', 'No Unit Standard Credits Achieved' and 'Incorrect Mix of Unit Standard Credits Achieved' groups is able to provide a clear description of the data in the categories. Further, a comparison across the three cluster analyses shows that ETQE identifiers 1105 and 1106 are featured in all three groups and ETQE identifiers 1111, 1112 and 1126 are featured in two categories.

Finally, the summary of semantic infringements by ETQE, which shows the percentage of infringements of this semantic business rule calculated as a percentage of the number of qualification enrolment records submitted to the NLRD by ETQE, shows clear trends of a systemic nature at some ETQEs.

Specific recommendations in regard to data records that are not compliant to this semantic business rule are provided in Appendix P.9.

#### **4.11 Data quality affinity**

The previous sections present analyses of specific semantic business rules as they apply to data records found in specific tables. Although these analyses present a detailed view of the nature and extent of data that does not conform to a specific semantic business rule, these analyses do not describe any co-existent relationships of non-conformance to more than one semantic business rule in the same data record.

This section presents the results of the analysis of learner enrolment records that contravene one or more of the semantic business rules in order to determine whether there are any associations and connections between the contraventions of semantic business rules.

This section presents the results of this type of analysis for learnership enrolment records, qualification enrolment records and unit standard enrolment records.

##### ***4.11.1 Learnership enrolments***

In order to determine any data quality affinity in learnership enrolment records, any learnership enrolment record that contains one of the following values, per semantic business rule, is included in the data set for analysis:

1. Contravention in regard to whether the ETQE was accredited for the duration of the learner's active enrolment on the learnership (ETQE\_IND, Appendix C.2)

Start Before, End During

Start During, End After

2. Contravention in regard to whether the provider was accredited for the duration of the learner's active enrolment on the learnership (PROV\_IND, Appendix C.2)

Start Before, End Before

No Accreditation

Start Before, End During

Start After, End After

Start During, End After

Start Before, End After

3. Contravention in regard to whether the assessor was registered at the time of the completion of the learnership (ASOR\_IND, Appendix C.2)

Lshp Completed After Assessor Registration

Lshp Completed Before Assessor Registration

No Registration

4. Contravention in relation to whether the intrinsic relationship between the completion of a learnership and achievement of its related qualification has been upheld (QENROL\_IND, Appendix C.2)

No Qual Enrolment

Lshp Enrolled, Qual Achieved (Derived)

Lshp Enrolled, Qual Achieved

Lshp Completed, Qual Enrolled

Lshp Completed, Qual Enrolled (Derived)

Lshp Completed Before Qual (Derived)

Lshp Completed Before Qual

Lshp Completed After Qual (Derived)

Lshp Completed After Qual

The resultant data set contains 30.59% of the learnership enrolment records.

The association rule data mining technique, as described in Appendix I.4, is applied to this data set in an effort to determine whether there are associations and connections between the contraventions of semantic business rules in learnership enrolment records.

The data mining effort results in the development of one association rule that meets the minimum criteria as defined for the evaluation of the association rules. The rule states, with a minimum confidence level of 96.64% and a minimum support level of 1.72%, that:

IF

ETQE\_IND = Start Before, End During

THEN

QENROL\_IND = No Qual Enrolment

The analysis of the “Start Before, End During” category for ETQE\_IND (Section 4.2.1) notes that 95.93% of the records belong to a single ETQE and learnership. Further investigation finds that these denote specific situations in which the ETQE, which resulted after an amalgamation, has found that a previous ETQE had not submitted data in regard to a specific learnership and related learnership enrolment records to the NLRD. The current ETQE has, on request of the Director of the NLRD, submitted the missing learnership and learnership enrolment records to the NLRD. In this specific case there is no data in the NLRD that defines a relationship between the learnership, the previous ETQE and the current ETQE. In consultation with the Director of the NLRD it was decided that these types of records need to be assumed as correct for the purposes of this research.

The association rule is significant in that it suggests that the ETQE that has submitted these missing learnership enrolment records has failed to submit the qualification enrolment records that are related to the learnership enrolment.

#### ***4.11.2 Qualification enrolments***

In order to determine any data quality affinity in qualification enrolment records, any qualification enrolment record that contained one of the following values, per semantic business rule, is included in the data set for analysis:

1. Contravention in regard to whether the ETQE was accredited for the duration of the learner’s active enrolment on the qualification (ETQE\_IND, Appendix E.2)

Start Before, End During

Start Before, End Before

Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End Before

2. Contravention in regard to whether the ETQE was accredited to quality assure the qualification for the duration of the learner's active enrolment on the qualification (ETQE\_ACCRED\_IND, Appendix E.2)
  - No Accreditation
  - Start After, End After
  - Start Before, End Before
  - Start Before, End During
  - Start During, End After
  - Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End Before
  - Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End During
  - Submitting ETQE: Start During, End After, Other ETQE: Start After, End After
  - Submitting ETQE: Start During, End After, Other ETQE: Start During, End After
3. Contravention in regard to whether the provider was accredited for the duration of the learner's active enrolment on the qualification (PROV\_IND, Appendix E.2)
  - Start Before, End Before
  - No Accreditation
  - Start Before, End During
  - Start After, End After
  - Start During, End After
  - Start Before, End After
4. Contravention in regard to whether the provider was accredited to offer the qualification for the duration of the learner's active enrolment on the qualification (PROV\_ACCRED\_IND, Appendix E.2)
  - Start Before, End Before
  - No Accreditation
  - Start Before, End During
  - Start After, End After
  - Start During, End After
  - Start Before, End After
5. Contravention in regard to whether the assessor was registered at the time of the completion of the qualification (ASOR\_IND, Appendix E.2)
  - Qual Achieved After Assessor Registration
  - Qual Achieved Before Assessor Registration
  - No Registration

6. Contravention in regard to whether the assessor was registered to assess the qualification at the time of the completion of the qualification (ASOR\_IND, Appendix E.2)

Qual Achieved After Assessor Registration

Qual Achieved Before Assessor Registration

No Registration

7. Contravention in regard to whether the qualification was registered for the duration of the learner's active enrolment on the qualification (QUAL\_REGSTR\_IND, Appendix E.2)

Start After, End After

Start After, End During

Start Before, End Before

Start Before, End During

Start During, End After

8. Contravention in regard to whether the learner achieved (USTD\_MIX\_IND, Appendix E.2):

- the minimum required number of credits for the qualification, on or before the achievement of the qualification, based on the achievement of unit standards related to the qualification, and
- the correct range of credits for the qualification, achieved on or before the achievement of the qualification, based on the achievement of unit standards that are been defined as core, fundamental and elective unit standards for the qualification

Insufficient Credits Achieved, Core Credits OK, Fundamental Credits OK, Insufficient Elective Credits

Insufficient Credits Achieved, Core Credits OK, Insufficient Fundamental Credits, Elective Credits OK

Insufficient Credits Achieved, Core Credits OK, Insufficient Fundamental Credits, Insufficient Elective Credits

Insufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK

Insufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Insufficient Elective Credits

Insufficient Credits Achieved, Insufficient Core Credits, Insufficient Fundamental Credits, Elective Credits OK

Insufficient Credits Achieved, Insufficient Core Credits, Insufficient Fundamental Credits, Insufficient Elective Credits

No Unit Standards Achieved

Sufficient Credits Achieved, Core Credits OK, Fundamental Credits OK, Insufficient Elective Credits

Sufficient Credits Achieved, Core Credits OK, Insufficient Fundamental Credits, Elective Credits OK

Sufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK

Sufficient Credits Achieved, Insufficient Core Credits, Insufficient Fundamental Credits, Elective Credits OK

The resultant data set contains 33.26% of the qualification enrolment records.

The association rule data mining technique, as described in Appendix I.4, is applied to this data set in an effort to determine whether there are associations and connections between the contraventions of semantic business rules in qualification enrolment records.

The data mining effort results in the development of six association rules that meet the minimum criteria as defined for the evaluation of the association rules:

1. Relationship between ETQE accreditation to quality assure a qualification and the registration of the qualification.

This rule states, with a minimum confidence level of 100% and a minimum support level of 0.48%, that:

IF

ETQE\_ACCRED\_IND = Start Before, End During

THEN

QUAL\_REGSTR\_IND = Start Before, End During

This rule succinctly points to the intrinsic relationship between qualifications and the ETQEs that have been accredited to quality assure them. Many ETQE accreditations to

quality assure qualifications have start and end dates that correspond directly with the start and end dates of the qualification.

2. Relationship between the registration status of the assessor and registration status of the same assessor to assess the qualification.

This rule states, with a minimum confidence of 95.93% and a minimum support level of 0.40% that:

IF

ASOR\_IND = Qual Achieved Before Assessor Registration

THEN

ASOR\_REGSTR\_IND = Qual Achieved Before Assessor Registration

This rule succinctly points to the intrinsic relationship between assessor registrations and assessor registrations to assess qualifications. Many assessor registrations have start dates that coincide with the start date for the same assessor to quality assure a specific qualification.

3. Relationship between the accreditation of an ETQE and the achievement of unit standards towards a unit standards based qualification.

This rule states, with a minimum confidence of 95.10% and a minimum support level of 1.18% that:

IF

ETQE\_IND = Start Before, End Before

THEN

UNIT\_STD\_MIX\_IND = No Unit Standards Achieved

The analysis of the “Start Before, End before” category for ETQE\_IND (Section 4.2.2) notes that 96.42% of the records belong to a single ETQE and 3 qualifications. Further investigation finds that these denote specific situations in which the ETQE, which resulted after an amalgamation, has found that a previous ETQE had not submitted data in regard to a specific qualification and related qualification enrolment records to the NLRD. The current ETQE has, on request of the Director of the NLRD, submitted the missing qualification and qualification enrolment records to the NLRD. In this specific case there is no data in the NLRD that defines a relationship between the qualification, the previous ETQE and the current ETQE. In consultation with the Director of the

NLRD it was decided that these types of records need to be assumed as correct for the purposes of this research.

The association rule is significant in that it suggests that the ETQE that has submitted these missing qualification enrolment records has failed to submit the unit standard enrolment records that are related to the qualification enrolment.

4. Relationship between the registration of the qualification and ETQE accreditation to quality assure a qualification.

This rule states, with a minimum confidence of 91.85% and a minimum support level of 1.75% that:

IF

QUAL\_REGSTR\_IND = Start After, End After

THEN

ETQE\_ACCRED\_IND = Start After, End After

As with the first association rule, this rule succinctly points to the intrinsic relationship between qualifications and the ETQEs that have been accredited to quality assure them. Many ETQE accreditations to quality assure qualifications have start and end dates that correspond directly with the start and end dates of the qualification.

5. Relationship between the registration of the qualification and ETQE accreditation to quality assure a qualification.

This rule states, with a minimum confidence of 90.05% and a minimum support level of 0.48% that:

IF

QUAL\_REGSTR\_IND = Start Before, End During

THEN

ETQE\_ACCRED\_IND = Start Before, End During

As with the first and fourth association rules, this rule succinctly points to the intrinsic relationship between qualifications and the ETQEs that have been accredited to quality assure them. Many ETQE accreditations to quality assure qualifications have start and end dates that correspond directly with the start and end dates of the qualification.

6. Relationship between the accreditation of the provider, the achievement of unit standards towards a unit standards based qualification and the accreditation of the provider to offer the qualification.

This rule states, with a minimum confidence of 87.07% and a minimum support level of 0.31% that:

IF

PROV\_IND = Start Before, End During

AND

UNIT\_STD\_MIX\_IND = No Unit Standards Achieved

THEN

PROV\_ACCRED\_IND = Start Before, End During

This rule shows that when a qualification enrolment's active enrolment period starts prior to the provider's accreditation time period then the qualification enrolment's active enrolment period also precedes the provider's accreditation to offer the qualification time period. This rule may be highlighting that when providers offer qualifications prior to their accreditation and accreditation to offer the qualification the provider may be offering unit standards based qualifications incorrectly.

#### ***4.11.3 Unit Standard enrolments***

In order to determine any data quality affinity in unit standard enrolment records, any unit standard enrolment record that contained one of the following values, per semantic business rule, is included in the data set for analysis:

1. Contravention in regard to whether the ETQE was accredited for the duration of the learner's active enrolment on the unit standard (ETQE\_IND, Appendix G.2)

Start Before, End Before

Start Before, End During

Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End During

Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End During

Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End Before

Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End Before

2. Contravention in regard to whether the ETQE was accredited to quality assure the unit standard for the duration of the learner's active enrolment on the unit standard (ETQE\_ACCRED\_IND, Appendix G.2)

No Accreditation

Start After, End After

Start Before, End Before

Start Before, End During

Start During, End After

Submitting ETQE: No Accreditation, Other ETQE: Start After, End After

Submitting ETQE: No Accreditation, Other ETQE: Start Before, End Before

Submitting ETQE: No Accreditation, Other ETQE: Start Before, End During

Submitting ETQE: Start After, End After, Other ETQE: Start After, End After

Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End Before

Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End After

Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End Before

Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End During

Submitting ETQE: Start During, End After, Other ETQE: Start After, End After

Submitting ETQE: Start During, End After, Other ETQE: Start During, End After

3. Contravention in regard to whether the provider was accredited for the duration of the learner's active enrolment on the unit standard (PROV\_IND, Appendix G.2)

No Accreditation

Start After, End After

Start Before, End After

Start Before, End Before

Start Before, End During

Start During, End After

4. Contravention in regard to whether the provider was accredited to offer the unit standard for the duration of the learner's active enrolment on the unit standard (PROV\_ACCRED\_IND, Appendix G.2)

No Accreditation

Start After, End After

Start Before, End After

Start Before, End Before

Start Before, End During

Start During, End After

5. Contravention in regard to whether the assessor was registered at the time of the completion of the unit standard (ASOR\_IND, Appendix G.2)

No Registration

UStd Achieved After Assessor Registration

UStd Achieved Before Assessor Registration

6. Contravention in regard to whether the assessor was registered to assess the unit standard at the time of the completion of the unit standard (ASOR\_IND, Appendix G.2)

No Registration

UStd Achieved After Assessor Registration

UStd Achieved Before Assessor Registration

7. Contravention in regard to whether the unit standard was registered for the duration of the learner's active enrolment on the unit standard (USTD\_REGSTR\_IND, Appendix G.2)

Start After, End After

Start After, End During

Start Before, End After

Start Before, End Before

Start Before, End During

Start During, End After

The resultant data set contains 26.12% of the unit standard enrolment records.

The association rule data mining technique, as described in Appendix I.4, is applied to this data set in an effort to determine whether there are associations and connections between the contraventions of semantic business rules in unit standard enrolment records.

The data mining effort results in the development of thirteen association rules that meet the minimum criteria as defined for the evaluation of the association rules:

1. Relationship between the ETQE accreditation to quality assure a unit standard and the registration of the unit standard.

This rule states, with a minimum confidence level of 100% and a minimum support level of 4.10%, that:

IF

ETQE\_ACCRED\_STATUS-Start Before, End Before

THEN

USTD\_REGSTR\_STATUS-Start Before, End Before

This rule succinctly points to the intrinsic relationship between unit standards and the ETQEs that have been accredited to quality assure them. Many ETQE accreditations to quality assure unit standards have start and end dates that correspond directly with the start and end dates of the unit standard.

2. Relationship between the ETQE accreditation to quality assure a unit standard and the registration of the unit standard.

This rule states, with a minimum confidence level of 100% and a minimum support level of 0.55%, that:

IF

ETQE\_ACCRED\_STATUS-Submitting ETQE: Start Before, End During, Other  
ETQE: Start Before, End During

THEN

USTD\_REGSTR\_STATUS-Start Before, End During

As with the first association rule this rule points to the intrinsic relationship between unit standards and the ETQEs that have been accredited to quality assure them. Many ETQE accreditations to quality assure unit standards have start and end dates that correspond directly with the start and end dates of the unit standard.

3. Relationship between the ETQE accreditation to quality assure a unit standard and the registration of the unit standard.

This rule states, with a minimum confidence level of 100% and a minimum support level of 0.44%, that:

IF

ETQE\_ACCRED\_STATUS-Submitting ETQE: Start Before, End Before, Other  
ETQE: Start Before, End Before

THEN

USTD\_REGSTR\_STATUS-Start Before, End Before

As with the first two association rules this rule points to the intrinsic relationship between unit standards and the ETQEs that have been accredited to quality assure them.

Many ETQE accreditations to quality assure unit standards have start and end dates that correspond directly with the start and end dates of the unit standard.

4. Relationship between the ETQE accreditation, the ETQE accreditation to quality assure a unit standard and the registration of the unit standard.

This rule states, with a minimum confidence level of 100% and a minimum support level of 0.41%, that:

IF

ETQE\_ACCRED\_STATUS-Submitting ETQE: Start Before, End Before, Other  
ETQE: Start Before, End Before AND ETQE\_STATUS-Start Before, End Before

THEN

USTD\_REGSTR\_STATUS-Start Before, End Before

Unlike the preceding three association rules this rule includes the accreditation of the ETQE in conjunction with the accreditation of the ETQE to quality assure the unit standard and the registration of the unit standard. However similarly the rule points to the intrinsic relationship between unit standards and the ETQEs that have been accredited to quality assure them and the ETQE's accreditation. ETQE accreditations to quality assure unit standards have start and end dates that correspond directly with the start and end dates of the unit standard. ETQE accreditations to quality assure unit standards have start and end dates that correspond directly with the start and end dates of the ETQEs accreditation.

5. Relationship between the ETQE accreditation, the ETQE accreditation to quality assure a unit standard and the registration of the unit standard.

This rule states, with a minimum confidence of 100% and a minimum support level of 0.32% that:

IF

ETQE\_ACCRED\_STATUS-Submitting ETQE: Start Before, End During, Other  
ETQE: Start Before, End During AND ETQE\_STATUS-Start Before, End During

THEN

USTD\_REGSTR\_STATUS-Start Before, End During

As with the fourth association rule this rule points to the intrinsic relationship between unit standards and the ETQEs that have been accredited to quality assure them and the ETQE's accreditation. ETQE accreditations to quality assure unit standards have start and end dates that correspond directly with the start and end dates of the unit standard. ETQE accreditations to quality assure unit standards have start and end dates that correspond directly with the start and end dates of the ETQEs accreditation.

6. Relationship between the registration status of the assessor and registration status of the same assessor to assess the unit standard.

This rule states, with a minimum confidence of 98.68% and a minimum support level of 0.81% that:

IF

ASOR\_STATUS-UStd Achieved Before Assessor Registration

THEN

ASOR\_REGSTR\_STATUS-UStd Achieved Before Assessor Registration

This rule succinctly points to the intrinsic relationship between assessor registrations and assessor registrations to assess unit standards. Many assessor registrations have start dates that coincide with the start date for the same assessor to quality assure a specific unit standard.

7. Relationship between the accreditation of the provider to offer the unit standard, the registration of the unit standard and the accreditation of the provider.

This rule states, with a minimum confidence of 98.66% and a minimum support level of 0.87% that:

IF

PROV\_ACCRED\_STATUS-Start Before, End During AND  
USTD\_REGSTR\_STATUS-Start During, End After

THEN

PROV\_STATUS-Start Before, End During

Although this rule points to the intrinsic relationship between provider accreditations and the provider's accreditation to offer a unit standard, the expression of this rule is awkward when conjoined with the registration of the unit standard. The rule shows a

misalignment between the registration of the unit standard and the accreditation of a provider to offer the unit standard. The rule seemingly suggests that when the provider accreditation to offer the unit standard was awarded, the provider had already enrolled learners on the unit standard and the ETQE had expected the registration of the unit standard to be extended.

8. Relationship between the ETQE accreditation to quality assure a unit standard and the registration of the unit standard.

This rule states, with a minimum confidence level of 97.42% and a minimum support level of 1.72%, that:

IF

ETQE\_ACCRED\_STATUS-Start Before, End During

THEN

USTD\_REGSTR\_STATUS-Start Before, End During

As with the first two association rules this rule points to the intrinsic relationship between unit standards and the ETQEs that have been accredited to quality assure them. Many ETQE accreditations to quality assure unit standards have start and end dates that correspond directly with the start and end dates of the unit standard.

9. Relationship between the accreditation of the provider to offer the unit standard, the registration of the unit standard and the accreditation of the provider.

This rule states, with a minimum confidence of 95.97% and a minimum support level of 0.31% that:

IF

USTD\_REGSTR\_STATUS-Start Before, End Before AND

PROV\_ACCRED\_STATUS-Start Before, End Before

THEN

PROV\_STATUS-Start Before, End Before

As with the seventh association rule, although this rule points to the intrinsic relationship between provider accreditations and the provider's accreditation to offer a unit standard, the expression of this rule is awkward when conjoined with the registration of the unit standard. Given that both the provider accreditation and unit

standard registration succeeded the unit standard enrolments it seems that the ETQE was not aware that the unit standard's registrations had expired.

10. Relationship between ETQE accreditation to quality assure a unit standard and the accreditation of the ETQE.

This rule states, with a minimum confidence level of 92.41% and a minimum support level of 0.41%, that:

IF

ETQE\_ACCRED\_STATUS-Submitting ETQE: Start Before, End Before, Other  
ETQE: Start Before, End Before

THEN

ETQE\_STATUS-Start Before, End Before

This rule succinctly points to the intrinsic relationship between the accreditation of an ETQE and the accreditation of an ETQE to quality assure a unit standard. The rule suggests that either, ETQEs are quality assuring unit standard enrolments that took place before the ETQE was accredited and before the ETQE was accredited to quality assure the unit standard or the dates on these enrolment records are incorrect.

11. Relationship between ETQE accreditation to quality assure a unit standard, the registration of the unit standard and the accreditation of the ETQE.

This rule states, with a minimum confidence level of 92.41% and a minimum support level of 0.41%, that:

IF

ETQE\_ACCRED\_STATUS-Submitting ETQE: Start Before, End Before, Other  
ETQE: Start Before, End Before AND USTD\_REGSTR\_STATUS-Start Before, End  
Before

THEN

ETQE\_STATUS-Start Before, End Before

As with the tenth association rule, this rule points to the intrinsic relationship between the accreditation of an ETQE and the accreditation of an ETQE to quality assure a unit standard. However this rule also includes the registration of the unit standards. Given that the accreditation of the ETQE, the accreditation of the ETQE to quality assure the

unit standard and the registration of the unit standard succeeded these enrolments the likelihood that the dates on the enrolment records is incorrect seems extremely high.

12. Relationship between the accreditation of the provider to offer the unit standard, the registration of the unit standard and the accreditation of the provider.

This rule states, with a minimum confidence of 88.84% and a minimum support level of 0.31% that:

IF

USTD\_REGSTR\_STATUS-Start Before, End Before AND PROV\_STATUS-Start Before, End Before

THEN

PROV\_ACCRED\_STATUS-Start Before, End Before

This association rule is very similar to the ninth association rule and similarly points to the intrinsic relationship between provider accreditations and the provider's accreditation to offer a unit standard. However the expression of this rule is awkward when conjoined with the registration of the unit standard. Given that both the provider accreditation and unit standard registration succeeded the unit standard enrolments it seems that the ETQE was not aware that the unit standard's registrations had expired.

13. Relationship between the accreditation of the provider to offer the unit standard, the registration of the unit standard and the accreditation of the provider.

This rule states, with a minimum confidence of 86.90% and a minimum support level of 0.87% that:

IF

USTD\_REGSTR\_STATUS-Start During, End After AND PROV\_STATUS-Start Before, End During

THEN

PROV\_ACCRED\_STATUS-Start Before, End During

This association rule is very similar to the seventh association rule and similarly points to the intrinsic relationship between provider accreditations and the provider's accreditation to offer a unit standard. However the expression of this rule is awkward when conjoined with the registration of the unit standard. The rule shows a

misalignment between the registration of the unit standard and the accreditation of a provider to offer the unit standard. The rule seemingly suggests that when the provider accreditation to offer the unit standard was awarded, the provider had already enrolled learners on the unit standard and the ETQE had expected the registration of the unit standard to be extended.

#### ***4.11.4 Conclusion***

This section presents the results of the analysis of learner enrolment records that contravene one or more of the semantic business rules in order to determine whether there are any associations and connections between the contraventions of semantic business rules.

The results of the association data mining technique for learnership enrolment records yields only one rule. The single rule however highlights a data issue that had not been made apparent in the exploratory or clustering data mining techniques.

The qualification enrolment association data mining results yields six different rules. Of these six rules, four rules describe intrinsic relationships between ETQEs and their accreditations, providers and their accreditations and qualification registrations. One of the rules brings further insight into an issue that had already been noted during the exploratory and clustering data mining of these records, whereas another rule highlights an issue that was not apparent during the exploratory and clustering data mining of these records.

The results of the association data mining technique for unit standard enrolment records yields thirteen rules. All of these rules bring some insight into the unit standard enrolment records that did not comply to the semantic business rules for these records, however most of these insights relate to the intrinsic relationship between ETQEs and their accreditations, providers and their accreditations and unit standard registrations.

Specific recommendations in regard to what the association rules highlighted are provided in Appendix P.10.

#### **4.12 Chapter summary**

This chapter presents the results of the analysis of the learnership, qualification and unit standard enrolments in relation to the nineteen (19) applicable semantic business rules as

defined in Appendix A.7. The results of the analysis are presented by semantic business rule and enrolment type.

The analysis shows that overall, as a percentage of records submitted by the ETQE, for learnership, qualification and unit standard enrolments ETQE identifiers 1116, 1123 and 1115 have the highest percentage of records that infringe the semantic business rules. The percentage of records submitted by the ETQE per type of enrolment record however shows that the following ETQE identifiers have the highest percentage of records that infringe the semantic business rules for the enrolment type:

- Learnership enrolment records – ETQE Identifiers 1123, 1104 and 1115
- Qualification enrolment records – ETQE Identifiers 1108, 1119 and 1105
- Unit Standard enrolment records – ETQE Identifiers 1116, 1100 and 1122

The analysis of the data in regard to adherence of ETQE accreditation and ETQE accreditation to quality assure a qualification/unit standard do not reveal any possible systemic issues. However the analysis does highlight problems associated with the amalgamation of ETQEs where SAQA was required to request the backloading of data that the transient ETQE had not submitted to the NLRD (see Sections 4.2.4 and 4.3.3).

The results of the analysis of provider accreditations suggests that there are systemic issues in regard to the accreditation of providers. On review of the overall compliance of provider accreditations it is found that 12.99% learnership, 10.65% qualification and 12.64% unit standard enrolment records infringe on this semantic business rule. The scope and volume of records require that the analysis of this data be extended to a more in-depth review of the data which include cluster data mining (see Appendix I.1).

The analysis of provider accreditations suggests that no significant mechanism exists that facilitates the exchange of information in regard to providers between ETQEs. This conclusion is drawn as a result of the high percentage of records that do not comply to the provider accreditation related semantic business rules when the analysis focuses on providers that are not accredited by the submitting ETQE (see Section 4.4.4). As described in Section 3.8.3.5, although a provider may be accredited to offer qualifications by more than one ETQE, a provider may only have one primary ETQE. In practice it seems that although the ETQE is able to identify the primary ETQE of the provider:

- the ETQE is either not able to determine the accreditation start date, end date and status of the provider, or
- the information system of the ETQE does not have access to the accreditation start date, end date and status of the provider.

As with the analysis of provider accreditation, the percentage of records that fail to comply with the semantic business rules in regard to provider accreditations to offer qualifications or unit standards suggests that there are systemic issues in this regard (see Section 4.5.3). Due to the scope and volume of the records that infringe on these rules this data mining analysis is extended to include cluster data mining techniques (see I.1).

The assessor registrations and assessor registrations to assess qualifications and unit standards semantic business rule analysis suggests that there are no systemic issues in regard to these semantic business rules (see Section 4.6.4). The reader should note however that the provision of assessor related data for an achieved record is not a requirement. The number of achieved records that have been submitted without assessor-related data is significantly high and the reasons for assessor-related data not being provided is unclear.

The analysis of learnership enrolment records, in regard to whether the intrinsic relationship between the completion of a learnership and achievement of its related qualification has been upheld, highlights the possibility of a number of systemic issues. Accordingly the analysis of this data is also extended to include cluster data mining techniques (see Appendix P.7). The evolving understanding of the relationship between learnership enrolment records and qualification enrolment records in combination with the late introduction of the submission of learnership enrolment records to the NLRD may have contributed to this result (see Section 4.8.11).

Although the analysis of the qualification registration semantic business rule did highlight some issues, the overall results show that there are no systemic issues in regard to adherence to these rules (see Section 4.9.3). The analysis of the unit standard registration semantic business rule shows that the results are overall statistically significant. However, the volume and diversity of these results suggest that the issues are not systemic in nature (see Section 4.9.3).

The analysis of qualification enrolment records in regard to whether, in the case where the learner has achieved the qualification, and the qualification is unit standards based, the learner has achieved the correct number and mix of credits for the qualification, highlights the possibility of systemic issues in this regard (see Section 4.10.5). Accordingly the analysis of this data is also extended to include cluster data mining techniques.

Finally, a further analysis of learner enrolment records that contravened one or more of the semantic business rules is conducted in order to determine whether there are any associations and connections between the contraventions of semantic business rules (see Section 4.11). Generally the association rules that were generated describe intrinsic relationships between ETQEs and their accreditations, providers and their accreditations and qualification/unit standard registrations. Although these rules seem prosaic, they offer insight into how testing for compliance to the semantic business rules could be streamlined. Two of the association rules however provide additional insight into the data that was not been made evident utilizing exploratory or cluster data mining techniques (see Section 4.11.4).

In most instances the analysis of the compliance of data in accordance to a specific semantic business rule results in a recommendation, for SAQA, in regard to further steps that can be taken.

## 5 Chapter 5 – Conclusions and recommendations

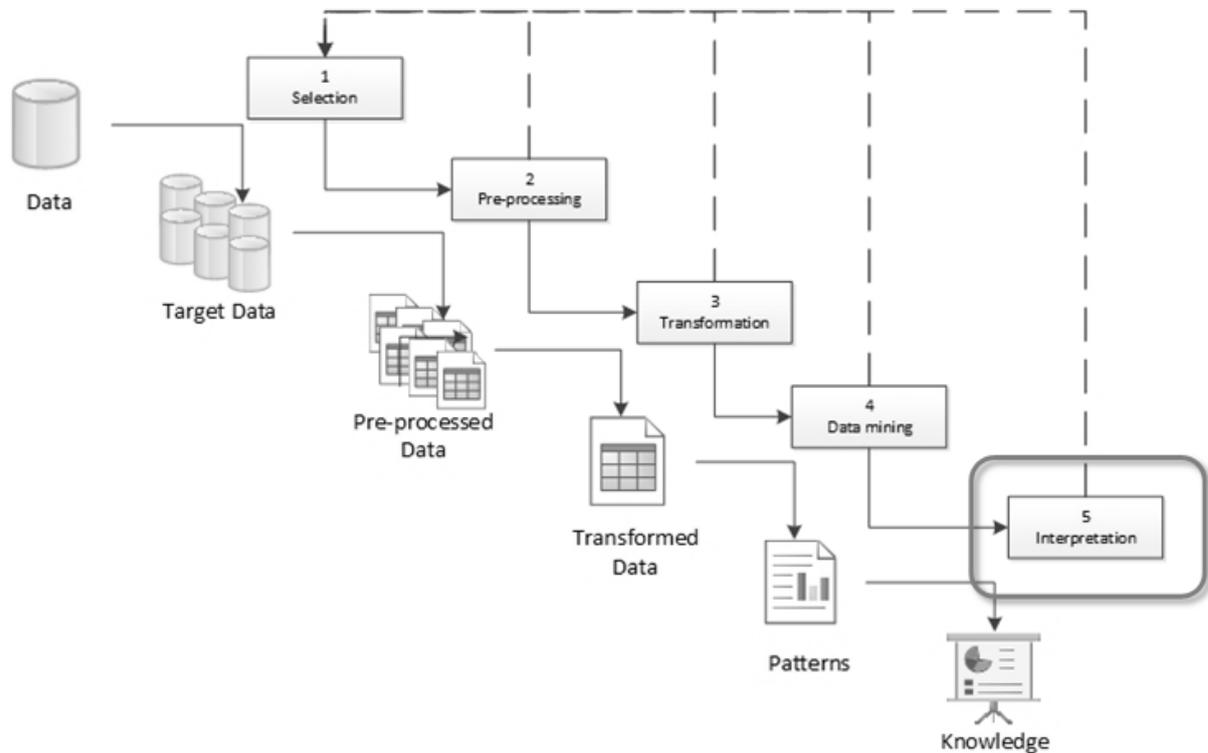


Figure 5.1 KDD phases – Interpretation

### 5.1 Response to research objectives and questions

This research helps to answer the research question:

How can data quality mining be used to describe, monitor and evaluate the scope and impact of semantic data quality problems in the learner enrolment data on the National Learners' Records Database?

There is a substantial amount of research related to DQM which focuses on one of four classes of DQM, namely; duplicate detection/record matching, instance conflict resolution, missing/incomplete data, and data staleness (Berti-Equille, 2007, p. 106). This research however differs in that it addresses how to data mine data quality in relation to the semantic business rules of the data. This research shows that data quality data mining can be used to describe, monitor and evaluate the scope and impact of semantic data quality problems in the learner enrolment data on the NLRD.

An objective of this research is to determine which data mining techniques are best suited for the identification, measurement and description of semantic data quality deficiencies. The literature review for this research identifies EDM, association, classification and clustering data mining techniques as best suited for the identification, measurement and description of semantic data quality deficiencies (see Section 2.4).

Having identified which data mining techniques are best suited for the identification, measurement and description of semantic data quality deficiencies, the next objective of the research is to develop a standard set of data mining techniques that can identify, measure and describe semantic data quality deficiencies related to learner enrolment records in the NLRD data warehouse.

Although the research conducted by McCarthy and Earp (McCarthy & Earp, 2009) on the 2002 Census of Agriculture clearly addressed the need to measure and explain data quality deficiencies, the research utilized classification data mining techniques to identify the characteristics of specific errors in the data. Classification data mining techniques are not used to mine the data in the NLRD data sets because the data mining technique is a supervised data mining technique (see Section 2.3) which, in conjunction with the fact that the data in the NLRD had never been interrogated from the vantage point of compliance to the semantic business rules described in Section 3.6.2, makes the implementation of this technique overly uncertain.

The research however extensively uses EDM, association and clustering data mining techniques for the identification, measurement and description of semantic data quality deficiencies. Overall these data mining techniques did highlight interesting patterns in the data sets that were not apparent to SAQA prior to the research being conducted and the research has resulted in actionable information that can and has been used by SAQA to improve compliance to the semantic business rules and the quality of the data being submitted to the NLRD.

The literature review highlights that EDM would produce simple and fast analyses and summaries of the data in order to reveal the characteristics of the data (Dasu & Johnson, 2003). The research found that EDM techniques did allow for the summarization of the data and identification of hidden relationships with relative ease and provided a concise overview

of the data that allows for the identification of issues that required further investigation. The NLRD has never been interrogated from the vantage point of compliance to semantic business rules (see Section 2.4) and the implementation of EDM techniques that describe compliance to semantic business rules provides previously unattainable insight into the data. As a result each of the data analyses (see Section 3.8.1) contain a substantial amount of EDM related results.

Clustering data mining finds hidden patterns in data (Khosravani, 2012, p. 10) and as a result assisted in the further investigation and description of the data. Clustering data mining is an unsupervised data mining technique and as a result reduced the uncertainty of mining the data in the NLRD, which had never been interrogated from the vantage point of compliance to the semantic business rules. The clustering data mining technique was employed when compliance to a specific semantic business rule exceeded 5%.

Furthermore, association rule data mining techniques are implemented on all data records that do not comply to one or more semantic business rule. Association rule mining is an important aspect of data mining (García, Romero, Ventura, & Calders, 2007, p. 13) and provides further insights into the data by showing relationships between variables in the dataset that neither the EDM or clustering data mining techniques had discovered. One of the shortcomings of association rule data mining noted during the literature review (Moreno, Segreña, & López, 2005, p. 317) was evident in that some of the rules were uninteresting, as a number of the rules generated pointed to the intrinsic relationship between ETQEs and their accreditations, providers and their accreditations and qualification or unit standard registrations.

The utilization of these data mining techniques did highlight an unexpected shortcoming in the selection of these data mining techniques:

The extent of learner enrolment records at one ETQE can differ remarkably from the extent of learner enrolment records at another ETQE. Similarly, the scope of the semantic business rule issues at one ETQE can differ remarkable from the scope of semantic business rule issues at another ETQE. As an example, ETQE A may have 1 000 000 learner enrolment records of which 2 000 records fail to comply with a specific semantic business rule whereas ETQE B may have 1 000 learner enrolment records of which 1 000 records fail to comply with a specific semantic business rule. SAQA places priority on

the fact that ETQE B fails completely to adhere to the specific semantic business rule than that 0.2% of the records from ETQE A that fail to adhere to the specific semantic business rule. Although this example highlights differences between ETQEs, the same applies to other core aspects of the data such as providers, assessors, qualifications and unit standards. In order to address this type of scenario the dataset must be balanced/weighted in order adjust the data to account for the volume bias.

The off-the-shelf product used in this research does not offer a built-in feature with which to implement this type of population weighting in either of the data mining techniques used for this research. As a result, in order to resolve this issue the population weighting would need to be addressed in the pre-processing phase of the KDD process (see Section 3.5). In the instance of the ETQE example above, it would mean that a number of data records would need to be duplicated in order to balance the dataset, resulting an in substantially larger dataset. Additional datasets would also need to be created to balance/weight the data by provider, assessor, qualification and unit standard. Each of these datasets in turn would need to be data mined separately. This would extend the scope of the analysis of the data from one analysis effort to five.

Further, in the analysis of the data SAQA did indicate that in some instances the degree to which a semantic business rule was not complied with had greater importance. For example an enrolment record with a provider that has never been accredited (see Appendix J.1.2) is of the greatest concern to SAQA. In retrospection the weighting of specific values in specific data fields may have further enhanced the analysis of the data.

The requirement for weighted cluster data mining algorithms for missing population data is relatively unresearched (Saarela, 2015, p. 337). Although Saarela and Kärkkäinen have proposed a modified k-spatial medians algorithm that can be utilized for weighted clustering (Saarela, 2015, p. 341) the proposed solution is not a feature in mainstream data mining software applications. The requirement for weighted association data mining algorithms is well researched, with some of the earliest research in this regard being done in 1998 by Cai, Fu, Cheng, et al. (Cai, 1998) and a more formalized weighted association rule model being proposed by Wang, Yang and Philip (Wang, 2000, p. 270). Unfortunately, these solutions have also not been implemented as features in mainstream data mining software applications.

A further objective of this research is to determine which aspects of existing data quality mining methods and models can be utilized in order to define semantic data quality deficiencies. The literature review identifies the academic KDD model developed by Fayyad et al. (Fayyad, Piatetsky-Shapiro, & Smyth, The KDD Process for Extracting Useful Knowledge from Volumes of Data, 1996, p. 30) as the most appropriate model for the research. In summary the model has 5 different phases, namely:

1. Selection
2. Pre-processing
3. Transformation
4. Data-mining
5. Interpretation and evaluation

The initial understanding of this model assumed that the model may need to be amended to accommodate derived data elements that describe compliance to the semantic business rules into the dataset (see Section 2.2).

The selection phase is defined as the phase that focuses on selecting a sub-set of the data that is concise enough to be processed within a reasonable time period whilst also large enough to contain a representation of the specific data quality dimension being studied. Within the context of this research the selection process is complicated by the following aspects:

- a. The data for the research is stored in a discrete database that is inaccessible to the researcher, as a result the data needed to be extracted from the system (see Section 3.8.1).
- b. The nature of the analysis to be conducted requires the development of records that concisely contain the duration of a records active status (see Section 3.8.2).
- c. The data for the research originates from a relational database, the transmission of all related lookup fields to a record would have been extremely complicated and as a result the data needed to be denormalized (see Section 3.8.2) to some extent.
- d. The data needed to be de-identified and the de-identification needed to be conducted in such a manner as to ensure that SAQA could re-identify the data if required (see Section 3.8.1).

The above-mentioned complexities required that the SAQA DBA needed to conduct the selection of the data, based on specifications developed by the researcher in consultation with the Director of the NLRD. Further selection related activities were completed for the purposes of the research once the research commenced (see Sections C.3.1, E.3.1 and G.3.1). Items a, c and d above are attributable to the fact that the data being mined is not a privately owned and easily accessible database. Item b above on the other hand is attributable to technical attributes of the data and how the data must be extracted in line with the subsequent utilization of the data in regard to semantic business rule compliance.

As a result, in practice the KDD model for this research is modified to contain a 2-step selection process, the first step is conducted remotely at the source of the data and the second step is conducted locally by the researcher. This modification however is only applicable if further data mining of the NLRD is conducted by an external researcher.

The pre-processing phase is defined as the phase that addresses the cleaning of data in regard to missing data values and the removal of statistical noise (i.e. unexplained variations in the data). Whereas the transformation phase focuses on determining which data fields need to be utilized in order to provide meaningful patterns in regard to the data quality dimension being studied.

As highlighted in Section 2.2, by definition, neither of these phases accommodates the deriving of data elements that is required for the mining of data to determine semantic business rule compliance.

As a result, a new phase was introduced in the KDD process which relates to the derivation of data fields that describe compliance to semantic business rules. A derived data element is defined as “... *derived from other data elements using a mathematical, logical, or other type of transformation, e.g. arithmetic formula, composition, aggregation.*” (UNECE, 2000). Similarly, the derivation phase can be described as the phase in which data elements are derived from other data elements using mathematical, logical and other types of transformation. The derivation phases for learnership, qualification and unit standard enrolments records are described in Sections C.3.2 to C.3.7, E.3.2 to E.3.12 and G.3.2 to G.3.11 respectively.

The pre-processing phase followed the derivation phase as can be seen in Sections C.3.8, E.3.13 and G.3.12 in which problem records are removed from the data sets, i.e. “the removal of statistical noise”. The transformation phase followed just prior to the data mining phase and is conducted in line with the type of data mining technique being utilized (see Sections I.2, I.3 and I.4). This phase is followed by the data mining and interpretation phases of the academic KDD model.

The resultant modified KDD phases utilized for this research is as follows:

1. Selection
2. Derivation
3. Pre-processing
4. Transformation
5. Data-mining
6. Interpretation and evaluation

The above modified KDD model addresses the objective of the research to develop an adapted data quality mining method and model that can be utilized by the NLRD data warehouse to continuously assess data quality deficiencies in learner enrolment records in the NLRD data warehouse. Further to this objective the first five phases of the research process is documented as follows:

- The steps completed for the selection phase are documented as SQL scripts that can be submitted to the NLRD as and when required. The scripts are been developed in such a manner as to accommodate further data mining which is done internally by SAQA or externally by an external researcher or service provider.
- The steps for the derivation and pre-processing phases are documented as SQL scripts that can be submitted to the data extracted in the selection phase as and when required.
- The steps for the transformation and data mining phases are documented within the data mining tool as projects and workflows. Each type of learner enrolment record (learnership, qualification and unit standard) has workflows that correspond to each semantic business rule. The projects and workflows are implemented in such a manner that they can be reused on data that is prepared in the derivation and pre-processing steps.

## **5.2 Limitations**

This study focuses specifically on applying data mining in order to describe, monitor and evaluate the scope and impact of semantic data quality problems in the learner enrolment data on the NLRD. It emphasises the application of specific data mining techniques and a KDD model to develop a standard set of data mining techniques that can identify, measure and describe semantic data quality deficiencies related to learner enrolment records in the NLRD data warehouse.

The identified data mining techniques are applied to the data utilizing a specific software application. The specific software application is selected considering the existing software applications utilized by the NLRD data warehouse. The data mining technique shortcoming discussed in Section 5.1 in regard to population and data value weighting may be specific to the software application utilized. Other data mining software applications may have the type of functionality available that allows the user to specify how the data should be weighted during the data mining phase. The results of the same data mining steps, utilizing the same data mining techniques with the ability to weight the data set would produce different results in the analysis of the data.

## **5.3 Recommendations for future research**

The data in the NLRD had not been data mined prior to this research. The focus and scope of the data mining activities for this research are specifically directed at compliance to semantic business rules of the data in the NLRD. Given this narrow focus, and the resultant wealth of actionable information that is produced by this research, consideration must be given to data mining the data in the NLRD in relation to its education-related data. The modified two step selection process of the KDD model (see Section 5.1) provides a structural basis from which additional studies can model the selection phase of the KDD model for research purposes.

The research presents a material modification to the academic KDD model to be applied for the data mining of semantic business rules. This amendment should be applicable to any data mining projects that focus on the data mining of semantic business rules. The next step is to test and improve this amendment and to develop a KDD model specific to semantic business rule data mining.

## 6 Acronyms

CHE	Council on Higher Education
DBA	Database Administrator
DHET	Department of Higher Education and Training
DoL	Department of Labour
DVU	Data Validation Utility
EDM	Exploratory Data Mining
ETL	Extract-Transform-Load
ETQA	Education and Training Quality Assurance Body
ETQE	Education and Training Quality Entity
FASSET Authority	Finance and Accounting Services Sector Education and Training Authority
KDD	Knowledge Discovery in Databases
MRDM	Multi-Relational Data Mining
NLRD	National Learners' Records Database
NLRD specifications Database	Specifications for Load Files for the National Learners' Records Database
	Version 2.0
NQF	National Qualifications Framework
NSB	National Standards Body
QC	Quality Council
QCTO	Quality Council for Trades and Occupations
QP	Quality Partner
SAQA	South African Qualifications Authority
SETA	Sector Education and Training Authority

## 7 Bibliography

- Agrawal, H., Chafle, G., Goyal, S., Mittal, S., & Mukherjea, S. (2008). An Enhanced Extract-Transform-Load System for Migrating Data in Telecom Billing. *ICDE*, 1277-1286.
- Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the 20th International Conference on Very Large Data Bases*, (pp. 487-499). Santiago.
- Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, (pp. 207-217). California.
- Alizamini, F., Pedram, M., Alishahi, M., & Badie, K. (2010). Data Quality Improvement using Fuzzy Association Rules. *International Conference on Electronics and Information Engineering*, 1, 468-472.
- Alpar, P., & Winkelsträter, S. (2014). Assessment of data quality in accounting data with association rules. *Expert Systems with Applications*(41), 2259-2268.
- Apiletti, D., Bruno, G., Ficarra, E., & Baralis, E. (2006). Data Cleaning and Semantic Improvement in Biological Databases. *Journal of Integrative Bioinformatics*.
- Berthold, M., Borgelt, C., Höppner, F., & Klawonn, F. (2010). *Guide to Intelligent Data Analysis*. London: Springer-Verlag.
- Berti-Equille, L. (2007). Measuring and Modelling Data Quality for Quality-Awareness in Data Mining. In F. Guillet, & H. Hamilton, *Quality Measures in Data Mining*. Berlin: Springer.
- Cai, C. F. (1998). Mining association rules with weighted items. *Database Engineering and Applications Symposium, 1998. Proceedings. IDEAS'98*, (pp. 68-77). Cardiff.
- Chaovalit, P., & Zhou, L. (2005). Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches. *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*.
- Creswell, J. (2011). *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*. Boston: Pearson Education.
- Crios, K., Pedryzc, W., Swiniarski, R., & Kurgan, L. (2007). *Data Mining: A Knowledge Discovery Approach*. New York: Springer Science and Business Media.
- Crnkovic, G. (2010). Model-Based Reasoning in Science and Technology. In *Studies in Computational Intelligence* (pp. 359-380). Berlin: Springer.

- Crotty, M. (1998). *The foundations of social research: Meaning and perspective in the research process*. Sage.
- Das, S., & Saha, B. (2009). Data quality mining using genetic algorithm. *International Journal of Computer Science and Security*, 3(2), 105-112.
- Dasu, T., & Johnson, T. (2003). *Exploratory Data Mining and Data Cleaning: An Overview*. New Jersey: John Wiley & Sons, Inc.
- De Bie, T. (2011). An Information Theoretic Framework for Data Mining. *Proc. of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD11)*, 564-572.
- De Bie, T., & Spyropoulou, E. (2013). *A Theoretical Framework for Exploratory Data Mining: Recent Insights and Challenges Ahead*. Berlin.
- Department of Labour. (2013). *Learnerships*. Pretoria: Department of Labour.
- Dzeroski, S. (2003). Multi-Relational Data Mining: An introduction. *SIGKDD Explorations*, 5, 1-16.
- Easterby-Smith, M., Thorpe, R., & Lowe, A. (2002). *Management Research: An Introduction*. London: SAGE Publications.
- Eckerson, W., & White, C. (2003). *Evaluating ETL and Data Integration Platforms*. Seattle: The Data Warehousing Institute.
- Farzi, S., & Dastjerdi, A. (2010, February 1). Data Quality Measurement using Data Mining. *International Journal of Computer Theory and Engineering*, pp. 115-118.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3), 37-54.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3), 37-54.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. *KDD-96 Proceedings*, 82-88.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM*, 39(11), 27-34.
- García, E., Romero, C., Ventura, S., & Calders, T. (2007). Drawbacks and solutions of applying association rule mining in learning management systems. *Proceedings of the International Workshop on Applying Data Mining in e-Learning (ADML 2007)*, 13-22.
- Grüning, F. (2007). Data Quality Mining: Employing Classifiers for Assuring consistent Datasets. *Information Technologies in Environmental Engineering*, 85-94.

- Gupta, K. (2011). *Introduction to Data Mining with Case Studies* (2nd ed.). New Delhi: Prentice Hall India.
- Han, J., & Kamber, M. (2001). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.
- Hipp, J., Guntzer, U., & Grimmer, U. (2001). Data Quality Mining: Making a Virtue of Necessity. 52-57. Santa Barbara, CA. Retrieved April 12, 2013, from [http://www.cs.cornell.edu/johannes/papers/dmkd2001-papers/p5\\_hipp.pdf](http://www.cs.cornell.edu/johannes/papers/dmkd2001-papers/p5_hipp.pdf)
- Hipp, J., Muller, M., Hohendorff, J., & Naumann, F. (2007). Rule-based measurement of data quality in nominal data. *Paper presented at the 12th International Conference on Information Quality, Cambridge*.
- Hoffer, J., Prescott, M., & McFadden, F. (2008). *Modern Database Management*. New Jersey: Pearson Prentice Hall.
- Houshmand, M., & Alishahi, M. (2011). Improve the classification and sales management of products using multi-relational data mining. *Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on*.
- Inmon, W. (2005). *Building the Data Warehouse* (4 ed.). Indianapolis: Wiley Publishing.
- Januzaj, E., & Januzaj, V. (2009). An Application of Data Mining to Identify Data. *Third International Conference on Advanced Engineering Computing and Applications in Sciences*, 17-22.
- Khosravani, H. (2012, July). Proposing an Improved Semantic and Syntactic Data Quality Mining Method using Clustering and Fuzzy. *International Journal of Applied Information Systems*, 3(3), pp. 8-22.
- Krönner, H. (2005). National Qualifications Frameworks in the SADC Region. *BREDA Seminar*. Addis.
- MacQueen, J. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297). California: University of California Press.
- Madnick, S., Wang, R., & Lee, Y. (2009). Overview and Framework for Data and Information Quality Research. *Journal of Data and Information Quality*, 1(1), Article No. 2.
- McCarthy, J., & Earp, M. (2009). *Who Makes Mistakes? Using Data Mining Techniques to Analyze Reporting Errors in Total Acres Operated*. Washington: United States Department of Agriculture.

- Mehta, R., & Rajalakshmi, S. (2014). Semantic Integrity Constraint Rule Discovery and Outlier Detection in Relational Data as a Data Quality Mining Technique. *International Journal of Computer Applications*, 88(6), 23-26.
- Mehta, V., Sankarasubramaniam, B., & Rajalakshmi, S. (2012). An algorithm for fuzzy-based sentence-level document clustering for micro-level contradiction analysis. *International Conference on Advances in Computing, Communications and Informatics*, 102-105.
- Ministry in the Office of the President. (Act 101 of 1997). Higher Education.
- Ministry in the Office of the President. (Act 58 of 1995). South African Qualifications Authority Act.
- Ministry in the Office of the President. (Act 58 of 2001). General and Further Education and Training Quality Assurance.
- Ministry in the Office of the President. (Act 67 of 2008). National Qualifications Framework.
- Ministry in the Office of the President. (Act 97 of 1998). Skills Development Act.
- Ministry in the Office of the President. (Notice 1040 of 2012). Determination of the sub-frameworks that comprise the National Qualifications Framework.
- Moreno, M., Segre, S., & López, V. (2005). Association Rules: Problems, solutions and new applications. *Actas del III Taller Nacional de Minería de Datos y Aprendizaje*, 317-325.
- Myatt, J. (2006). *Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining*. New Jersey: Wiley.
- Natarajan, K., Li, J., & Koronios, A. (2009). Data Mining Techniques for Data Cleaning. *Proceedings of the 4th World Congress on Engineering Asset Management*, 796 - 804.
- Natarajan, K., Li, J., & Koronios, A. (2012). Use Rule Based to Predict Dirty Values. *Engineering Asset Management and Infrastructure Sustainability*, 693-703.
- Nisbet, R., Elder, J., & G., M. (2009). *Handbook of Statistical Analysis and Data Mining Applications*. Amsterdam: Elsevier.
- Orlikowski, W., & Baroudi, J. (1991). Studying Information Technology in Organizations: Research Approaches and Assumptions.
- Padhy, N., & Panigrahi, R. (2012). Multi Relational Data Mining Approaches: A Data Mining Technique. *International Journal of Computer Applications*, 5(17), 23-32.
- Praxis Computing. (2012, March 22). *Edu.Dex User Manual*. Retrieved April 15, 2014, from [http://nlrdinfo.octoplus.co.za/documents/UserManual.Edu.Dex.Level.2\\_v3.pdf](http://nlrdinfo.octoplus.co.za/documents/UserManual.Edu.Dex.Level.2_v3.pdf)

- Rahm, E., & Hai Do, H. (2000). Data Cleaning: Problems and Current Approaches. *IEEE Data Engineering Bulletin*.
- Saarela, M. K. (2015). Weighted Clustering of Sparse Educational Data. *25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, At Bruges, Belgium*. Bruges.
- Sheng, V., Provost, F., & Ipeirotis, P. (2008). Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 614-622). ACM.
- South African Qualifications Authoritya. (1998). Regulations under the South African Qualifications Authority Act (Act No.58 of 1995). (R 1127).
- South African Qualifications Authorityb. (1998). Regulations under the South African Qualifications Authority Act (Act No.58 of 1995). (R 452).
- South African Qualifications Authorityc. (2011, March 16). *NLRD Information for Data Suppliers*. Retrieved April 15, 2014, from [http://nlrdinfo.octopus.co.za/documents/NLRD\\_Data\\_Loads\\_Min\\_Standard\\_for\\_v2\\_of\\_NLRD\\_2011-03-16.doc](http://nlrdinfo.octopus.co.za/documents/NLRD_Data_Loads_Min_Standard_for_v2_of_NLRD_2011-03-16.doc)
- South African Qualifications Authorityd. (2013, July 11). *NLRD Information for Data Suppliers*. Retrieved April 15, 2014, from [http://nlrdinfo.octopus.co.za/documents/loadspecs\\_rel2%202013%2007%2011.doc](http://nlrdinfo.octopus.co.za/documents/loadspecs_rel2%202013%2007%2011.doc)
- South African Qualifications Authoritye. (2007, March 12). *Transitional Period for Implementation, where qualifications were replaced by new Qualifications and/or significant changes to Qualifications were effected*. Retrieved July 12, 2014, from South African Qualifications Authority: <http://www.sqa.org.za/show.php?id=5416#sthash.lGnt5LmY.dpuf>
- Sumathi, S., & Sivanandam, S. (2006). Data Mining & KDD. *Studies in Computational Intelligence*, 29, 231–241.
- Thiprungsri, S., & Vasarhelyi, M. (2011). Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach. *The International Journal of Digital Accounting Research*, 11, 69-84.
- UNECE. (2000). Terminology on Statistical Metadata. *Conference of European Statisticians Statistical Standards and Studies, No. 53*. Geneva.

- Vizhi, J., & Bhuvaneshwari, T. (2012, January). Data quality measurement on categorical data using genetic algorithm. *International Journal of Data Mining & Knowledge Management Process*, 2(1), 33-42.
- Wang, W. Y. (2000). Efficient Mining of Weighted Association Rules (WAR). *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 270-274). Boston.
- Witten, I., Frank, E., & Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques* (3rd ed.). Amsterdam: Elsevier.

## Appendix A

### A.1 Introduction

This section entails a review of the legislative, policy and regulatory framework that relates to the learner enrolment records that can be found on the NLRD. The review introduces six discrete concepts that form part of the South African National Qualification Framework (NQF) namely; Education and Training Quality Assurance Bodies, providers, assessor, learnerships, qualifications and unit standards (see Figure A.1). The section then goes on to describe how these six concepts relate to learner enrolment records and form the basis of ten (10) discrete semantic business rules that form the core of this research.

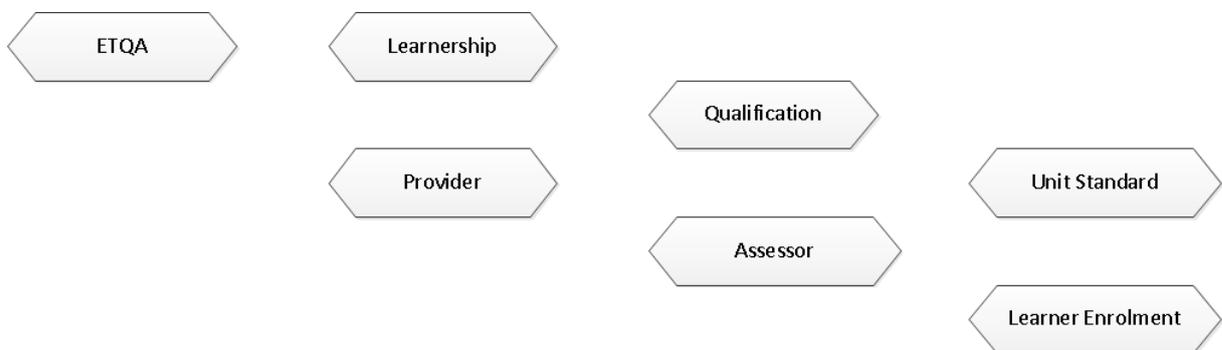


Figure A.1 Conceptual diagram of seven concepts of the National Qualifications Framework

### A.2 The National Qualifications Framework, Qualifications and Unit Standards

This subsection focuses on the definition of qualifications and unit standards and their relationship as contemplated in specific acts and regulations (see Figure A.2).



Figure A.2 Conceptual diagram of qualifications and unit standards

In 1995, the South African Qualifications Authority Act (No. 58 of 1995) was promulgated to establish the South African Qualifications Authority (SAQA) whose main task was to

oversee the development and implementation of the NQF (Ministry in the Office of the President, South African Qualifications Authority Act, Act 58 of 1995, p. 1). In this regard the Act declared that SAQA would perform the following functions (Ministry in the Office of the President, South African Qualifications Authority Act, Act 58 of 1995, p. 3):

*... oversee the development of the National Qualifications Framework; and formulate and publish policies and criteria for the registration of bodies responsible for establishing education and training standards or qualifications; and the accreditation of bodies responsible for monitoring and auditing achievements in terms of such standards or qualifications...*

*... oversee the implementation of the National Qualifications Framework, including the registration or accreditation of bodies and the assignment of functions to them; the registration of national standards and qualifications; steps to ensure compliance with provisions for accreditation; ...*

The Act made a clear distinction between two types of sub-structures, namely bodies that were “responsible for establishing education and training standards or qualifications” and bodies that were “responsible for monitoring and auditing achievements in terms of such standards or qualifications”. The implementation of the NQF in regard to the monitoring and auditing of achievements in terms of standards and or qualifications is the main focus of the further analysis provided in this section and the research to be conducted as a whole.

The Act defined a qualification to mean the “formal recognition of the achievement of the required number and range of credits and such other requirements at specific levels of the National Qualifications Framework as may be determined by the relevant bodies registered for such purpose by the South African Qualifications Authority” (Ministry in the Office of the President, South African Qualifications Authority Act, Act 58 of 1995, p. 1). The Act defined standards to mean “registered statements of desired education and training outcomes and their associated assessment criteria” (Ministry in the Office of the President, South African Qualifications Authority Act, Act 58 of 1995, p. 1). Furthermore the Act defined registered as having the meaning of “registered in terms of the National Qualifications Framework” (Ministry in the Office of the President, South African Qualifications Authority Act, Act 58 of 1995, p. 1).

The National Standards Bodies Regulations, 1998, provided the supporting and regulatory context for the implementation of education and training standards or qualifications required by the NQF. This regulation defined that National Standards Bodies (NSBs) would be “registered in terms of section 5(1)(a)(ii) of the Act, responsible for establishing education and training standards or qualifications, and to which specific functions relating to the registration of national standards and qualifications have been assigned in terms of section 5(1)(b)(i) of the Act” (South African Qualifications Authorityb, 1998, p. 2) The regulation defined a unit standard to mean “registered statements of desired education and training outcomes and their associated assessment criteria together with administrative and other information as specified in NSB regulations” (South African Qualifications Authorityb, 1998, p. 2).

As per the NSB regulations, unit standards and qualifications were registered on the NQF. Both unit standards and qualifications were allocated a registration status, a discrete term and a credit value (one credit is equivalent to ten notional hours of learning and 120 credits represent a year of study).

- The achievement of a unit standard was determined by the achievement of the specific outcomes and competencies, against which assessment criteria were set, of the unit standard.
- The achievement of a qualification is determined by the achievement of the number and range of unit standards defined for the qualification (a unit standard based qualification) or the achievement of the exit level outcomes defined for the qualification. A unit standard based qualification required that the learner not only achieved specific unit standards that had been defined for the qualification but also that the correct mix of unit standards was achieved in regard to core, fundamental and elective unit standards. When a learner engaged on a unit standard as a standalone unit of learning (i.e. outside of the scope of wanting to achieve a qualification) the registration status and term of the unit standard was applicable. When a learner engaged on a unit standard in order to acquire the required number of credits for a specific qualification then the registration status and term of the qualification is applicable.

### ***A.3 Learnerships***

This subsection focuses on the definition of learnerships and how they relate to qualifications as contemplated in the Skills Development Act, 1998 (see Figure A.3).

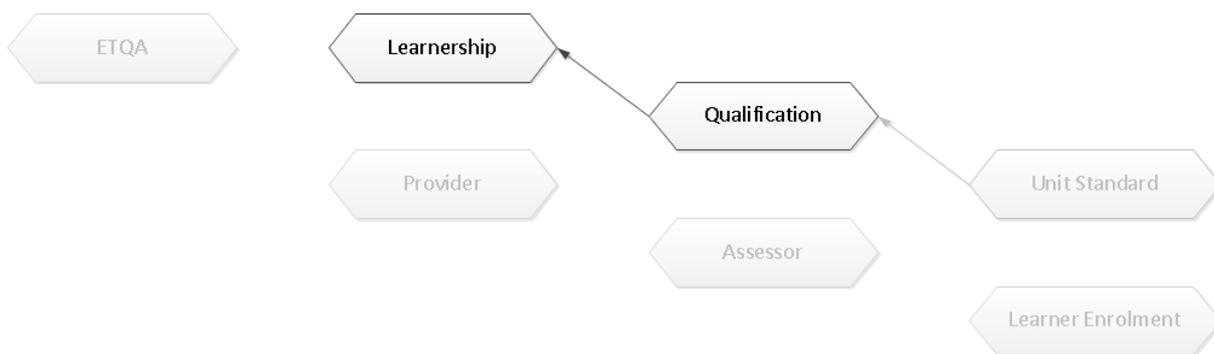


Figure A.3 Conceptual diagram of learnerships and their relationship to qualifications

The Skills Development Act, 1998, provided the supporting and regulatory context for the implementation of national, sectoral and workplace strategies, integrated with the NQF, to develop and improve the skills of the South African workforce (Ministry in the Office of the President, Skills Development Act, Act 97 of 1998, p. 8). This act defined a learnership to mean a structured learning component with practical work experience of a specific nature and duration, registered with the Director General of Labour, which would lead to the achievement of a qualification registered by SAQA (Ministry in the Office of the President, Skills Development Act, Act 97 of 1998, p. 20).

#### ***A.4 Education and Training Quality Assurance Bodies, providers and assessors***

This subsection focuses on the definition of Education and Training Quality Assurance Bodies (ETQAs), providers and assessors and their relationships as contemplated in the Education and Training Quality Assurance Bodies Regulations, 1998.

The Education and Training Quality Assurance Bodies Regulations, 1998, provided the supporting and regulatory context for the implementation of quality assurance systems and processes required by the NQF. This regulation defined that Education and Training Quality Assurance Bodies (ETQAs) would be “accredited in terms of section 5(1)(a)(ii) of the Act, responsible for monitoring and auditing achievements in terms of national standards or qualifications, and to which specific functions relating to the monitoring and auditing of national standards or qualifications have been assigned in terms of section 5(1)(b)(i) of the Act” (South African Qualifications Authority, 1998, p. 3).

As per the ETQA regulations, ETQAs were issued with a certificate of accreditation by SAQA (South African Qualifications Authority, 1998, p. 6). Further, ETQAs were also issued with a certificate of accreditation that stipulated the specific standards or qualifications for which accreditation had been granted (South African Qualifications Authority, 1998, p. 6). ETQA accreditations had a status (for example application, full or provisional etc.) and were allocated a discrete term. ETQA accreditations for specific standards or qualifications were also allocated a discrete term and, as a result of standards or qualifications being registered after the start date of an ETQA accreditation, could start after the accreditation of an ETQA, but would end on or before the end of an ETQA accreditation (see Figure A.4.1).

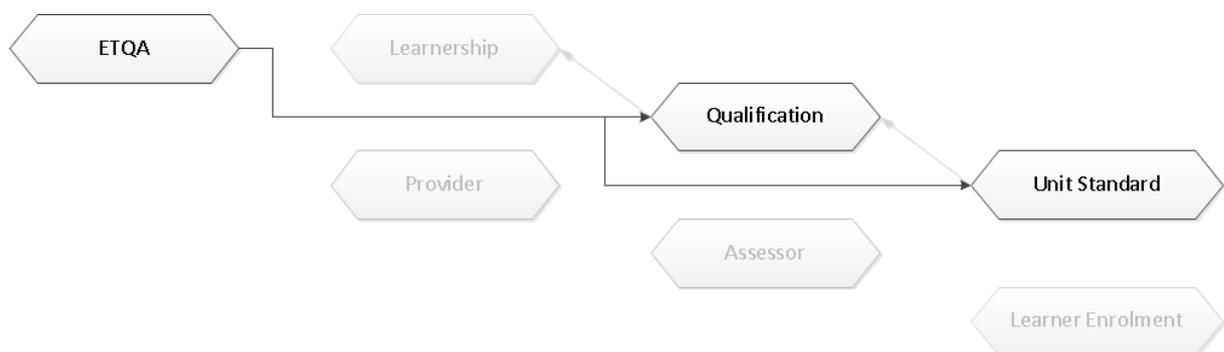


Figure A.4.1 Conceptual diagram of ETQAs and their relationship to qualifications and unit standards

The ETQA regulations stipulated that the functions of an ETQA included (South African Qualifications Authority, 1998, p. 8) (see Figure A.4.2):

- The accreditation of constituent providers for specific standards or qualifications registered on the NQF
- The monitoring of provision by constituent providers
- The registration of constituent assessors for specified registered standards or qualifications
- The maintenance of a database acceptable to SAQA
- The submission of reports to SAQA in accordance with the requirements of SAQA

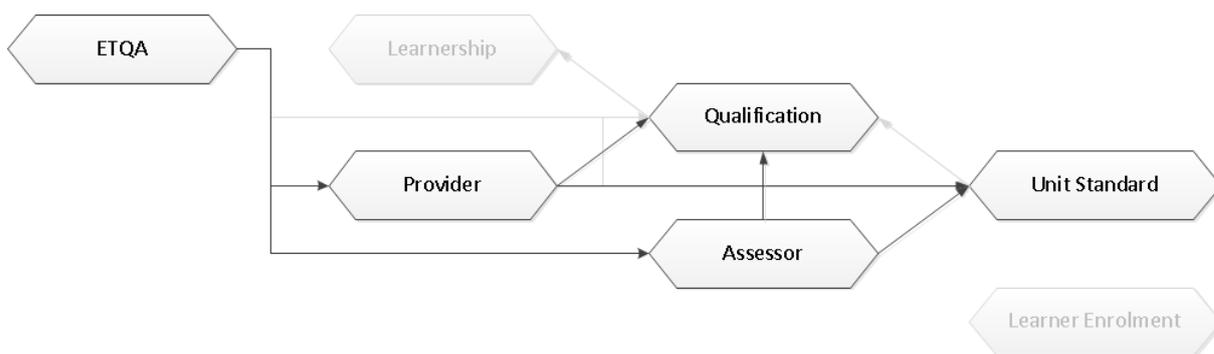


Figure A.4.2 Conceptual diagram of ETQAs and their relationship to providers and assessors

As per the ETQA regulations, providers were issued with a certificate of accreditation by the ETQA (South African Qualifications Authority, 1998, p. 10). Further, providers were also issued with a certificate of accreditation that stipulated the specific standards or qualifications for which accreditation had been granted (South African Qualifications Authority, 1998, p. 10). Provider accreditations had a status (for example application, full or provisional etc.) and were allocated a discrete term. Provider accreditations for specific standards or qualifications were also allocated a discrete term which could start on or after the start of the provider accreditation and would end on or before the end of the provider accreditation.

The requirement that every ETQA must submit data to the NLRD at least twice a year (South African Qualifications Authority, 2011, p. 1) in part addressed the requirements related to an ETQA needing to maintain a database that is acceptable to SAQA and submitting reports to SAQA in accordance with SAQA's requirements.

The Specifications for Load Files for the National Learners' Records Database Version 2.0 (NLRD specifications) provides the most descriptive and detailed interpretation of what SAQA required in regard to the accreditation of providers and the registrations of assessors.

In the NLRD specifications SAQA clearly defined the requirement for the accreditation of providers with specific accreditation statuses and a start and end date for the accreditation (South African Qualifications Authority, 2013, p. 7). Further, the NLRD specification detailed the requirement for the accreditation of providers to offer a specific learnership and/or qualification and/or unit standard with specific accreditation statuses and a start and end date for the accreditation (South African Qualifications Authority, 2013, p. 11). The NLRD specification also detailed the requirement to register assessors with specific

registration statuses and a start and end date for the registration (South African Qualifications Authorityd, 2013, p. 15) and the requirement to register assessors to assess a specific learnership and/or qualification and/or unit standard with specific registration statuses and a start and end date for the registration (South African Qualifications Authorityd, 2013, p. 16).

### A.5 *First version of the semantic business rules*

This subsection focuses on the learner enrolment and/or achievement and how they relate to concepts such as ETQAs, providers, assessors, learnerships, qualifications and unit standards (see Figure A.5). Further, based on the review conducted thus far a first version of the semantic business rules that form the core of this research is presented.

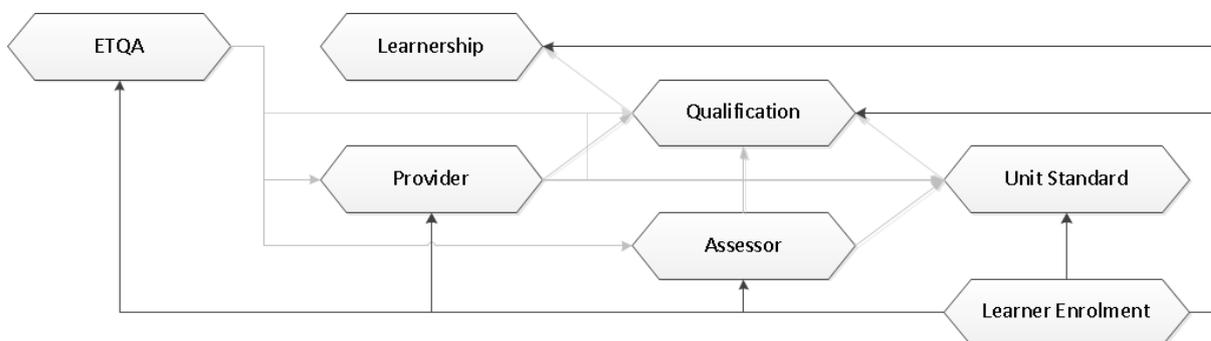


Figure A.5 Conceptual diagram of learner enrolments and their relationship to ETQAs, providers, assessors, learnerships, qualifications and unit standards

The discrete process of accrediting providers and registering assessors was not the overall goal for the implementation of quality assurance policies and mechanisms for the NQF. Rather the goal was to ensure that learners who were enrolled on or had achieved NQF registered qualifications and unit standards had received their training at a provider that had been accredited to offer the qualification or unit standard, and had been assessed by an assessor that was registered to assess the qualification or unit standard. To this end the NLRD specifications also required that the details of the provider and the assessor were recorded against each learnership enrolment record (South African Qualifications Authorityd, 2013, p. 18), qualification enrolment record (South African Qualifications Authorityd, 2013, p. 19) and unit standard enrolment record (South African Qualifications Authorityd, 2013, p. 20).

The requirements discussed thus far can be interpreted in the inverse to mean that for all learnership, qualification and unit standard enrolment records found on the NLRD the following semantic business rules should apply:

- that the ETQA that submitted the record
  - was accredited for the duration of the learner's active enrolment on the learnership/qualification/unit standard
  - was accredited to quality assure the qualification/unit standard for the duration of the learner's active enrolment on the learnership/qualification/unit standard
- that the provider
  - was accredited for the duration of the learner's active enrolment on the learnership/qualification/unit standard
  - was accredited to offer the qualification/unit standard for the duration of the learner's active enrolment on the learnership/qualification/unit standard
- that if the learner has completed the learnership or achieved the qualification/unit standard and the details of the assessor are supplied, that the assessor
  - was registered at the time of the completion of the learnership or achievement of the qualification/unit standard
  - was registered to assess the qualification/unit standard at the time of the completion of the learnership or achievement of the qualification/unit standard
- that the qualification/unit standard was registered for the duration of the learner's active enrolment on the qualification/unit standard
- that if the learner has completed the learnership, that due to the intrinsic nature of a learnership and qualification the learner would have achieved the qualification on or before the completion of the learnership
- that if the learner has achieved the qualification, and the qualification is a unit standards based qualification
  - the learner would have achieved the minimum required number of credits for the qualification, on or before the achievement of the qualification, based on the achievement of unit standards related to the qualification, and
  - the learner would have achieved the correct range of credits for the qualification, on or before the achievement of the qualification, based on the achievement of unit standards that have been defined as core, fundamental and elective unit standards for the qualification.

#### ***A.6 The impact of the National Qualifications Framework Act on the semantic business rules***

This subsection addresses the manner in which the National Qualifications Framework Act (No. 67 of 2008) impacts on the definition of the first version of the semantic business rules as defined in Appendix A.5.

The objective of the National Qualifications Framework Act (No. 67 of 2008), which succeeds the South African Qualifications Authority Act (No. 58 of 1995), is to provide for the further development of the NQF and applies to all education programmes or learning programmes that lead to qualifications or part qualifications that are offered in the Republic of South Africa by education institutions and skills development providers (Ministry in the Office of the President, National Qualifications Framework, Act 67 of 2008, p. 3).

The National Qualifications Framework Act (No. 67 of 2008) defines that, amongst others, SAQA's function is to:

*“... oversee the implementation of the NQF and ensure the achievement of its objectives.”* (Ministry in the Office of the President, National Qualifications Framework, Act 67 of 2008, p. 6)

*“... maintain a national learners' records database comprising registers of national qualifications, part qualifications, learner achievements, recognised professional bodies, professional designations and associated information;”* (Ministry in the Office of the President, National Qualifications Framework, Act 67 of 2008, p. 8)

The National Qualifications Framework Act (No. 67 of 2008) stipulates that all qualifications or part qualifications must be registered on the NQF (Ministry in the Office of the President, National Qualifications Framework, Act 67 of 2008, p. 3). The National Qualifications Framework Act (No. 67 of 2008) differs from the South African Qualifications Authority Act (No. 58 of 1995) in that it defines a qualification to mean a national qualification that has been registered on the NQF and has replaced the concept of a standard, as referred to in the South African Qualifications Authority Act (No. 58 of 1995), with the concept of a part qualification which means an assessed unit of learning that is registered as part of a qualification (which can be a module or a unit standard).

The National Qualifications Framework Act (No. 67 of 2008) indicates that the NQF is a single integrated system which comprises three qualification sub-frameworks namely General and Further Education Training and Training (as contemplated in the General and Further Education and Training Quality Assurance Act (Act 58 of 2001)), Higher Education (as contemplated in the Higher Education Act (Act 101 of 1997)) and Trades and Occupations (as contemplated in the Skills Development Act (Act 97 of 1998)) (Ministry in the Office of the President, National Qualifications Framework, Act 67 of 2008, p. 4), commonly known as the Occupational Qualifications Sub-framework (Ministry in the Office of the President, Determination of the sub-frameworks that comprise the National Qualifications Framework, Notice 1040 of 2012, p. 4). The National Qualifications Framework Act (No. 67 of 2008) goes further to introduce Quality Councils (QCs) for each of these sub-frameworks namely; Umalusi for General and Further Education Training and Training, the Council on Higher Education (CHE) for Higher Education and the Quality Council for Trades and Occupations (QCTO) (Ministry in the Office of the President, National Qualifications Framework, Act 67 of 2008, p. 13).

The National Qualifications Framework Act (No. 67 of 2008) defines that, amongst others, a QC's function is to (Ministry in the Office of the President, National Qualifications Framework, Act 67 of 2008, p. 13):

- “...perform its functions subject to this Act and the law by which the QC is established”*
- “...maintain a database of learner achievements and related matters...”*
- “...submit such data in a format determined in consultation with the SAQA for recording on the national learners' records database”*

The above indicates that implementation of ETQAs has been replaced with the implementation of QCs. Both Umalusi and the CHE were previously accredited as ETQAs whereas the QCTO is a newly established entity.

- A review of the General and Further Education and Training Quality Assurance Act (Act 58 of 2001) in regard to the semantic business rules, defined thus far based on the South African Qualifications Authority Act (No. 58 of 1995), indicates that concepts such as the registration of qualifications, accreditation of providers and the achievement of qualifications are similar to concepts described in the South African Qualifications Authority Act (No. 58 of 1995). The act however does not make specific reference to

assessors; rather the act makes reference to assessment bodies and providers that conduct assessments. The NLRD does not collect information in regard to assessment bodies and providers that conduct assessments.

- A review of the Higher Education Act (Act 101 of 1997) in regard to the semantic business rules, defined thus far based on the South African Qualifications Authority Act (No. 58 of 1995), indicates that concepts such as the registration of qualifications, accreditation of providers (referred to as institutions) and the achievement of qualifications are similar to concepts described in the South African Qualifications Authority Act (No. 58 of 1995). The act however does not make specific reference to an assessor; rather the act makes reference to the assessment of achievements by the institution. As stated previously the NLRD does not collect information in regard to institutions that conduct assessments.
- A review of the Skills Development Act (Act 97 of 1998) indicates that SETA ETQAs would continue to perform their duties as per the South African Qualifications Authority Act (No. 58 of 1995) until the QCTO delegated powers and functions to the SETA (Ministry in the Office of the President, Skills Development Act, Act 97 of 1998, p. 70). The QCTO delegated powers and functions to the SETAs as Quality Partners (QPs) in 2012 which included amongst others the accreditation of providers, registration of assessors, evaluation of the achievement of qualifications, the maintenance of a comprehensive learner management system and the submission of data to the NLRD. The QCTO has retained the function of registering qualifications and part qualifications on the NQF.

Since the implementation of the National Qualifications Framework Act (No. 67 of 2008) the NLRD specifications have only been amended to highlight the fact that the term “ETQA” is no longer an official acronym.

The above shows that the semantic business rules, defined thus far based on the South African Qualifications Authority Act (No. 58 of 1995), need only be adjusted to accommodate the National Qualifications Framework Act (No. 67 of 2008) in regard to the term “ETQA” (South African Qualifications Authority, 2013, p. 3).

Considering that the research that will be conducted will entail the analysis of data collected whilst both acts were in effect, a suitable term must be defined that encapsulates the ETQAs

that were established under the auspices of the South African Qualifications Authority Act (No. 58 of 1995), the QCs that have been established under the auspices of the National Qualifications Framework Act (No. 67 of 2008) and the QPs that have been delegated specific powers and functions by the QCTO. For the purposes of this research all three types of entities will be referred to as Education and Training Quality Entities (ETQEs) (see Figure A.6).

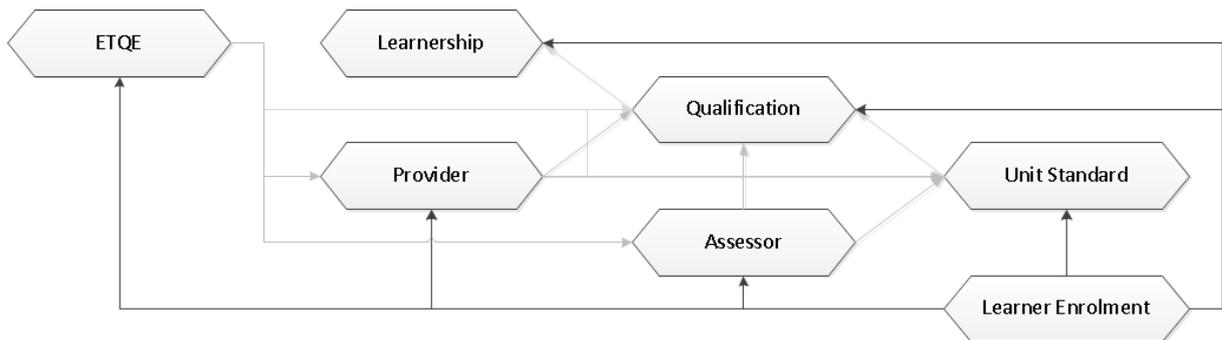


Figure A.6 Conceptual diagram of learner enrolments and their relationship to ETQEs, providers, assessors, learnerships, qualifications and unit standards

### A.7 Final version of the semantic business rules

This subsection details the final version of the semantic business rules that form the core of this research.

In consideration to the impact, as described in Appendix A.6 that the National Qualifications Framework Act (No. 67 of 2008) has on the first version of the semantic business rules, as described in Appendix A.6, the final version of the semantic business rules that form the core of this research for each learner enrolment record stored in the NLRD are as follows:

- that the ETQE that submitted the record
  - a. was accredited for the duration of the learner’s active enrolment on the learnership/qualification/unit standard
  - b. was accredited to quality assure the qualification/unit standard for the duration of the learner’s active enrolment on the qualification/unit standard
- that the provider
  - a. was accredited for the duration of the learner’s active enrolment on the learnership/qualification/unit standard

- b. was accredited to offer the qualification/unit standard for the duration of the learner's active enrolment on the learnership/qualification/unit standard
- that if the learner has completed the learnership or achieved the qualification/unit standard and the details of the assessor are supplied, that the assessor
  - a. was registered at the time of the completion of the learnership or achievement of the qualification/unit standard
  - b. was registered to assess the qualification/unit standard at the time of the completion of the learnership or achievement of the qualification/unit standard
- that the qualification/unit standard was registered for the duration of the learner's active enrolment on the qualification/unit standard
- that if the learner has completed the learnership, then due to the intrinsic nature of a learnership and qualification the learner would have achieved the qualification on or before the completion of the learnership
- that if the learner has achieved the qualification, and the qualification is a unit standards based qualification
  - a. the learner would have achieved the minimum required number of credits for the qualification, on or before the achievement of the qualification, based on the achievement of unit standards related to the qualification
  - b. the learner would have achieved the correct range of credits for the qualification, on or before the achievement of the qualification, based on the achievement of unit standards that have been defined as core, fundamental and elective unit standards for the qualification

#### ***A.8 Inflections of the semantic business rules in actuality***

This subsection addresses the implementation of the semantic business rules in actuality whilst considering the dynamic nature of the functioning the NQF concepts of ETQEs, providers, assessors, learnerships, qualifications and unit standards and their relationship to a learner's record of enrolment and/or achievement.

The semantic business rules provided in Appendix A.7 seem easily attainable in an environment where there is a segregated relationship between other ETQEs and a specific ETQE, its providers and assessors and the relationships between the learnerships,

qualifications and part qualifications that the ETQE quality assures. The dynamic nature of the NQF however allows for situations in which:

- A learner may engage on a learnership registered at one ETQE which requires the achievement of a qualification at another ETQE.
- A learner may engage on a qualification at a provider accredited by the specific ETQE which requires the achievement of unit standards that are offered by another provider which may be accredited by a different ETQE.
- A learner may engage on a qualification that is assessed by an assessor registered by the specific ETQE, which requires the achievement of unit standards that are assessed by other assessors which may be registered by a different ETQE.

In practice the above types of scenarios require that ETQEs need to communicate with each other to ensure that the enrolment or completion/achievement of a learnership, qualification or unit standard meets the quality requirements of the NQF.

Due to the complexity of these semantic business rules, combined with the volume of data that each ETQE manages, the implementation of a sophisticated information system at each ETQE is required. Further, due to the dynamic nature of the NQF, such information systems must also be able to accommodate the processing of the requirements within a framework where data records that support specific requirements do not originate from within the information system.

As already stated, every ETQE must submit data to the NLRD at least twice a year (South African Qualifications Authority, 2011, p. 1). The NLRD was established in 1999 by SAQA to support the implementation, monitoring and evaluation of the NQF in the Republic of South Africa. The NLRD provides SAQA with functionality that allows SAQA to:

- Combine education and training into a single framework, the NQF. This aspect of the NLRD is best described as an operational information system.
- Monitor and evaluate the implementation of the NQF. This aspect of the NLRD is best described as a data warehouse which is populated with data submitted to SAQA by ETQAs at least twice a year (South African Qualifications Authority, 2011, p. 1).

The operational information system and data warehousing components of the NLRD are integrated into a single information system.

The data warehouse aspect of the NLRD specifically collects data in regard to legacy and current learner achievement records, as well as all their associated records (learner enrolments, provider, provider accreditation, assessor, assessor registrations and professional designation data), from both the public and private education sectors in South Africa. The data is collected by means of an Extract-Transform-Load (ETL) process with the following discrete steps:

1. The data is extracted and transformed at the ETQE information system.
2. The data is verified by the ETQE to ensure that it conforms to the minimum data standard requirements of the NLRD. If the data fails to meet the minimum data standard requirements of the NLRD then the ETQE cleans the data submission and re-verifies the submission until the data submission achieves the minimum data standards required.
3. The data submission is then transferred to the NLRD data warehouse domain.
4. Once received by the NLRD data loading team the data submission is quality assured and the data is loaded into the NLRD.

The second step of the NLRD ETL is achieved with the implementation of a data validation application called Edu.Dex (Praxis Computing, 2012, p. 3). The main objective of this application is to ensure that data submissions to the NLRD meet the minimum requirements of the NLRD. The minimum requirements of the NLRD, documented as the Specifications for Load Files for the National Learners' Records Database Version 2.0, do not include rules related to the inverse requirements described in this document. This is not an accidental omission on the part of SAQA, but rather a decision that was made in regard to the overall nature of collection of data for the NLRD which includes salient considerations such as:

- The NQF and all its supporting structures and policies were new structures that were implemented over a number of years. Further, a common understanding of what these new policies and structures entailed was only gained after their practical implementation within the education sector. As a result, data that was created on or around the inception of many of these structures may not necessarily comply with the requirements.
- The NLRD ETL process needs to accommodate the processing of legacy and current learner enrolments and achievements. Legacy records include records that were created prior to the implementation of the acts and regulations that stipulate these requirements and as a result cannot be expected to meet such requirements.
- In order to eliminate duplication of data within the NLRD, each ETQE may only submit data records pertaining to its own domain of operation. This requirement, in combination

with the dynamic nature of the NQF however means that a submission from an ETQE may include references to other data records that are not included in the data submission.

It is reasonable to assume that the NLRD contains records that legitimately infringe the inverse requirements described in this document. However no research has been conducted to determine whether there are records that illegitimately infringe these requirements, the nature and scope of these types of records and the reasons why such records exist.

### ***A.9 Appendix summary***

This section provided a review of the legislative, policy and regulatory framework that relates to the learner enrolment records that can be found on the NLRD. In this regard this section detailed the following discrete NQF concepts as contemplated in various acts, legislations and regulations:

- The definition of qualifications and unit standards and the relationship between qualifications and unit standards.
- The definition of learnerships and how they relate to qualifications.
- The definition of ETQAs, providers and assessors and the relationship between these concepts.

Having defined these concepts, their relationship to learner enrolment/achievement records was explored, and based on this understanding the ten (10) semantic business rules that form the core of this research were defined. Finally, this section briefly touched on the challenges faced with the implementation of the semantic business rules in actuality which in turn forms the basis of research problem for this research.

## Appendix B

This appendix provides a detailed specification of the raw data tables received from the NLRD for this research.

Table Name	Column Name	Data Type	Data Length	Allow NULLs	Comment
DM_ASOR	ASSESSOR_ID	NUMBER	22	Y	De-identified
DM_ASOR	START_DATE	DATE	7	Y	
DM_ASOR	END_DATE	DATE	7	Y	
DM_ASOR_REGSTR	ASSESSOR_ID	NUMBER	22	Y	De-identified
DM_ASOR_REGSTR	LEARNERSHIP_ID	NUMBER	22	Y	De-identified
DM_ASOR_REGSTR	QUALIFICATION_ID	NUMBER	22	Y	De-identified
DM_ASOR_REGSTR	UNIT_STANDARD_ID	NUMBER	22	Y	De-identified
DM_ASOR_REGSTR	START_DATE	DATE	7	Y	
DM_ASOR_REGSTR	END_DATE	DATE	7	Y	
DM_ETQE	ETQE_ID	NUMBER	22	Y	De-identified
DM_ETQE	START_DATE	DATE	7	Y	
DM_ETQE	END_DATE	DATE	7	Y	
DM_ETQE	ORGANISATION_TYPE_ID	NUMBER	22	Y	
DM_ETQE	ORGANISATION_TYPE_DESC	VARCHAR2	60	Y	
DM_ETQE_ACCRED	ETQE_ID	NUMBER	22	Y	De-identified
DM_ETQE_ACCRED	QUALIFICATION_ID	NUMBER	22	Y	De-identified
DM_ETQE_ACCRED	UNIT_STANDARD_ID	NUMBER	22	Y	De-identified
DM_ETQE_ACCRED	START_DATE	DATE	7	Y	
DM_ETQE_ACCRED	END_DATE	DATE	7	Y	
DM_ETQE_START	ETQE_ID	NUMBER	22	N	De-identified
DM_ETQE_START	START_DATE	DATE	7	Y	
DM_LSHP	LEARNERSHIP_ID	NUMBER	22	N	De-identified
DM_LSHP	ETQE_ID	NUMBER	22	Y	De-identified
DM_LSHP	NQF_LEVEL_ID	NUMBER	22	Y	
DM_LSHP	NQF_LEVEL_DESC	VARCHAR2	26	Y	
DM_LSHP_ENROL	LEARNER_ENROLMENT_ID	NUMBER	22	N	De-identified
DM_LSHP_ENROL	LEARNER_ID	NUMBER	22	N	De-identified
DM_LSHP_ENROL	LEARNERSHIP_ID	NUMBER	22	Y	De-identified
DM_LSHP_ENROL	ETQE_ID	NUMBER	22	N	De-identified
DM_LSHP_ENROL	PROVIDER_ID	NUMBER	22	Y	De-identified
DM_LSHP_ENROL	ASSESSOR_ID	NUMBER	22	Y	De-identified
DM_LSHP_ENROL	ENROL_STATUS_ID	NUMBER	22	N	
DM_LSHP_ENROL	ENROL_STATUS_DESC	VARCHAR2	26	N	
DM_LSHP_ENROL	ENROL_TYPE_ID	NUMBER	22	N	
DM_LSHP_ENROL	ENROL_TYPE_DESC	VARCHAR2	50	N	
DM_LSHP_ENROL	ENROL_DATE	DATE	7	Y	
DM_LSHP_ENROL	ACHIEVE_DATE	DATE	7	Y	
DM_LSHP_ENROL	DERIVED_START_DATE	DATE	7	Y	

Table Name	Column Name	Data Type	Data Length	Allow NULLs	Comment
DM_LSHP_ETQE	ETQE_ID	NUMBER	22	Y	De-identified
DM_LSHP_ETQE	LEARNERSHIP_ID	NUMBER	22	N	De-identified
DM_PROV	PROVIDER_ID	NUMBER	22	Y	De-identified
DM_PROV	START_DATE	DATE	7	Y	
DM_PROV	END_DATE	DATE	7	Y	
DM_PROV	ETQE_ID	NUMBER	22	Y	De-identified
DM_PROV	PROVIDER_TYPE_ID	NUMBER	22	Y	
DM_PROV	PROVIDER_TYPE_DESC	VARCHAR2	26	Y	
DM_PROV	PROVIDER_CLASS_ID	NUMBER	22	Y	
DM_PROV	PROVIDER_CLASS_DESC	VARCHAR2	50	Y	
DM_PROV	PROVINCE_CODE	VARCHAR2	10	Y	
DM_PROV	PROVINCE_DESC	VARCHAR2	60	Y	
DM_PROV_ACCRED	PROVIDER_ID	NUMBER	22	Y	De-identified
DM_PROV_ACCRED	LEARNERSHIP_ID	NUMBER	22	Y	De-identified
DM_PROV_ACCRED	QUALIFICATION_ID	NUMBER	22	Y	De-identified
DM_PROV_ACCRED	UNIT_STANDARD_ID	NUMBER	22	Y	De-identified
DM_PROV_ACCRED	START_DATE	DATE	7	Y	
DM_PROV_ACCRED	END_DATE	DATE	7	Y	
DM_QUAL	QUALIFICATION_ID	NUMBER	22	Y	De-identified
DM_QUAL	START_DATE	DATE	7	Y	
DM_QUAL	END_DATE	DATE	7	Y	
DM_QUAL	CREDITS	NUMBER	22	Y	
DM_QUAL	TRANSITION_PERIOD	NUMBER	22	Y	
DM_QUAL	TRAIN_OUT_PERIOD	NUMBER	22	Y	
DM_QUAL	NQF_LEVEL_ID	NUMBER	22	Y	
DM_QUAL	NQF_LEVEL_DESC	VARCHAR2	26	Y	
DM_QUAL	QUALIFICATION_TYPE_ID	NUMBER	22	Y	
DM_QUAL	QUALIFICATION_TYPE_DESC	VARCHAR2	30	Y	
DM_QUAL	DEFAULT_G2_TRAIN_OUT_PERIOD	NUMBER	22	Y	
DM_QUAL	QUALIFICATION_CLASS_ID	NUMBER	22	Y	
DM_QUAL	QUALIFICATION_CLASS_DESC	VARCHAR2	26	Y	
DM_QUAL	FIELD_ID	NUMBER	22	Y	
DM_QUAL	FIELD_DESC	VARCHAR2	60	Y	
DM_QUAL	SUBFIELD_ID	NUMBER	22	Y	
DM_QUAL	SUBFIELD_DESC	VARCHAR2	80	Y	
DM_QUAL_ENROL	LEARNER_ENROLMENT_ID	NUMBER	22	N	De-identified
DM_QUAL_ENROL	LEARNER_ID	NUMBER	22	N	De-identified
DM_QUAL_ENROL	QUALIFICATION_ID	NUMBER	22	Y	De-identified
DM_QUAL_ENROL	LEARNERSHIP_ID	NUMBER	22	Y	De-identified
DM_QUAL_ENROL	ETQE_ID	NUMBER	22	N	De-identified
DM_QUAL_ENROL	PROVIDER_ID	NUMBER	22	Y	De-identified
DM_QUAL_ENROL	ASSESSOR_ID	NUMBER	22	Y	De-identified
DM_QUAL_ENROL	ENROL_STATUS_ID	NUMBER	22	N	
DM_QUAL_ENROL	ENROL_STATUS_DESC	VARCHAR2	26	N	

Table Name	Column Name	Data Type	Data Length	Allow NULLs	Comment
DM_QUAL_ENROL	ENROL_TYPE_ID	NUMBER	22	N	
DM_QUAL_ENROL	ENROL_TYPE_DESC	VARCHAR2	50	N	
DM_QUAL_ENROL	ENROL_DATE	DATE	7	Y	
DM_QUAL_ENROL	ACHIEVE_DATE	DATE	7	Y	
DM_QUAL_ENROL	DERIVED_START_DATE	DATE	7	Y	
DM_QUAL_LSHP	LEARNERSHIP_ID	NUMBER	22	Y	De-identified
DM_QUAL_LSHP	QUALIFICATION_ID	NUMBER	22	Y	De-identified
DM_QUAL_REPL	OLD_QUAL_ID	NUMBER	22	N	De-identified
DM_QUAL_REPL	NEW_QUAL_ID	NUMBER	22	N	De-identified
DM_USTD	UNIT_STANDARD_ID	NUMBER	22	Y	De-identified
DM_USTD	START_DATE	DATE	7	Y	
DM_USTD	END_DATE	DATE	7	Y	
DM_USTD	CREDITS	NUMBER	22	Y	
DM_USTD	TRANSITION_PERIOD	NUMBER	22	Y	
DM_USTD	TRAIN_OUT_PERIOD	NUMBER	22	Y	
DM_USTD	NQF_LEVEL_ID	NUMBER	22	Y	
DM_USTD	NQF_LEVEL_DESC	VARCHAR2	26	Y	
DM_USTD	UNIT_STD_TYPE_ID	NUMBER	22	Y	
DM_USTD	UNIT_STD_TYPE_DESC	VARCHAR2	26	Y	
DM_USTD	FIELD_ID	NUMBER	22	Y	
DM_USTD	FIELD_DESC	VARCHAR2	60	Y	
DM_USTD	SUBFIELD_ID	NUMBER	22	Y	
DM_USTD	SUBFIELD_DESC	VARCHAR2	80	Y	
DM_USTD_ENROL	LEARNER_ENROLMENT_ID	NUMBER	22	N	De-identified
DM_USTD_ENROL	LEARNER_ID	NUMBER	22	N	De-identified
DM_USTD_ENROL	UNIT_STANDARD_ID	NUMBER	22	Y	De-identified
DM_USTD_ENROL	QUALIFICATION_ID	NUMBER	22	Y	De-identified
DM_USTD_ENROL	ETQE_ID	NUMBER	22	N	De-identified
DM_USTD_ENROL	PROVIDER_ID	NUMBER	22	Y	De-identified
DM_USTD_ENROL	ASSESSOR_ID	NUMBER	22	Y	De-identified
DM_USTD_ENROL	ENROL_STATUS_ID	NUMBER	22	N	
DM_USTD_ENROL	ENROL_STATUS_DESC	VARCHAR2	26	N	
DM_USTD_ENROL	ENROL_TYPE_ID	NUMBER	22	N	
DM_USTD_ENROL	ENROL_TYPE_DESC	VARCHAR2	50	N	
DM_USTD_ENROL	ENROL_DATE	DATE	7	Y	
DM_USTD_ENROL	ACHIEVE_DATE	DATE	7	Y	
DM_USTD_ENROL	DERIVED_START_DATE	DATE	7	Y	
DM_USTD_QUAL	QUALIFICATION_ID	NUMBER	22	Y	De-identified
DM_USTD_QUAL	UNIT_STANDARD_ID	NUMBER	22	Y	De-identified
DM_USTD_QUAL	USTD_QUAL_TYPE_CODE	VARCHAR2	2	Y	
DM_USTD_QUAL	USTD_QUAL_TYPE_DESC	VARCHAR2	26	Y	
DM_USTD_REPL	OLD_USTD_ID	NUMBER	22	N	De-identified
DM_USTD_REPL	NEW_USTD_ID	NUMBER	22	N	De-identified

## Appendix C

## ***C.1 Introduction***

This section details the initial selection, pre-processing and derivation of the learnership enrolment records, received from the NLRD in the table DM\_LSHP\_ENROL, into a format that is suitable for data mining.

The specific semantic business rules that are applicable to learnership enrolment records are identified in Appendix C.2. The analysis and data mining of these semantic business rules requires the implementation of four (4) semantic business rule indicators. Appendix C.3 describes the selection, pre-processing and derivation steps required for the implementation of these semantic business rule indicators. Appendix C.3.1 and Appendix C.3.2 describe the type of logic developed for the selection and pre-processing of the data. Whereas Appendix C.3.4 to Appendix C.3.7 describe the type of logic used for the derivation of the data.

The selection, pre-processing and derivation logic resulted in the implementation of a final version of the learnership enrolment data as a new table called DM\_LSHP\_ENROL\_FINAL, described in Appendix C.3.8.

## ***C.2 Applicable semantic business rules and their indicator fields***

A review of the final version of the semantic business rules (see Section 3.6.2) shows that the business rules that are applicable to learnership enrolment records are as follows:

1. that the ETQE that submitted the record
  - a. was accredited for the duration of the learner's active enrolment on the learnership/qualification/unit standard
  - ...
2. that the provider
  - a. was accredited for the duration of the learner's active enrolment on the learnership/qualification/unit standard
  - ...
3. that if the learner has completed the learnership or achieved the qualification/unit standard and the details of the assessor are supplied, that the assessor
  - a. was registered at the time of the completion of the learnership or achievement of the qualification/unit standard
  - ...

- ...
- that if the learner has completed the learnership, then due to the intrinsic nature of a learnership and qualification the learner would have achieved the qualification on or before the completion of the learnership
- ...

The main purpose of the derivation of the learnership enrolment data for analysis and data mining therefore focused on the development of four (4) semantic business rule indicators (each consisting of a data code and a description) that described the compliance of the record in accordance with these rules:

- ETQE\_IND

Denotes whether the ETQE was accredited for the duration of the learner's active enrolment on the learnership.

- PROV\_IND

Denotes whether the provider was accredited for the duration of the learner's active enrolment on the learnership.

- ASOR\_IND

Denotes whether the assessor was registered at the time of the completion of the learnership.

- QENROL\_IND

Denotes whether, when a learner has completed a learnership, a corresponding qualification achievement record has been submitted to the NLRD.

### ***C.3 Semantic business rule indicator development steps***

#### ***C.3.1 Pre derivation data collection***

Determining compliance of a data record in regard to all of the semantic business rules that are applicable to learnership enrolment records required that each learnership enrolment record have an active enrolment time period. An active enrolment time period needed to be derived for learnership enrolment records that did not have an enrolment date and/or a completion date (Section 3.6.4.1). Deriving the active enrolment time period for these types of enrolment records was accomplished utilizing the calculation of credits to notional hours

(Appendix A.2) using the credits of the qualification that the learnership is linked to (Section 1.4.1).

As a result the first step of the development of the semantic business rule indicators for the learnership enrolment records focused on determining which qualification the learnership enrolment record is linked to. The determination of which qualification is linked to a learnership is guided by the data stored in the table DM\_QUAL\_LSHP (see Appendix E.3.13). The utilization of this type of data however needed to consider following:

- a learnership could be linked to more than one qualification,
- the qualification that a learnership is linked to may not be active (and as a result the credits of the qualification cannot be determined), and
- the qualification that a learnership is linked to may have been replaced.

To ensure that the logic had access to records for both the replaced qualifications and the qualifications that replaced them, additional qualification learnership link records were created using records found in DM\_QUAL\_REPL (the table that records qualification replacements, see Section 3.8.2.14). Where the qualification that a learnership was linked to was found as an “old” qualification (OLD\_QUAL\_ID) in DM\_QUAL\_REPL, a new qualification learnership link was created using the “new” qualification (NEW\_QUAL\_ID). These derived qualification learnership links and the original qualification learnership links found in DM\_QUAL\_LSHP were saved to a new table called DM\_QUAL\_LSHP\_FINAL. The contents of this new table were used to determine the link between a learnership and a qualification.

The implementation of the table DM\_QUAL\_LSHP\_FINAL successfully addressed learnerships linked to qualifications that have been replaced. The implementation of this table also partially addressed learnerships that are linked to qualifications that were not active. However, the implementation of this table also aggravated the number of qualifications that a learnership is linked to. As a result a large amount of the derivation logic had to focus on the elimination of the resultant duplications in a manner that ensured that the learnership enrolment record was linked to the correct or most desirable qualification record for processing. The elimination of these duplicates included the following processing logic, in the order provided, for all learnership enrolment records that linked to more than one qualification:

- If one of the qualifications did not have any credits, then the qualification with credits was retained for processing.
- A check was completed to see whether the learner had enrolled on any of the remaining qualification links. The match was completed as follows:
  - The LEARNER\_IDs, LEARNERSHIP\_IDs and QUALIFICATION\_IDs of the learnership enrolment record and qualification enrolment record are the same.
  - The LEARNER\_IDs and QUALIFICATION\_IDs of the learnership enrolment record and the qualification enrolment record are the same (in other words the LEARNERSHIP\_ID for the qualification enrolment record is either NULL or has a value other than the LEARNERSHIP\_ID of the learnership enrolment record).

The qualification link that resolved to a qualification enrolment record was retained and any qualification links that did not resolve to a qualification enrolment record were discarded.

This strategy highlights that a number of learners had possibly enrolled on two qualifications for the same learnership. For the purposes of the learnership enrolment analysis these records were allocated a PROBLEM\_ID code of 1 and excluded from the further processing of the data.

For records where no qualification enrolment records could be found, the following type of logic was implemented to eliminate duplicate qualification links:

- Any unduplicated qualification link, where the learner enrolled on the learnership after the registration start date of the qualification and before the end date of the registration date of the qualification, was retained.
- Any unduplicated qualification link with the highest credit value was retained.
- For all remaining duplicates, the qualification link with the newest qualification identifier value was retained.

The reader should note that the above-mentioned logic was implemented only to determine the credits for the qualification linked to a learnership in order to utilize this value to derive the start date and end date of the learnership enrolment record. The overall number of records that were retained for the analysis of learnership enrolment records was only

impacted by those records that were allocated a `PROBLEM_ID` code of 1. These records constituted 1.31% of the learnership enrolment records initially extracted from the NLRD.

As stated above the implementation of the table `DM_QUAL_LSHP_FINAL` only eliminated some of the issues related to learnerships being linked to qualifications that were not active. Learnership enrolment records where the learnership has either not been linked to a qualification or the linkage between a learnership and qualification has never been active was allocated a `PROBLEM_ID` of 2 and excluded from the further processing of the data. These records constituted 3.93% of the learnership enrolment records initially extracted from the NLRD. Further, any learnership enrolment records where the learnership has been linked to a qualification that has never been active was allocated a `PROBLEM_ID` of 3 and excluded from the further processing of the data. These records constituted 1.91% of the learnership enrolment records initially extracted from the NLRD.

The linking of the qualification to the learnership enrolment record resulted in the addition of the following data fields to the table `DM_LSHP_ENROL`:

1. `QUALIFICATION_ID`: the qualification identifier of the qualification utilized to derive the active enrolment time period of the learnership enrolment record if required.
2. `CREDITS`: the credits for the qualification utilized to derive the active enrolment time period of the learnership enrolment record if required.
3. `QUAL_START_DATE`: the active registration start date of the qualification utilized to derive the active enrolment time period of the learnership enrolment record if required.
4. `QUAL_END_DATE`: the active registration end date of the qualification utilized to derive the active enrolment time period of the learnership enrolment record if required
5. `PROBLEM_ID`: A nominal data value used to indicate problem records that needed to be excluded from further processing and analysis.

### ***C.3.2 Deriving the active enrolment time period***

Once an active qualification was linked to the learnership enrolment record, the derivation logic focused on deriving the active enrolment period for the learnership enrolment record.

Two new indicators were created namely; a nominal data value and a corresponding descriptive data value used to record whether the start date of the learnership enrolment record represented;

1. the enrolment date as provided in the learnership enrolment record,
2. the learnership enrolment record did not have an enrolment date and was as a result derived from the combination of the learnership achievement date and the qualification credits (see Section 3.6.4.3.a), or
3. that the learnership enrolment record did not have an enrolment date or an achievement date and was as a result derived from the derived start date of the enrolment (see Section 3.6.4.3.a).

Additionally a new data field was created to store the derived start date of the learnership enrolment record based on the above.

Once a start date was implemented as described above, an end date for the active enrolment time period was implemented either as:

- the actual completion date of the learnership enrolment record, or
- a derived end date calculated using the combination of the start date for the enrolment record and the qualification credits (see Section 3.6.4.1.b).

This resulted in the addition of the following indicators and data fields on the table DM\_LSHP\_ENROL:

6. START\_DATE\_ID: A nominal data code, and
7. START\_DATE\_DESC: A corresponding descriptive data value indicating whether the value in START\_DATE represents:
  - an enrolment date (ENROL\_DATE),
  - a derived value utilizing the achievement date (ACHIEVE\_DATE) for the enrolment record and the qualification credits (CREDITS), or
  - a derived value utilizing the derived start date (DERIVED\_START\_DATE) of the record.
8. START\_DATE: The start date of the active enrolment time period.
9. END\_DATE: The derived end date of the active enrolment time period, representing either the value found in ACHIEVE\_DATE or a value derived from START\_DATE and CREDITS.

A number of the semantic business rules are dependent on the active enrolment period of the record. Additionally an analysis of the qualification data (DM\_QUAL) shows that the earliest registration for a qualification occurred on 30 June 2000. As a result of the intrinsic

nature between learnerships and qualifications it can be deduced that no learnership could have been registered before 30 June 2000. As a direct consequence it could further be deduced that any learnership enrolment record with a start date less than 30 June 2000 was an outlier record. Such records by their very nature were considered erroneous and needed to be excluded from the research. These records were allocated a PROBLEM\_ID code of 4 and excluded from the further processing of the data. These records constituted 0.04% of the learnership enrolment records initially extracted from the NLRD.

### *C.3.3 Core data required for the development of the indicator fields and additional data values*

Having established the active enrolment time period of the learnership enrolment record, the derivation process focused on the:

- collection of the core data required for the development of the semantic business rule indicators described in Appendix C.2, and
- the collection of additional data fields that may prove valuable to analysis of the learnership enrolment data as described in Section 3.6.5.

The reader should note that the data received from the NLRD was, with the exception of lookup values, provided in a format that closely represents a relational database design. As an example, even though the learnership enrolment table (DM\_LSHP\_ENROL) contained a unique identifier for the learnership (LEARNERSHIP\_ID), the learnership enrolment table did not contain additional data fields that describe the learnership, for example the NQF Level of the learnership.

This section describes which data fields sourced from other tables were added to the learnership enrolment table (DM\_LSHP\_ENROL) and how the linkage between the learnership enrolment record and the other tables were implemented.

Data that describes the learnership was obtained from the learnership table (DM\_LSHP) using the unique identifier of the learnership (LEARNERSHIP\_ID).

10. LSHP\_ETQE\_ID (ETQE\_ID on DM\_LSHP): The ETQE that is mandated to implement the learnership.

11. NQF\_LEVEL\_ID and NQF\_LEVEL\_DESC: The data code and corresponding description of the NQF Level of the learnership.

Data that describes the accreditation of the ETQE was obtained from the ETQE accreditation table (DM\_ETQE) using the unique identifier of the ETQE (ETQE\_ID).

12. ETQE\_START\_DATE (START\_DATE on DM\_ETQE): The start date of the accreditation of the ETQE that submitted the enrolment record to the NLRD.
13. ETQE\_END\_DATE (END\_DATE on DM\_ETQE): The end date of the accreditation of the ETQE that submitted the enrolment record to the NLRD.

Data that describes the provider and its accreditation as obtained from the provider accreditation table (DM\_PROV) using the unique identifier of the provider (PROVIDER\_ID).

14. PROV\_START\_DATE (START\_DATE on DM\_PROV): The start date of the accreditation of the provider that offered the learnership.
15. PROV\_END\_DATE (END\_DATE on DM\_PROV): The end date of the accreditation of the provider that offered the learnership.
16. PROV\_ETQE\_ID (ETQE\_ID on DM\_PROV): The primary ETQE of the provider.
17. PROVIDER\_TYPE\_ID and PROVIDER\_TYPE\_DESC: The data code and corresponding description of the provider type.
18. PROVIDER\_CLASS\_ID and PROVIDER\_CLASS\_DESC: The data code and corresponding description of the provider class.
19. PROV\_PROVINCE\_CODE (PROVINCE\_CODE on DM\_PROV) and PROV\_PROVINCE\_DESC (PROVINCE\_DESC on DM\_PROV): The data code and corresponding description of the province that the provider is located in.

Data that describes the registration of the assessor as obtained from the assessor registration table (DM\_ASOR) using the unique identifier of the assessor (ASSESSOR\_ID).

20. ASOR\_START\_DATE (START\_DATE on DM\_ASOR): The start date of the registration of the assessor that assessed the learnership completion.
21. ASOR\_END\_DATE (END\_DATE on DM\_ASOR): The end date of the registration of the assessor that assessed the learnership completion.

Data that describes the qualification enrolment that is linked to the learnership as obtained from the qualification enrolment table (DM\_QUAL\_ENROL) using the unique identifier of the learner (LEARNER\_ID) and the following type of logic; in order of processing (the

logic described below is required because the correct linkage between a learnership enrolment record and its related qualification enrolment record utilizing the data field LEARNERSHIP\_ID in the qualification enrolment record is not guaranteed. This is a data management issue that is actively being addressed by the NLRD Director with the ETQEs):

The qualification identifier (QUALIFICATION\_ID) as derived in Appendix C.3.1 is assumed to be correct:

- A match is sought on learnership enrolment QUALIFICATION\_ID and the qualification enrolment QUALIFICATION\_ID and a match between the learnership enrolment LEARNERSHIP\_ID and qualification enrolment LEARNERSHIP\_ID.
- A mismatch between the learnership enrolment QUALIFICATION\_ID and the qualification enrolment QUALIFICATION\_ID and a match between the learnership enrolment LEARNERSHIP\_ID and qualification enrolment LEARNERSHIP\_ID.
- A match is sought on learnership enrolment QUALIFICATION\_ID and the qualification enrolment QUALIFICATION\_ID where the qualification enrolment LEARNERSHIP\_ID is NULL.
- A match is sought on learnership enrolment QUALIFICATION\_ID and the qualification enrolment QUALIFICATION\_ID and there is a mismatch between the learnership enrolment LEARNERSHIP\_ID and qualification enrolment LEARNERSHIP\_ID.

The qualification identifier (QUALIFICATION\_ID) as derived in Appendix C.3.1 is assumed to be incorrect because the learner has enrolled on a replacement qualification, therefore matching is performed using the table DM\_QUAL\_LSHP\_FINAL

22. A match is sought on learnership enrolment QUALIFICATION\_ID and the qualification enrolment QUALIFICATION\_ID where the qualification enrolment LEARNERSHIP\_ID is NULL.
23. A match is sought on learnership enrolment QUALIFICATION\_ID and the qualification enrolment QUALIFICATION\_ID and there is a mismatch between the learnership enrolment LEARNERSHIP\_ID and qualification enrolment LEARNERSHIP\_ID.
24. QENROL\_LEARNERSHIP\_ID (LEARNERSHIP\_ID on DM\_QUAL\_ENROL): The learnership identifier recorded against the qualification enrolment record.

25. QENROL\_QUALIFICATION\_ID (QUALIFICATION\_ID on DM\_QUAL\_ENROL):  
The qualification identifier of the qualification enrolment record.
26. QENROL\_ENROL\_STATUS\_ID (ENROL\_STATUS\_ID on DM\_QUAL\_ENROL)  
and QENROL\_ENROL\_STATUS\_DESC (ENROL\_STATUS\_DESC on  
DM\_QUAL\_ENROL): The enrolment status code and corresponding description of the  
qualification enrolment record.
27. QENROL\_ENROL\_DATE (ENROL\_DATE on DM\_QUAL\_ENROL): The enrolment  
date of the qualification enrolment record.
28. QENROL\_ACHIEVE\_DATE (ACHIEVE\_DATE on DM\_QUAL\_ENROL): The  
achievement date of the qualification enrolment record.

The date on which an ETQE submitted its first full data submission to the NLRD as obtained from the table DM\_ETQE\_START using the unique identifier of the ETQE (ETQE\_ID) (see Section 3.8.3.1).

29. ETQE\_FIRST\_DATE (START\_DATE on DM\_ETQE\_START): The first date on  
which the ETQE submitted a full submission to the NLRD.

The date on which the primary ETQE of the provider submitted its first full data submission to the NLRD as obtained from the table DM\_ETQE\_START using the unique identifier of the ETQE of the provider (PROV\_ETQE\_ID) (see Section 3.8.3.1 and Section 3.8.3.5).

30. PROV\_ETQE\_START (START\_DATE on DM\_ETQE\_START): The first date on  
which the primary ETQE of the provider submitted a full submission to the NLRD.

The date of the most recent NLRD data submission cycle as obtained from the Director of the NLRD (see Section 3.8.3.2).

31. CYCLE\_DATE (variable that is set at execution of the script): The date of the most  
recent NLRD data submission cycle.

Two calculated indicators that represent the time delta between the first registration date of a qualification on the NQF and the start and end date of the learnership enrolment record (Section 3.6.4.4). In other words these fields represent the time that has elapsed from the first time that a qualification was registered on the NQF and the start and end date of the learnership enrolment record:

32. START\_DATE\_IND: The difference in whole months, as a rounded down value, between START\_DATE and 2000/06/30.

33. END\_DATE\_IND: The difference in whole months, as a rounded down value, between END\_DATE and 2000/06/30.

### C.3.4 Development of ETQE\_IND

As detailed in Appendix C.2, the development of ETQE\_IND required the implementation of an indicator that denotes whether the ETQE was accredited for the duration of the learner's active enrolment on the learnership. In order to make the analysis of the data more suitable for data mining purposes the indicator was implemented as a data code and corresponding description.

Figure C.3.4.1 illustrates the manner in which ETQE\_IND was developed using five example learnership enrolment records, for an ETQE that has not been amalgamated. The figure shows how:

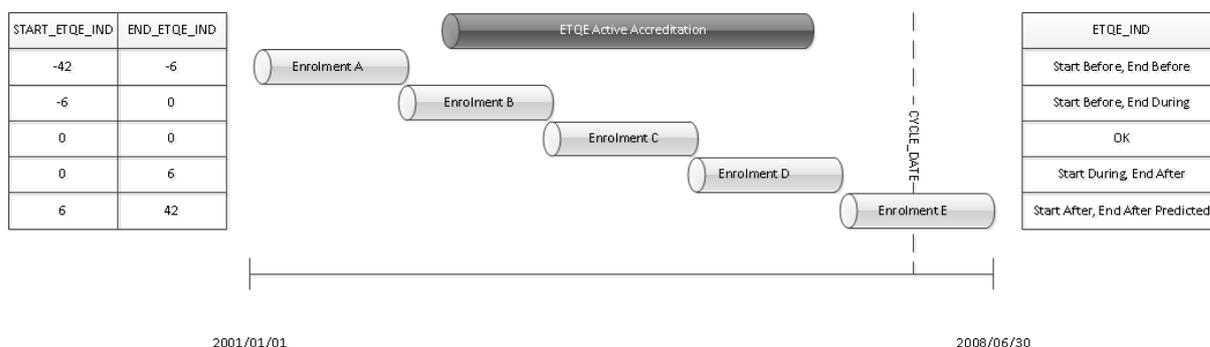


Figure C.3.4.1 Illustrative diagram of ETQE\_IND development

- a learnership enrolment record (Enrolment A) with a start date prior to the accreditation period of the ETQE and an end date prior to the accreditation period of the ETQE is allocated an ETQE\_IND value of 'Start Before, End Before',
- a learnership enrolment record (Enrolment B) with a start date prior to the accreditation period of the ETQE and an end date during the accreditation period of the ETQE is allocated an ETQE\_IND value of 'Start Before, End During',
- a learnership enrolment record (Enrolment C) with a start date during the accreditation period of the ETQE and an end date during the accreditation period of the ETQE is allocated an ETQE\_IND value of 'OK',

- a learnership enrolment record (Enrolment D) with a start date during the accreditation period of the ETQE and an end date after the accreditation period of the ETQE is allocated an ETQE\_IND value of ‘Start During, End After’, and
- a learnership enrolment record (Enrolment E) with a start date after the accreditation period of the ETQE and an end date after the accreditation period of the ETQE is allocated an ETQE\_IND value of ‘Start After, End After’, and because the end of the active enrolment time period exceeds the latest data submission cycle date the word ‘Predicted’ is appended to the value.

The development of the ETQE\_IND indicator required the implementation of two additional indicators. These indicators assisted in the development of and further description of the value allocated to ETQE\_IND. Both of these two additional indicators were developed as representations of data in relation to a point in time as discussed in Section 3.6.4.4.

The first indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the start date (START\_DATE) of the active enrolment time period and the start date of the ETQE’s active accreditation time period (ETQE\_START\_DATE and ETQE\_END\_DATE), where;

- a learnership enrolment record with a start date before the start date of the ETQE’s accreditation would be given a negative value of the number of months between these two values (for example Enrolment A on Figure C.3.4.1),
- a learnership enrolment record with a start date that falls between the start and end dates of the ETQE’s accreditation is given a value of 0 (for example Enrolment C on Figure C.3.4.1), and
- a learnership enrolment record with a start date that is after the end date of the ETQE’s accreditation would be given a positive value of the number of months between these two values (for example Enrolment E on Figure C.3.4.1).

The second indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the end date (END\_DATE) of the active enrolment time period and the end date of the ETQE’s active accreditation time period (ETQE\_START\_DATE and ETQE\_END\_DATE), where;

- a learnership enrolment record with an end date before the start date of the ETQE's accreditation would be given a negative value of the number of months between these two values (for example Enrolment A on Figure C.3.4.1),
- a learnership enrolment record with an end date that falls between the start and end dates of the ETQE's accreditation is given a value of 0 (for example Enrolment C on Figure C.3.4.1), and
- a learnership enrolment record with an end date that is after the end date of the ETQE's accreditation would be given a positive value of the number of months between these two values (for example Enrolment E on Figure C.3.4.1).

This logic resulted in the addition of the following new indicators on the table DM\_LSHP\_ENROL:

34. START\_ETQE\_IND: Numeric value indicating the distance between the start date of the learnership enrolment record and the ETQE accreditation.
35. END\_ETQE\_IND: Numeric value indicating the distance between the end date of the learnership enrolment record and the ETQE accreditation.

Using the values in these two fields in conjunction with each other it was possible to derive a code and corresponding description for ETQE\_IND by:

- Allocating a value of 'OK' to records where START\_ETQE\_IND is equal to 0 and END\_ETQE\_IND is equal to 0.
- For all remaining records
- Allocating a value of 'Start Before' to records where START\_ETQE\_IND was less than 0, 'Start During' to records where START\_ETQE\_IND is equal to 0 and 'Start After' were START\_ETQE\_IND was greater than 0.
- Allocating a value 'End Before' to records where END\_ETQE\_IND was less than 0, 'End During' to records where END\_ETQE\_IND is equal to 0 and 'End After' where END\_ETQE\_IND was greater than 0.

This logic resulted in the addition of the ETQE\_IND indicator code and corresponding description on the table DM\_LSHP\_ENROL.

36. ETQE\_IND\_ID and ETQE\_IND\_DESC: Code and corresponding description denoting a record's compliance in regard to whether the ETQE was accredited for the duration of the learner's active enrolment on the learnership

The above mentioned logic however did not take into consideration the accreditation of the ETQE that was also mandated to implement the learnership in situations where ETQEs had been amalgamated (see Section 3.8.3.3). In order to address this issue, the logic determined whether the submitting ETQE differed from the ETQE defined on the learnership record (ETQE\_ID on DM\_LSHP) or whether a different ETQE had in the past been mandated to implement the learnership (DM\_LSHP\_ETQE).

The ETQE identifier, start date and end date of the ETQE that was also mandated to implement the learnership was amended to the table DM\_LSHP\_ENROL:

37. OTHR\_ETQE\_ID: The ETQE identifier.
38. OTHR\_ETQE\_START\_DATE (START\_DATE on DM\_ETQE): The start date of the accreditation of the ETQE.
39. OTHR\_ETQE\_END\_DATE (END\_DATE on DM\_ETQE): The end date of the accreditation of the ETQE.

Four indicators were developed in the same manner as described for START\_ETQE\_IND, END\_ETQE\_IND, ETQE\_IND\_ID and ETQE\_IND\_DESC, using the indicators OTHR\_ETQE\_START\_DATE, OTHR\_ETQE\_END\_DATE, OTHR\_START\_ETQE\_IND and OTHR\_END\_ETQE\_IND in place of the indicators ETQE\_START\_DATE, ETQE\_END\_DATE, START\_ETQE\_IND and END\_ETQE\_IND.

This logic resulted in the addition of the following new indicators on the table DM\_LSHP\_ENROL:

40. OTHR\_START\_ETQE\_IND: Numeric value indicating the distance between the start date of the learnership enrolment record and the other ETQE's accreditation.
41. OTHR\_END\_ETQE\_IND: Numeric value indicating the distance between the end date of the learnership enrolment record and the other ETQE's accreditation.
42. OTHR\_ETQE\_IND\_ID and OTHR\_ETQE\_IND\_DESC: Code and corresponding description denoting a record's compliance in regard to whether the other ETQE was accredited for the duration of the learner's active enrolment on the learnership.

The results of both the ETQE\_IND\_ID and ETQE\_IND\_DESC fields and the OTHR\_ETQE\_IND\_ID and OTHR\_ETQE\_IND\_DESC fields were then consolidated into the ETQE\_IND indicators in the following manner:

- Any record that was found to be compliant based on the value stored in ETQE\_IND\_ID or OTHR\_ETQE\_IND\_ID was marked as compliant.
- Any record that was found to be non-compliant based on both the value stored in ETQE\_IND\_ID and OTHR\_ETQE\_IND\_ID was provided a modified code and corresponding description that show the results of the compliance result of both ETQE\_IND\_ID and OTHR\_ETQE\_IND\_ID.

The final derivation step entailed amending the ETQE\_IND data code and corresponding description to differentiate records with a calculated end date that is greater than the latest data submission cycle date from other records (see Section 3.8.3.2). As a result the data code was amended and the word ‘Predicted’ was appended to the ETQE\_IND indicator description for any records with an END\_DATE value greater than CYCLE\_DATE (for example Enrolment E on Figure C.3.4.1).

### ***C.3.5 Development of PROV\_IND***

As detailed in Appendix C.2, the development of PROV\_IND required the implementation of an indicator that denotes whether the provider was accredited for the duration of the learner’s active enrolment on the learnership. In order to make the analysis of the data more suitable for data mining purposes the indicator was implemented as a data code and corresponding description.

Figure C.3.5.1 illustrates the manner in which PROV\_IND was developed using five example learnership enrolment records, for a provider that was accredited and is not an ‘ETQE Provider’. The figure shows how:

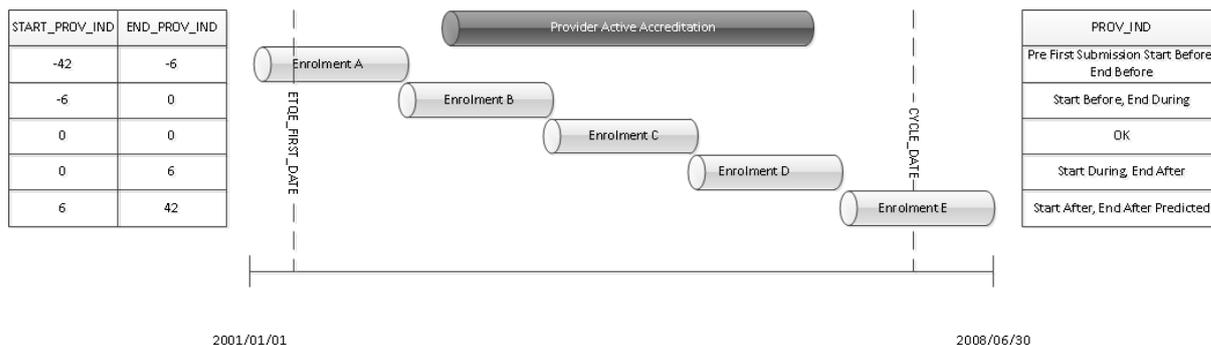


Figure C.3.5.1 Illustrative diagram of PROV\_IND development

- a learnership enrolment record (Enrolment A) with a start date prior to the provider's accreditation period and an end date prior to the provider's accreditation period is allocated a PROV\_IND value of 'Start Before, End Before'. The start of the active enrolment time period precedes the date on which the ETQE submitted its first full data submission to the NLRD, as a result the words 'Pre First Submission' are appended to the value,
- a learnership enrolment record (Enrolment B) with a start date prior to the accreditation period of the provider and an end date during the accreditation period of the provider is allocated a PROV\_IND value of 'Start Before, End During',
- a learnership enrolment record (Enrolment C) with a start date during the accreditation period of the provider and an end date during the accreditation period of the provider is allocated a PROV\_IND value of 'OK',
- a learnership enrolment record (Enrolment D) with a start date during the accreditation period of the provider and an end date after the accreditation period of the provider is allocated a PROV\_IND value of 'Start During, End After', and
- a learnership enrolment record (Enrolment E) with a start date after the accreditation period of the provider and an end date after the accreditation period of the provider is allocated a PROV\_IND value of 'Start After, End After'. The end of the active enrolment time period exceeds the latest data submission cycle date, as a result the word 'Predicted' is appended to the value.

The development of the PROV\_IND indicator required the implementation of two additional indicators. These indicators assisted in the development of and further description of the value allocated to PROV\_IND. Both of these two additional indicators

were developed as representations of data in relation to a point in time as discussed in Section 3.6.4.4.

The first indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the start date (START\_DATE) of the active enrolment time period and the start date of the provider's active accreditation time period (PROV\_START\_DATE and PROV\_END\_DATE), where;

- a learnership enrolment record with a start date before the start date of the provider's accreditation would be given a negative value of the number of months between these two values (for example Enrolment A on Figure C.3.5.1),
- a learnership enrolment record with a start date that falls between the start and end dates of the provider's accreditation is given a value of 0 (for example Enrolment C on Figure C.3.5.1), and
- a learnership enrolment record with a start date that is after the end date of the provider's accreditation would be given a positive value of the number of months between these two values (for example Enrolment E on Figure C.3.5.1).

The second indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the end date (END\_DATE) of the active enrolment time period and the end date of the provider's active accreditation time period (PROV\_START\_DATE and PROV\_END\_DATE)), where;

- a learnership enrolment record with an end date before the start date of the provider's accreditation would be given a negative value of the number of months between these two values (for example Enrolment A on Figure C.3.5.1),
- a learnership enrolment record with an end date that falls between the start and end dates of the provider's accreditation is given a value of 0 (for example Enrolment C on Figure C.3.5.1), and
- a learnership enrolment record with an end date that is after the end date of the provider's accreditation would be given a positive value of the number of months between these two values (for example Enrolment E on Figure C.3.5.1).

This logic resulted in the addition of the following new indicators on the table DM\_LSHP\_ENROL:

43. START\_PROV\_IND: Numeric value indicating the distance between the start date of the learnership enrolment record and the provider accreditation.
44. END\_PROV\_IND: Numeric value indicating the distance between the end date of the learnership enrolment record and the provider accreditation.

Using the values in these two fields in conjunction with each other it was possible to derive a code and corresponding description for PROV\_IND as follows:

- Where a provider is an 'ETQE provider' (see Section 3.8.3.5), allocating a value of 'ETQE Provider'
- Where a provider accreditation did not exist, allocating a value of 'No Accreditation' to the record.
- Allocating a value of 'OK' to records where START\_PROV\_IND is equal to 0 and END\_PROV\_IND is equal to 0.

For all remaining records:

- Allocating a value of 'Start Before' to records where START\_PROV\_IND was less than 0, 'Start During' to records where START\_PROV\_IND is equal to 0 and 'Start After' where START\_PROV\_IND was greater than 0.
- Allocating a value 'End Before' to records where END\_PROV\_IND was less than 0, 'End During' to records where END\_PROV\_IND is equal to 0 and 'End After' where END\_PROV\_IND was greater than 0.

The final derivation steps included:

- Amending the data code and appending the words 'Pre First Submission' to the indicator value for any records with a START\_DATE value less than PROV\_ETQE\_FIRST\_DATE (for example Enrolment A on Figure C.3.5.1). In other words all records where the learner enrolled on the learnership prior to the first full data submission from the primary ETQE of the provider to the NLRD (see Section 3.8.3.1 and Section 3.8.3.5).
- Amending the data code and appending the word 'Predicted' to the indicator value for any records with a END\_DATE value greater than CYCLE\_DATE (for example Enrolment E on Figure C.3.5.1). In other words all records with a calculated end date that is greater than the latest data submission cycle date (see Section 3.8.3.2).

This logic resulted in the addition of the PROV\_IND indicator code and corresponding description on the table DM\_LSHP\_ENROL.

45. PROV\_IND\_ID and PROV\_IND\_DESC: Code and corresponding description denoting a record's compliance in regard to whether the provider was accredited for the duration of the learner's active enrolment on the learnership.

### C.3.6 Development of ASOR\_IND

As detailed in Appendix C.2, the development of ASOR\_IND required the implementation of an indicator that denotes whether the assessor was registered at the time of the completion of the learnership. In order to make the analysis of the data more suitable for data mining purposes the indicator was implemented as a data code and corresponding description.

Figure C.3.6.1 illustrates the manner in which ASOR\_IND was developed using four example completed learnership enrolment records, for an assessor that was registered. The figure shows how:

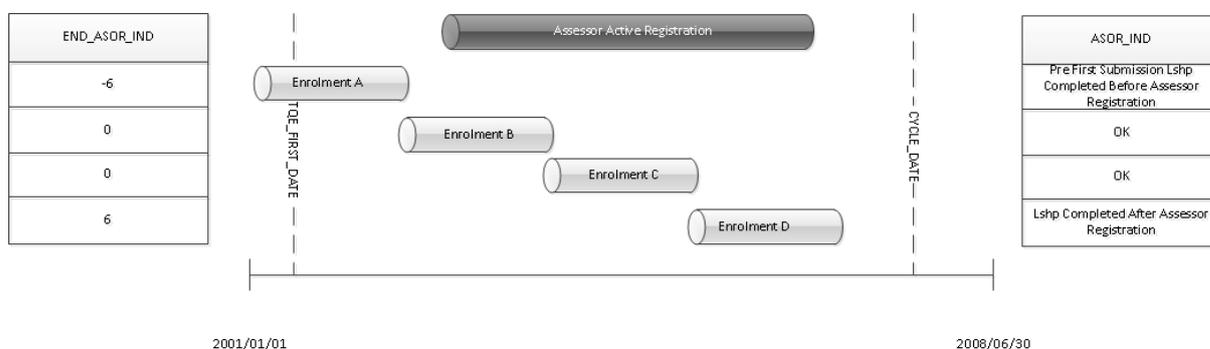


Figure C.3.6.1 Illustrative diagram of ASOR\_IND development

- a learnership enrolment record (Enrolment A) with an end date prior to the registration period of the assessor is allocated an ASOR\_IND value of 'Lshp Completed Before Assessor Registration'. The start of the active enrolment time period precedes the date on which the ETQE submitted its first full data submission to the NLRD, as a result the words 'Pre First Submission' are appended to the value,
- a learnership enrolment record (Enrolment B) with an end date during the registration period of the assessor is allocated an ASOR\_IND value of 'OK',

- a learnership enrolment record (Enrolment C) end date during the registration period of the assessor is allocated an ASOR\_IND value of 'OK', and
- a learnership enrolment record (Enrolment D) with an end date after the registration period of the assessor is allocated an ASOR\_IND value of 'Lshp Completed After Assessor Registration'.

The development of the ASOR\_IND indicator required the implementation of one additional indicator. This indicator assisted in the development of and further description of the value allocated to ASOR\_IND. This additional indicator was developed as a representation of data in relation to a point in time as discussed in Section 3.6.4.4.

The indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the end date (END\_DATE) of the active enrolment time period and assessor's active registration time period (ASOR\_START\_DATE and ASOR\_END\_DATE), where;

- a learnership enrolment record with an end date before the start date of the assessor's registration would be given a negative value of the number of months between these two values (for example Enrolment A on Figure C.3.6.1),
- a learnership enrolment record with an end date that falls between the start and end dates of the assessor's registration is given a value of 0 (for example Enrolment B on Figure C.3.6.1), and
- a learnership enrolment record with an end date that is after the end date of the assessor's registration would be given a positive value of the number of months between these two values (for example Enrolment D on Figure C.3.6.1).

This logic resulted in the addition of the following new indicator on the table DM\_LSHP\_ENROL:

46. END\_ASOR\_IND: Numeric value indicating the distance between the end date of the learnership enrolment record and the assessor registration.

Using the values in this field it was possible to derive a code and corresponding description for ASOR\_IND as follows:

- Where the learnership enrolment had not been completed, allocating a value of 'Not Completed'.

- Where the learnership enrolment had been completed but an assessor identifier had not been provided, allocation a value of 'No Assessor Provided'.
- Where an assessor identifier had been provided but an assessor registration did not exist, allocating a value of 'No Registration' to the record.
- Allocating a value of 'OK' to records where END\_ASOR\_IND is equal to 0.
- Allocating a value of 'Lshp Completed Before Assessor Registration' where END\_ASOR\_IND is less than 0.
- Allocating a value of 'Lshp Completed After Assessor Registration' where END\_ASOR\_IND is greater than 0.

The final derivation steps included:

- Amending the data code and appending the words 'Pre First Submission' to the indicator value for any records with a START\_DATE value less than ETQE\_FIRST\_DATE (for example Enrolment A on Figure C.3.6.1). In other words all records where the learner enrolled on the learnership prior to the first full data submission from the ETQE to the NLRD (see Section 3.8.3.1).

This logic resulted in the addition of the ASOR\_IND indicator code and corresponding description on the table DM\_LSHP\_ENROL.

47. ASOR\_IND\_ID and ASOR\_IND\_DESC: Code and corresponding description denoting a record's compliance in regard to whether the assessor was registered at the time of the completion of the learnership.

### ***C.3.7 Development of QENROL\_IND***

As detailed in Appendix C.2, the development of QENROL\_IND required the implementation of an indicator that denotes whether, when the learner has completed a learnership, a corresponding qualification achievement record has been submitted to the NLRD. In order to make the analysis of the data more suitable for data mining purposes the indicator was implemented as a data code and corresponding description.

Figure C.3.7.1 illustrates the manner in which QENROL\_IND was developed using five example learnership and qualification enrolment records, where the learnership has been completed and the qualification has been achieved. The figure shows how:

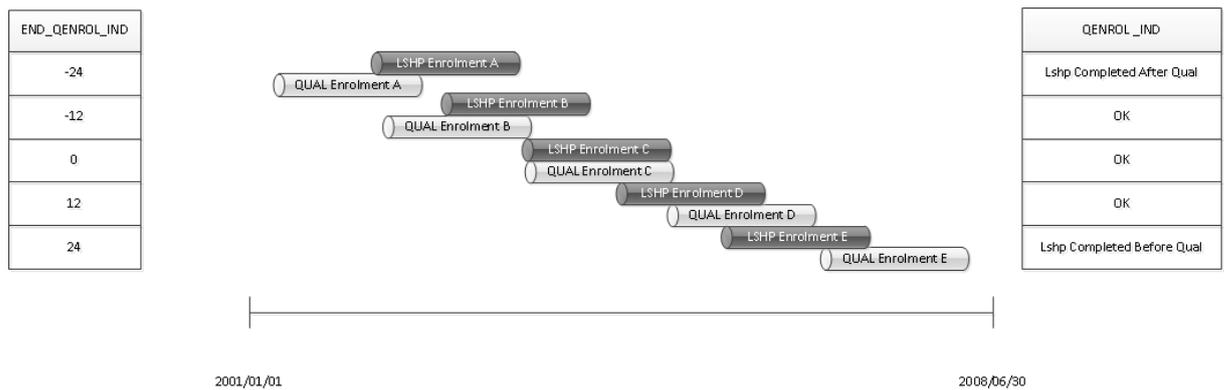


Figure C.3.7.1 Illustrative diagram of QUAL\_IND development

- a learnership enrolment record (LSHP Enrolment A) and its corresponding qualification enrolment record (QUAL Enrolment A) with a qualification enrolment end date that is more than a year prior to the learnership end date is allocated a QENROL\_IND value of 'Lshp Completed After Qual',
- a learnership enrolment record (LSHP Enrolment B) and its corresponding qualification enrolment record (QUAL Enrolment B) with a qualification enrolment end date that is within a year of the learnership enrolment end date is allocated a QENROL\_IND value of 'OK',
- a learnership enrolment record (LSHP Enrolment C) and its corresponding qualification enrolment record (QUAL Enrolment C) with a qualification enrolment end date that is within a year of the learnership enrolment end date is allocated a QENROL\_IND value of 'OK',
- a learnership enrolment record (LSHP Enrolment D) and its corresponding qualification enrolment record (QUAL Enrolment D) with a qualification enrolment end date that is within a year of the learnership enrolment end date is allocated a QENROL\_IND value of 'OK, and
- a learnership enrolment record (LSHP Enrolment E) and its corresponding qualification enrolment record (QUAL Enrolment E) with a qualification end date that is more than a year after the learnership enrolment end date is allocated a QENROL\_IND value of 'Lshp Completed Before Qual'.

The nature of the relationship between the completion of a learnership enrolment and the achievement of a qualification has a number of permutations that are not immediately

recognized on review of the semantic business rule. A complete overview of the possible scenarios is as follows:

- A related qualification enrolment record exists and the qualification enrolment record correctly records the learnership that it is linked to (see Section 3.6.3.5). In other words, the qualification enrolment record contains the same learnership identifier in its LEARNERSHIP\_ID field as the learnership enrolment record.
  - a) The learnership was completed and the qualification was achieved
    - i) The learnership was completed at the same time as the qualification was achieved
    - ii) The learnership was completed before the qualification was achieved
    - iii) The learnership was completed after the qualification was achieved
  - b) The learnership was completed and the qualification was not achieved
  - c) The learnership was not completed and the qualification was achieved
  - d) The learnership was not completed and the qualification was not achieved
- A related qualification enrolment record exists and the qualification enrolment record either incorrectly records the learnership that it is linked to or does not record the learnership that it is linked to (see Section 3.6.3.5). In other words, the qualification enrolment record does not contain the same learnership identifier in its LEARNERSHIP\_ID field as the learnership enrolment record or the LEARNERSHIP\_ID field on the qualification enrolment record is NULL.
  - a) The learnership was completed and the qualification was achieved
    - i) The learnership was completed at the same time as the qualification was achieved
    - ii) The learnership was completed before the qualification was achieved
    - iii) The learnership was completed after the qualification was achieved
  - b) The learnership was completed and the qualification was not achieved
  - c) The learnership was not completed and the qualification was achieved
  - d) The learnership was not completed and the qualification was not achieved

- A related qualification enrolment record does not exist.

In consultation with the Director of the NLRD it was decided that the logic should include a built in tolerance of 12 months for determining compliance to both instances of “The learnership was completed at the same time as the qualification was achieved” above. In other words a record was defined as being compliant for this business rule if the difference between the completion date of the learnership enrolment record and the achievement date of the qualification enrolment record has a distance of no more than 12 months. Further, it was decided that all compliance issues that were determined needed to differentiate between records where the linkage on the qualification enrolment record to the learnership had been correctly recorded and where the linkage on the qualification enrolment record to the learnership had been incorrectly recorded or omitted.

The development of this indicator focused on the learnership completion status, the qualification achievement status and the end dates of both the learnership enrolment record and the qualification enrolment record.

The development of the QENROL\_IND indicator required the implementation of one additional indicator. This indicator assisted in the development of and further description of the value allocated to QENROL\_IND. This additional indicator was developed as a representation of data in relation to a point in time as discussed in Section 3.6.4.4.

The indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the end date (END\_DATE) of the active enrolment time period of a completed learnership and the achievement date of the qualification enrolment record (QENROL\_ACHIEVE\_DATE), where;

- a completed learnership enrolment record with an end date before the achievement date of the linked qualification would be given a negative value of the number of months between these two values (for example LSHP Enrolment A and QUAL Enrolment A on Figure C.3.7.1),
- a completed learnership enrolment record with an end date that is equal to the achievement date of the linked qualification was given a value of 0 (for example LSHP Enrolment C and QUAL Enrolment C on Figure C.3.7.1), and

- a completed learnership enrolment record with an end date that is after the achievement date of the linked qualification would be given a positive value of the number of months between these two values (for example LSHP Enrolment E and QUAL Enrolment E on Figure C.3.7.1).

This logic resulted in the addition of the following new indicator on the table DM\_LSHP\_ENROL:

48. END\_QENROL\_IND: Numeric value indicating the distance between the end date of the learnership enrolment record and the qualification achievement date.

Using the values in these this field in conjunction with the learnership completion status, the qualification enrolment achievement status, the learnership identifier of the learnership enrolment record and the learnership identifier of the qualification enrolment record it was possible to derive a code and corresponding description for QENROL\_IND as follows:

- Where the learnership enrolment record was completed and the qualification enrolment record was achieved, allocating a value of 'OK' to records where END\_QENROL\_IND is greater than or equal to -12 and END\_QENROL\_IND is less than or equal to 12.
- Where the learnership enrolment record was completed and the qualification enrolment record was achieved and END\_QENROL\_IND had a value greater than 12, allocating a value of 'Lshp Completed After Qual'. If the learnership identifier of the learnership enrolment record differed from the learnership identifier on the qualification enrolment record the allocated value included the text '(Derived)'.
- Where the learnership enrolment record was completed and the qualification enrolment record was achieved and END\_QENROL\_IND had a value less than -12, allocating a value of 'Lshp Completed Before Qual'. If the learnership identifier of the learnership enrolment record differed from the learnership identifier on the qualification enrolment record the allocated value included the text '(Derived)'.
- Where the learnership enrolment record was not completed and the qualification enrolment record was not achieved, allocating a value of 'Both Lshp Not Completed and Qual Not Achieved'. If the learnership identifier of the learnership enrolment record differed from the learnership identifier on the qualification enrolment record the allocated value included the text '(Derived)'.

- Where the learnership enrolment record was not completed and the qualification enrolment record was achieved, allocating a value of 'Lshp Enrolled, Qual Achieved'. If the learnership identifier of the learnership enrolment record differed from the learnership identifier on the qualification enrolment record the allocated value included the text '(Derived)'.
- Where the learnership enrolment record was completed and the qualification enrolment record was not achieved, allocating a value of 'Lshp Completed, Qual Enrolled'. If the learnership identifier of the learnership enrolment record differed from the learnership identifier on the qualification enrolment record the allocated value included the text '(Derived)'.
- Where no qualification enrolment record could be found, allocating a value of 'No Qual Enrolment'.

This logic resulted in the addition of the QENROL\_IND indicator code and corresponding description on the table DM\_LSHP\_ENROL.

49. QENROL\_IND\_ID and QENROL\_IND\_DESC: Code and corresponding description denoting a record's compliance in regard to whether the intrinsic relationship between the completion of a learnership and achievement of its related qualification has been upheld.

### ***C.3.8 DM\_LSHP\_ENROL\_FINAL***

The derivation steps described from Appendix C.3.1 to Appendix C.3.7 were saved in a new data table called DM\_LSHP\_ENROL\_FINAL. This table included all of the data records initially received from the NLRD in the table DM\_LSHP\_ENROL, including the problem records described in Appendix C.3.1 and Appendix C.3.2 (i.e. records that have a value in the data field PROBLEM\_ID). The problem records were immediately communicated to SAQA, who in turn implemented processes and procedures in order to address these records.

A technical description of the table DM\_LSHP\_ENROL\_FINAL has been provided in C.1.

## ***C.4 Appendix summary***

This section detailed the development of the data table DM\_LSHP\_ENROL\_FINAL which will be used in the analysis and data mining of the learnership enrolment records received

from the NLRD. The section identified the semantic business rules that are applicable to learnership enrolment records and then described the selection, pre-processing and derivation steps implemented to establish the table DM\_LSHP\_ENROL\_FINAL, which contains the learnership enrolment records in a format that is suited for data mining.

## Appendix D

This appendix provides a detailed specification of the table DM\_LSHP\_ENROL\_FINAL.

Table Name	Column Name	Data Type	Data Length	Allow NULLs	Comment
DM_LSHP_ENROL_FINAL	LEARNER_ENROLMENT_ID	NUMBER	22	N	De-identified
DM_LSHP_ENROL_FINAL	LEARNER_ID	NUMBER	22	N	De-identified
DM_LSHP_ENROL_FINAL	LEARNERSHIP_ID	NUMBER	22	Y	De-identified
DM_LSHP_ENROL_FINAL	ETQE_ID	NUMBER	22	N	De-identified
DM_LSHP_ENROL_FINAL	PROVIDER_ID	NUMBER	22	Y	De-identified
DM_LSHP_ENROL_FINAL	ASSESSOR_ID	NUMBER	22	Y	De-identified
DM_LSHP_ENROL_FINAL	ENROL_STATUS_ID	NUMBER	22	N	
DM_LSHP_ENROL_FINAL	ENROL_STATUS_DESC	VARCHAR2	26	N	
DM_LSHP_ENROL_FINAL	ENROL_TYPE_ID	NUMBER	22	N	
DM_LSHP_ENROL_FINAL	ENROL_TYPE_DESC	VARCHAR2	50	N	
DM_LSHP_ENROL_FINAL	ENROL_DATE	DATE	7	Y	
DM_LSHP_ENROL_FINAL	ACHIEVE_DATE	DATE	7	Y	
DM_LSHP_ENROL_FINAL	DERIVED_START_DATE	DATE	7	Y	
DM_LSHP_ENROL_FINAL	QUALIFICATION_ID	NUMBER	22	Y	De-identified
DM_LSHP_ENROL_FINAL	CREDITS	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	QUAL_START_DATE	DATE	7	Y	
DM_LSHP_ENROL_FINAL	QUAL_END_DATE	DATE	7	Y	
DM_LSHP_ENROL_FINAL	PROBLEM_ID	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	START_DATE_ID	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	START_DATE_DESC	VARCHAR2	26	Y	
DM_LSHP_ENROL_FINAL	START_DATE	DATE	7	Y	
DM_LSHP_ENROL_FINAL	END_DATE	DATE	7	Y	
DM_LSHP_ENROL_FINAL	LSHP_ETQE_ID	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	NQF_LEVEL_ID	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	NQF_LEVEL_DESC	VARCHAR2	26	Y	
DM_LSHP_ENROL_FINAL	ETQE_START_DATE	DATE	7	Y	
DM_LSHP_ENROL_FINAL	ETQE_END_DATE	DATE	7	Y	
DM_LSHP_ENROL_FINAL	PROV_START_DATE	DATE	7	Y	
DM_LSHP_ENROL_FINAL	PROV_END_DATE	DATE	7	Y	
DM_LSHP_ENROL_FINAL	PROV_ETQE_ID	NUMBER	22	Y	De-identified
DM_LSHP_ENROL_FINAL	PROVIDER_TYPE_ID	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	PROVIDER_TYPE_DESC	VARCHAR2	26	Y	
DM_LSHP_ENROL_FINAL	PROVIDER_CLASS_ID	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	PROVIDER_CLASS_DESC	VARCHAR2	50	Y	
DM_LSHP_ENROL_FINAL	PROV_PROVINCE_CODE	VARCHAR2	10	Y	
DM_LSHP_ENROL_FINAL	PROV_PROVINCE_DESC	VARCHAR2	60	Y	
DM_LSHP_ENROL_FINAL	ASOR_START_DATE	DATE	7	Y	
DM_LSHP_ENROL_FINAL	ASOR_END_DATE	DATE	7	Y	

Table Name	Column Name	Data Type	Data Length	Allow NULLs	Comment
DM_LSHP_ENROL_FINAL	QENROL_LEARNERSHIP_ID	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	QENROL_QUALIFICATION_ID	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	QENROL_ENROL_STATUS_ID	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	QENROL_ENROL_STATUS_DE SC	VARCHAR2	32	Y	
DM_LSHP_ENROL_FINAL	QENROL_ENROL_DATE	DATE	7	Y	
DM_LSHP_ENROL_FINAL	QENROL_ACHIEVE_DATE	DATE	7	Y	
DM_LSHP_ENROL_FINAL	ETQE_FIRST_DATE	DATE	7	Y	
DM_LSHP_ENROL_FINAL	PROV_ETQE_FIRST_DATE	DATE	7	Y	
DM_LSHP_ENROL_FINAL	CYCLE_DATE	DATE	7	Y	
DM_LSHP_ENROL_FINAL	START_ETQE_IND	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	END_ETQE_IND	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	ETQE_IND_ID	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	ETQE_IND_DESC	VARCHAR2	134	Y	
DM_LSHP_ENROL_FINAL	OTHR_ETQE_ID	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	OTHR_ETQE_START_DATE	DATE	7	Y	
DM_LSHP_ENROL_FINAL	OTHR_ETQE_END_DATE	DATE	7	Y	
DM_LSHP_ENROL_FINAL	OTHR_START_ETQE_IND	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	OTHR_END_ETQE_IND	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	OTHR_ETQE_IND_ID	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	OTHR_ETQE_IND_DESC	VARCHAR2	24	Y	
DM_LSHP_ENROL_FINAL	START_DATE_IND	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	END_DATE_IND	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	START_PROV_IND	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	END_PROV_IND	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	PROV_IND_ID	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	PROV_IND_DESC	VARCHAR2	55	Y	
DM_LSHP_ENROL_FINAL	END_ASOR_IND	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	ASOR_IND_ID	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	ASOR_IND_DESC	VARCHAR2	64	Y	
DM_LSHP_ENROL_FINAL	END_QENROL_IND	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	QENROL_IND_ID	NUMBER	22	Y	
DM_LSHP_ENROL_FINAL	QENROL_IND_DESC	VARCHAR2	45	Y	

## Appendix E

### ***E.1 Introduction***

This section details the initial selection, pre-processing and derivation of the qualification enrolment records, received from the NLRD in the table DM\_QUAL\_ENROL, into a format that is suitable for data mining.

The specific semantic business rules that are applicable to qualification enrolment records are identified in Appendix E.2. The analysis and data mining of these semantic business rules requires the implementation of eight (8) semantic business rule indicators. Appendix E.3 describes the selection, pre-processing and derivation steps required for the implementation of these semantic business rule indicators. Appendix E.3.1 and Appendix E.3.2 describe the type of logic developed for the selection and pre-processing of the data. Whereas Appendix E.3.4 to Appendix E.3.11 describe the type of logic used for the derivation of the data.

The selection, pre-processing and derivation logic resulted in the implementation of a final version of the qualification enrolment data as a new table called DM\_QUAL\_ENROL\_FINAL, described in Appendix E.3.13.

### ***E.2 Applicable semantic business rules and their indicator fields***

A review of the final version of the semantic business rules (see Section 3.6.2) shows that the business rules that are applicable to qualification enrolment records are as follows:

1. that the ETQE that submitted the record
  - a. was accredited for the duration of the learner's active enrolment on the learnership/qualification/unit standard
  - b. was accredited to quality assure the qualification/unit standard for the duration of the learner's active enrolment on the qualification/unit standard
2. that the provider
  - a. was accredited for the duration of the learner's active enrolment on the learnership/qualification/unit standard
  - b. was accredited to offer the qualification/unit standard for the duration of the learner's active enrolment on the learnership/qualification/unit standard
3. that if the learner has completed the learnership or achieved the qualification/unit standard and the details of the assessor are supplied, that the assessor

- a. was registered at the time of the completion of the learnership or achievement of the qualification/unit standard
- b. was registered to assess the qualification/unit standard at the time of the completion of the learnership or achievement of the qualification/unit standard
- 4. that the qualification/unit standard was registered for the duration of the learner's active enrolment on the qualification/unit standard
- ...
- 6. that if the learner has achieved the qualification, and the qualification is a unit standards based qualification
  - c. the learner would have achieved the minimum required number of credits for the qualification, on or before the achievement of the qualification, based on the achievement of unit standards related to the qualification, and
  - d. the learner would have achieved the correct range of credits for the qualification, on or before the achievement of the qualification, based on the achievement of unit standards that have been defined as core, fundamental and elective unit standards for the qualification

The main purpose of the derivation of the qualification enrolment data for analysis and data mining therefore focused on the development of eight (8) semantic business rule indicators (each consisting of a data code and a description) that described the compliance of the record in accordance with these rules:

1. ETQE\_IND

Denotes whether the ETQE was accredited for the duration of the learner's active enrolment on the qualification.

2. ETQE\_ACCRED\_IND

Denotes whether the ETQE was accredited to quality assure the qualification for the duration of the learner's active enrolment on the qualification.

3. PROV\_IND

Denotes whether the provider was accredited for the duration of the learner's active enrolment on the qualification.

4. PROV\_ACCRED\_IND

Denotes whether the provider was accredited to offer the qualification for the duration of the learner's active enrolment on the qualification.

5. ASOR\_IND

Denotes whether the assessor was registered at the time of the achievement of the qualification.

6. ASOR\_REGSTR\_IND

Denotes whether the assessor was registered to assess the qualification at the time of the achievement of the qualification.

7. QUAL\_REGSTR\_IND

Denotes whether the qualification was registered for the duration of the learner's active enrolment on the qualification.

8. UNIT\_STD\_MIX\_IND

Denotes whether the learner achieved:

- a. the minimum required number of credits for the qualification, on or before the achievement of the qualification, based on the achievement of unit standards related to the qualification, and
- b. the correct range of credits for the qualification, achieved on or before the achievement of the qualification, based on the achievement of unit standards that have been defined as core, fundamental and elective unit standards for the qualification.

### ***E.3 Semantic business rule indicator development steps***

#### ***E.3.1 Pre derivation data collection***

By their very definition on the NQF, qualifications have start and end dates, denoted by the qualification's registration start and end dates. However, the end date of a qualification does not indicate the last date on which a learner may enrol on or achieve a qualification. Rather, qualifications have transition and train-out time periods that allow for the graceful ending of a qualification in order to allow all stakeholders and learners sufficient time to transition to a qualification's replacement qualification (Section 3.8.2.11).

Consequently, the first step of the development of the semantic business rule indicators for the qualification enrolment records focused on developing data fields that defined the last date on which a learner may have enrolled on a qualification and the last day on which a learner may have achieved a qualification. In order to achieve this a new table was created called DM\_QUAL\_FINAL which contains the same data fields as the table DM\_QUAL with the addition of two new derived data fields namely:

1. **MAX\_START\_DATE**: A calculated value indicating the last date on which a learner may enrol on a qualification. The value is calculated as the transition period added to the end date of the qualification.

$$\text{MAX\_START\_DATE} = \text{TRANSITION\_PERIOD} + \text{END\_DATE}$$

2. **MAX\_END\_DATE**: A calculated value indicating the last date on which a learner may achieve a qualification. The value is calculated as two years plus the train out period of the qualification added to the last date on which the learner may enrol on the qualification.

$$\text{MAX\_END\_DATE} = \text{MAX\_START\_DATE} + 2 + \text{TRAIN\_OUT\_PERIOD}$$

The resulting table **DM\_QUAL\_FINAL** was used instead of the table **DM\_QUAL** to source data values that described the qualification that the learner had enrolled on during the development of the semantic business rule indicators.

Determining compliance of a data record in regard to all of the semantic business rules that are applicable to qualification enrolment records required that each qualification enrolment record have an active enrolment time period. An active enrolment time period needed to be derived for qualification enrolment records that did not have an enrolment date and/or an achievement date (Section 3.6.4.1). Deriving the active enrolment time period for these types of enrolment records was accomplished utilizing the calculation of credits to notional hours (Appendix A.2) using the credits of the qualification.

As a result the next step of the development of the semantic business rule indicators focused on obtaining additional data, including the credits, for the qualification of the qualification enrolment record.

The linking of the qualification enrolment record to its qualification record in the table **DM\_QUAL\_FINAL** resulted in the addition of the following data fields to the table **DM\_QUAL\_ENROL**:

3. **CREDITS**: the credits for the qualification utilized to derive the active enrolment time period of the qualification enrolment record if required.
4. **QUAL\_START\_DATE**: the active registration start date of the qualification utilized to derive the active enrolment time period of the qualification enrolment record if required.
5. **QUAL\_END\_DATE**: the active registration end date of the qualification utilized to derive the active enrolment time period of the qualification enrolment record if required.

6. TRANSITION\_PERIOD: the transition period for the qualification.
7. TRAIN\_OUT\_PERIOD: the train out period for the qualification.
8. QUAL\_MAX\_START\_DATE: the last date on which a learner may enrol on the qualification.
9. QUAL\_MAX\_END\_DATE: the last date on which a learner may achieve the qualification.
10. NQF\_LEVEL\_ID and NQF\_LEVEL\_DESC: The data code and corresponding description of the NQF Level of the qualification.
11. QUALIFICATION\_TYPE\_ID and QUALIFICATION\_TYPE\_DESC: The data code and corresponding description of the Qualification Type of the qualification.
12. QUALIFICATION\_CLASS\_ID and QUALIFICATION\_CLASS\_DESC: The data code and corresponding description of the Qualification Class of the qualification.
13. FIELD\_ID and FIELD\_DESC: The data code and corresponding description of the Field of the qualification.
14. SUBFIELD\_ID and SUBFIELD\_DESC: The data code and corresponding description of the Subfield of the qualification.

Learners that complete their qualification via distance learning are given more time in which to complete their qualification. In this type of scenario, the train out period for the qualification must be multiplied by 1.5. As a result, the MAX\_END\_DATE value for records where ENROL\_TYPE\_DESC = 'Distance Learning' needed to be recalculated as follows:

$$\text{MAX\_END\_DATE} = \text{MAX\_START\_DATE} + 2 + (\text{TRAIN\_OUT\_PERIOD} * 1.5)$$

Any record that could not be linked to a qualification record in the table DM\_QUAL\_FINAL could not be further processed for the development of the semantic business rule indicators and as a result needed to be excluded from the research. Any records that could not be linked to a qualification record on the table DM\_QUAL\_FINAL were allocated a PROBLEM\_ID of 1 (no such records were found) and excluded from the further processing of the data.

Further, any record that had a 0 or NULL credit value, and was missing a value for the enrolment date or achievement date also needed to be excluded from the research because an active enrolment time period could not be derived for these records. As a result any

records that had a CREDITS value of 0 or NULL, and had a NULL value for the ENROL\_DATE or ACHIEVE\_DATE fields were allocated a PROBLEM\_ID of 2 and excluded from the further processing of the data (no such records were found).

### *E.3.2 Deriving the active enrolment time period*

Having obtained all the information in regard to the qualification, the derivation logic focused on deriving the active enrolment period for the qualification enrolment record.

Two new indicators were created namely; a nominal data value and a corresponding descriptive data value used to record whether the start date of the qualification enrolment record represented;

- the enrolment date as provided in the qualification enrolment record,
- the qualification enrolment record did not have an enrolment date and was as a result derived from the combination of the qualification achievement date and the qualification credits (see Section 3.6.4.1.a), or
- that the qualification enrolment record did not have an enrolment date or an achievement date and was as a result derived from the derived start date of the enrolment (see Section 3.6.4.1.a).

Additionally a new data field was created to store the derived start date of the qualification enrolment record based on the above.

Once a start date was implemented as described above, an end date for the active enrolment time period was implemented either as:

- the actual achievement date of the qualification enrolment record, or
- a derived end date calculated using the combination of the start date for the enrolment record and the qualification credits (see Section 3.6.4.1.b).

This resulted in the addition of the following indicators and data fields on the table DM\_QUAL\_ENROL:

15. START\_DATE\_ID: A nominal data code, and

16. START\_DATE\_DESC: a corresponding descriptive data value indicating whether the value in START\_DATE represents

- an enrolment date (ENROL\_DATE),

- a derived value utilizing the achievement date (ACHIEVE\_DATE) for the enrolment record and the qualification credits (CREDITS), or
- a derived value utilizing the derived start date (DERIVED\_START\_DATE) of the record.

17. START\_DATE: The start date of the active enrolment time period.

18. END\_DATE: The derived end date of the active enrolment time period, representing either the value found in ACHIEVE\_DATE or a value derived from START\_DATE and CREDITS.

A number of the semantic business rules are dependent on the active enrolment period of the record. Additionally an analysis of the qualification data (DM\_QUAL) shows that the earliest registration for a qualification occurred on 30 June 2000. As a direct result it could be deduced that any qualification enrolment record with a start date less than 30 June 2000 was an outlier record. Such records by their very nature were considered erroneous and needed to be excluded from the research. These records were allocated a PROBLEM\_ID code of 3 and excluded from the further processing of the data. These records constituted 0.53% of the qualification enrolment records initially extracted from the NLRD.

### ***E.3.3 Core data required for the development of the indicator fields and additional data values***

Having established the active enrolment time period of the qualification enrolment record, the derivation process focused on the:

- collection of the core data required for the development of the semantic business rule indicators described in Appendix E.2, and
- the collection of additional data fields that may prove valuable to analysis of the qualification enrolment data as described in Section 3.6.5.

The reader should note that the data received from the NLRD was, with the exception of lookup values, provided in a format that closely represents a relational database design. As an example, even though the qualification enrolment table (DM\_QUAL\_ENROL) contained a unique identifier for the qualification (QUALIFICATION\_ID), the qualification enrolment table did not contain additional data fields that describe the qualification, for example the NQF Level of the qualification.

This section describes which data fields sourced from other tables were added to the qualification enrolment table (DM\_QUAL\_ENROL) and how the linkage between the qualification enrolment record and the other tables were implemented.

Data that describes the accreditation of the ETQE was obtained from the ETQE accreditation table (DM\_ETQE) using the unique identifier of the ETQE (ETQE\_ID).

19. ETQE\_START\_DATE (START\_DATE on DM\_ETQE): The start date of the accreditation of the ETQE that submitted the enrolment record to the NLRD.
20. ETQE\_END\_DATE (END\_DATE on DM\_ETQE): The end date of the accreditation of the ETQE that submitted the enrolment record to the NLRD.

Data that describes the accreditation of the ETQE to quality assure the qualification was obtained from the ETQE qualification accreditation table (DM\_ETQE\_ACCRED) using the unique identifier of the ETQE (ETQE\_ID) and the unique identifier of the qualification (QUALIFICATION\_ID).

21. ETQE\_ACCRED\_START\_DATE (START\_DATE on DM\_ETQE\_ACCRED): The start date of the accreditation to quality assure the qualification of the ETQE that submitted the enrolment record to the NLRD.
22. ETQE\_ACCRED\_END\_DATE (END\_DATE on DM\_ETQE\_ACCRED): The end date of the accreditation to quality assure the qualification of the ETQE that submitted the enrolment record to the NLRD.

Data that describes the provider and its accreditation as obtained from the provider accreditation table (DM\_PROV) using the unique identifier of the provider (PROVIDER\_ID).

23. PROV\_START\_DATE (START\_DATE on DM\_PROV): The start date of the accreditation of the provider that offered the qualification.
24. PROV\_END\_DATE (END\_DATE on DM\_PROV): The end date of the accreditation of the provider that offered the qualification.
25. PROV\_ETQE\_ID (ETQE\_ID on DM\_PROV): The primary ETQE of the provider.
26. PROVIDER\_TYPE\_ID and PROVIDER\_TYPE\_DESC: The data code and corresponding description of the provider type.
27. PROVIDER\_CLASS\_ID and PROVIDER\_CLASS\_DESC: The data code and corresponding description of the provider class.

28. PROV\_PROVINCE\_CODE (PROVINCE\_CODE on DM\_PROV) and PROV\_PROVINCE\_DESC (PROVINCE\_DESC on DM\_PROV): The data code and corresponding description of the province that the provider is located in.

The raw data obtained from the NLRD in regard to provider accreditations contains two scenarios in which a provider may have been accredited for the same qualification. The provider may have been accredited for the qualification explicitly whereby the provider accreditation record only contains a QUALIFICATION\_ID value. The provider may also have been accredited for the same qualification implicitly whereby the provider accreditation record contains a LEARNERSHIP\_ID and a QUALIFICATION\_ID. As a result a new table called DM\_PROV\_ACCRED\_QUAL was created that contains a unique combination of PROVIDER\_ID and QUALIFICATION\_ID, the minimum START\_DATE for the accreditation for the qualification and the maximum END\_DATE for the accreditation for the qualification.

Data that describes the accreditation of the provider to offer qualifications as obtained from the provider qualification accreditation table (DM\_PROV\_ACCRED\_QUAL) using the unique identifier of the provider (PROVIDER\_ID) and the unique identifier of the qualification (QUALIFICATION\_ID).

29. PROV\_ACCRED\_START\_DATE (START\_DATE on DM\_PROV\_ACCRED\_QUAL): The start date of the accreditation to offer the qualification of the provider that offered the qualification.

30. PROV\_ACCRED\_END\_DATE (END\_DATE on DM\_PROV\_ACCRED\_QUAL): The end date of the accreditation to offer the qualification of the provider that offered the qualification.

Data that describes the registration of the assessor as obtained from the assessor registration table (DM\_ASOR) using the unique identifier of the assessor (ASSESSOR\_ID).

31. ASOR\_START\_DATE (START\_DATE on DM\_ASOR): The start date of the registration of the assessor that assessed the qualification achievement.

32. ASOR\_END\_DATE (END\_DATE on DM\_ASOR): The end date of the registration of the assessor that assessed the qualification achievement.

The raw data obtained from the NLRD in regard to assessor registrations contains two scenarios in which an assessor may have been registered for the same qualification. The assessor may have been registered for the qualification explicitly whereby the assessor registration record only contains a QUALIFICATION\_ID value. The assessor may also have been registered for the same qualification implicitly whereby the assessor registration record contains a LEARNERSHIP\_ID and a QUALIFICATION\_ID. As a result a new table called DM\_ASOR\_REGSTR\_QUAL was created that contains a unique combination of ASSESSOR\_ID and QUALIFICATION\_ID, the minimum START\_DATE for the assessor's registration for the qualification and the maximum END\_DATE for the assessor's registration for the qualification.

Data that describes the registration of the assessor to assess a qualification as obtained from the assessor qualification registration table (DM\_ASOR\_REGSTR\_QUAL) using the unique identifier of the assessor (ASSESSOR\_ID) and the unique identifier of the qualification (QUALIFICATION\_ID).

33. ASOR\_REGSTR\_START\_DATE (START\_DATE on DM\_ASOR\_REGSTR\_QUAL):

The start date of the registration to assess the qualification of the assessor that assessed the qualification achievement.

34. ASOR\_REGSTR\_END\_DATE (END\_DATE on DM\_ASOR\_REGSTR\_QUAL):

The end date of the registration to assess the qualification of the assessor that assessed the qualification achievement.

The required number of core, fundamental and elective unit standards for unit standards based qualifications as obtained from the table DM\_USTD\_QUAL. This required the implementation of an interim table called DM\_USTD\_QUAL\_CREDITS which contains, per qualification, four data fields that represent:

35. the total credits required for the qualification (the field CREDITS from the table DM\_QUAL) as USTD\_CREDITS\_TOTAL,

36. the total core credits required for the qualification (sum of CREDITS from DM\_USTD\_QUAL where the unit standard type (USTD\_QUAL\_TYPE\_CODE) is 'core') as USTD\_CREDITS\_CORE,

37. the total fundamental credits required for the qualification (sum of CREDITS from DM\_USTD\_QUAL where the unit standard type (USTD\_QUAL\_TYPE\_CODE) is 'fundamental') as USTD\_CREDITS\_FUND,

38. and the total elective credits required for the qualification (sum of CREDITS from DM\_USTD\_QUAL where the unit standard type (USTD\_QUAL\_TYPE\_CODE) is 'elective') as USTD\_CREDITS\_ELEC.

Data that describes the unit standard requirements for unit standard based qualifications from qualification unit standard table (DM\_USTD\_QUAL\_CREDITS) using the unique identifier of the qualification (QUALIFICATION\_ID).

39. USTD\_CREDITS\_TOTAL (USTD\_CREDITS\_TOTAL on DM\_USTD\_QUAL\_CREDITS): The required number of credits required for the achievement of the unit standards based qualification.

40. USTD\_CREDITS\_CORE (USTD\_CREDITS\_CORE on DM\_USTD\_QUAL\_CREDITS): The required number of core credits required for the achievement of the unit standards based qualification.

41. USTD\_CREDITS\_FUND (USTD\_CREDITS\_FUND on DM\_USTD\_QUAL\_CREDITS): The required number of fundamental credits required for the achievement of the unit standards based qualification.

42. USTD\_CREDITS\_ELEC (USTD\_CREDITS\_ELEC on DM\_USTD\_QUAL\_CREDITS): The defined number of elective credits that can be achieved for the achievement of the unit standards based qualification.

A check was completed to ensure that the credits recorded against each unit standards based qualification was less than or equal to the credits available to the qualification based on the unit standards that are linked to the qualification. Any enrolment records that belong to qualifications where the total credits required for the qualification was more than the sum of the credits for the unit standards linked to the qualification were allocated a PROBLEM\_ID of 4 and excluded from the further processing of the data. These records constituted 0.33% of the qualification enrolment records initially extracted from the NLRD.

The date on which an ETQE submitted its first full data submission to the NLRD as obtained from the table DM\_ETQE\_START using the ETQE\_ID of the enrolment record (see Section 3.8.3.1).

43. ETQE\_FIRST\_DATE (START\_DATE on DM\_ETQE\_START): The first date on which the ETQE submitted a full submission to the NLRD.

The date of the most recent NLRD data submission cycle as obtained from the Director of the NLRD (see Section 3.8.3.2).

44. CYCLE\_DATE (variable that is set at execution of the script): The date of the most recent NLRD data submission cycle.

### E.3.4 Development of ETQE\_IND

As detailed in Appendix E.2, the development of ETQE\_IND required the implementation of an indicator that denotes whether the ETQE was accredited for the duration of the learner’s active enrolment on the qualification. In order to make the analysis of the data more suitable for data mining purposes the indicator was implemented as a data code and corresponding description.

Figure E.3.4.1 illustrates the manner in which ETQE\_IND was developed using five example qualification enrolment records, for an ETQE that has not been amalgamated. The figure shows how:

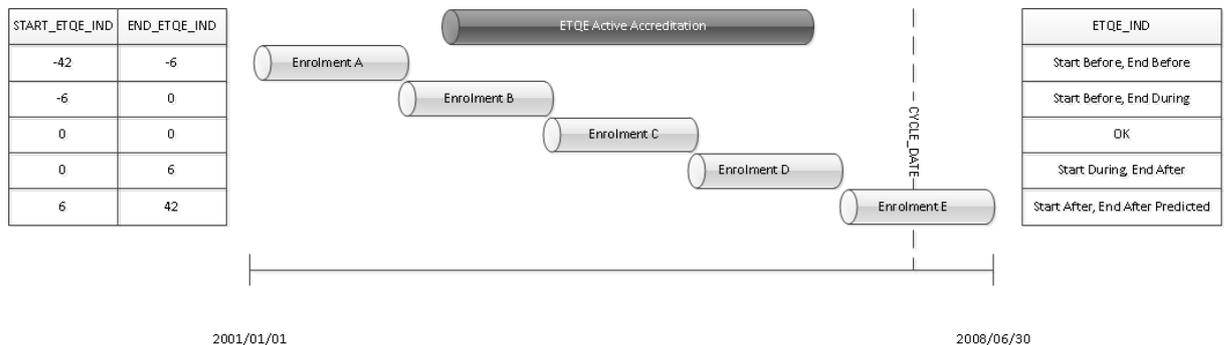


Figure E.3.4.1 Illustrative diagram of ETQE\_IND development

- a qualification enrolment record (Enrolment A) with a start date prior to the accreditation period of the ETQE and an end date prior to the accreditation period of the ETQE is allocated an ETQE\_IND value of ‘Start Before, End Before’,
- a qualification enrolment record (Enrolment B) with a start date prior to the accreditation period of the ETQE and an end date during the accreditation period of the ETQE is allocated an ETQE\_IND value of ‘Start Before, End During’,
- a qualification enrolment record (Enrolment C) with a start date during the accreditation period of the ETQE and an end date during the accreditation period of the ETQE is allocated an ETQE\_IND value of ‘OK’,

- a qualification enrolment record (Enrolment D) with a start date during the accreditation period of the ETQE and an end date after the accreditation period of the ETQE is allocated an ETQE\_IND value of ‘Start During, End After’, and
- a qualification enrolment record (Enrolment E) with a start date after the accreditation period of the ETQE and an end date after the accreditation period of the ETQE is allocated an ETQE\_IND value of ‘Start After, End After’, and because the end of the active enrolment time period exceeds the latest data submission cycle date the word ‘Predicted’ is appended to the value.

The development of the ETQE\_IND indicator required the implementation of two additional indicators. These indicators assisted in the development of and further description of the value allocated to ETQE\_IND. Both of these two additional indicators were developed as representations of data in relation to a point in time as discussed in Section 3.6.4.4.

The first indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the start date (START\_DATE) of the active enrolment time period and the start date of the ETQE’s active accreditation time period (ETQE\_START\_DATE and ETQE\_END\_DATE), where;

- a qualification enrolment record with a start date before the start date of the ETQE’s accreditation would be given a negative value of the number of months between these two values (for example Enrolment A on Figure E.3.4.1),
- a qualification enrolment record with a start date that falls between the start and end dates of the ETQE’s accreditation is given a value of 0 (for example Enrolment C on Figure E.3.4.1), and
- a qualification enrolment record with a start date that is after the end date of the ETQE’s accreditation would be given a positive value of the number of months between these two values (for example Enrolment E on Figure E.3.4.1).

The second indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the end date (END\_DATE) of the active enrolment time period and the end date of the ETQE’s active accreditation time period (ETQE\_START\_DATE and ETQE\_END\_DATE), where;

- a qualification enrolment record with an end date before the start date of the ETQE's accreditation would be given a negative value of the number of months between these two values (for example Enrolment A on Figure E.3.4.1),
- a qualification enrolment record with an end date that falls between the start and end dates of the ETQE's accreditation is given a value of 0 (for example Enrolment C on Figure E.3.4.1), and
- a qualification enrolment record with an end date that is after the end date of the ETQE's accreditation would be given a positive value of the number of months between these two values (for example Enrolment E on Figure E.3.4.1).

This logic resulted in the addition of the following new indicators on the table DM\_QUAL\_ENROL:

45. START\_ETQE\_IND: Numeric value indicating the distance between the start date of the qualification enrolment record and the ETQE accreditation.
46. END\_ETQE\_IND: Numeric value indicating the distance between the end date of the qualification enrolment record and the ETQE accreditation.

Using the values in these two fields in conjunction with each other it was possible to derive a code and corresponding description for ETQE\_IND by:

- Allocating a value of 'OK' to records where START\_ETQE\_IND is equal to 0 and END\_ETQE\_IND is equal to 0.

For all remaining records

- Allocating a value of 'Start Before' to records where START\_ETQE\_IND was less than 0, 'Start During' to records where START\_ETQE\_IND is equal to 0 and 'Start After' where START\_ETQE\_IND was greater than 0.
- Allocating a value 'End Before' to records where END\_ETQE\_IND was less than 0, 'End During' to records where END\_ETQE\_IND is equal to 0 and 'End After' where END\_ETQE\_IND was greater than 0.

This logic resulted in the addition of the ETQE\_IND indicator code and corresponding description on the table DM\_QUAL\_ENROL.

47. ETQE\_IND\_ID and ETQE\_IND\_DESC: Code and corresponding description denoting a record's compliance in regard to whether the ETQE was accredited for the duration of the learner's active enrolment on the qualification.

The above mentioned logic however did not take into consideration the accreditation of the ETQE that was also accredited to quality assure the qualification in situations where ETQEs had been amalgamated (see Section 3.8.3.3). In order to address this issue, the logic determined whether a different ETQE had in the past been accredited to quality assure the qualification (DM\_ETQE\_ACCRED).

The ETQE identifier, start date and end date of the ETQE that was also accredited to quality assure the qualification was amended to the table DM\_QUAL\_ENROL:

48. OTHR\_ETQE\_ID: The ETQE identifier of the ETQE.
49. OTHR\_ETQE\_START\_DATE (START\_DATE on DM\_ETQE): The start date of the accreditation of the ETQE.
50. OTHR\_ETQE\_END\_DATE (END\_DATE on DM\_ETQE): The end date of the accreditation of the ETQE.

Four indicators were developed in the same manner as described for START\_ETQE\_IND, END\_ETQE\_IND, ETQE\_IND\_ID and ETQE\_IND\_DESC, using the indicators OTHR\_ETQE\_START\_DATE, OTHR\_ETQE\_END\_DATE, OTHR\_START\_ETQE\_IND and OTHR\_END\_ETQE\_IND in place of the indicators ETQE\_START\_DATE, ETQE\_END\_DATE, START\_ETQE\_IND and END\_ETQE\_IND.

This logic resulted in the addition of the following new indicators on the table DM\_QUAL\_ENROL:

51. OTHR\_START\_ETQE\_IND: Numeric value indicating the distance between the start date of the qualification enrolment record and the other ETQE's accreditation.
52. OTHR\_END\_ETQE\_IND: Numeric value indicating the distance between the end date of the qualification enrolment record and the other ETQE's accreditation.
53. OTHR\_ETQE\_IND\_ID and OTHR\_ETQE\_IND\_DESC: Code and corresponding description denoting a record's compliance in regard to whether the other ETQE was accredited for the duration of the learner's active enrolment on the qualification.

The results of both the ETQE\_IND\_ID and ETQE\_IND\_DESC fields and the OTHR\_ETQE\_IND\_ID and OTHR\_ETQE\_IND\_DESC fields were then consolidated into the ETQE\_IND indicators in the following manner:

- Any record that was found to be compliant based on the value stored in ETQE\_IND\_ID or OTHR\_ETQE\_IND\_ID was marked as compliant.
- Any record that was found to be non-compliant based on both the value stored in ETQE\_IND\_ID and OTHR\_ETQE\_IND\_ID was provided a modified code and corresponding description that show the results of the compliance result of both ETQE\_IND\_ID and OTHR\_ETQE\_IND\_ID.

The final derivation step entailed amending the ETQE\_IND data code and corresponding description to differentiate records with a calculated end date that is greater than the latest data submission cycle date from other records (see Section 3.8.3.2). As a result the data code was amended and the word ‘Predicted’ was appended to the ETQE\_IND indicator description for any records with an END\_DATE value greater than CYCLE\_DATE (for example Enrolment E on Figure E.3.4.1).

### ***E.3.5 Development of ETQE\_ACCRED\_IND***

As detailed in Appendix E.2, the development of ETQE\_ACCRED\_IND required the implementation of an indicator that denotes whether the ETQE was accredited to quality assure the qualification for the duration of the learner’s active enrolment on the qualification. In order to make the analysis of the data more suitable for data mining purposes the indicator was implemented as a data code and corresponding description.

Figure E.3.5.1 illustrates the manner in which ETQE\_ACCRED\_IND was developed using five example qualification enrolment records, for an ETQE that has not been amalgamated. The figure shows how:

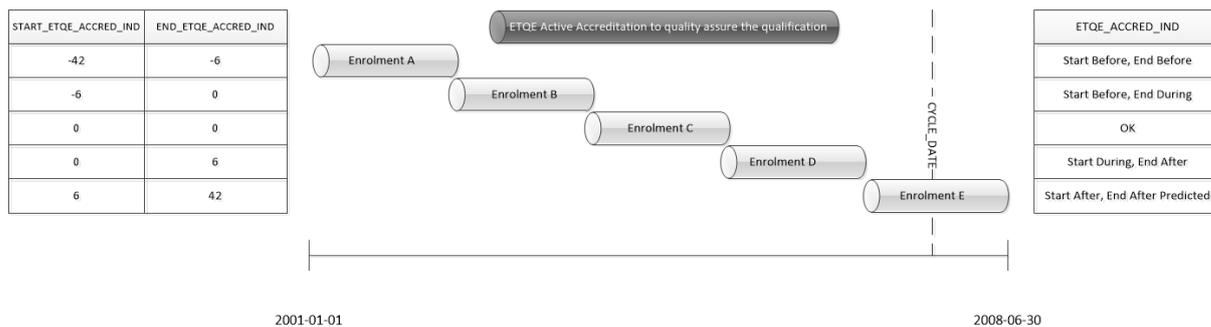


Figure E.3.5.1 Illustrative diagram of ETQE\_ACCRED\_IND development

- a qualification enrolment record (Enrolment A) with a start date prior to the qualification accreditation period of the ETQE and an end date prior to the qualification accreditation period of the ETQE is allocated an ETQE\_ACCRED\_IND value of ‘Start Before, End Before’,
- a qualification enrolment record (Enrolment B) with a start date prior to the qualification accreditation period of the ETQE and an end date during the qualification accreditation period of the ETQE is allocated an ETQE\_ACCRED\_IND value of ‘Start Before, End During’,
- a qualification enrolment record (Enrolment C) with a start date during the qualification accreditation period of the ETQE and an end date during the qualification accreditation period of the ETQE is allocated an ETQE\_ACCRED\_IND value of ‘OK’,
- a qualification enrolment record (Enrolment D) with a start date during the qualification accreditation period of the ETQE and an end date after the qualification accreditation period of the ETQE is allocated an ETQE\_ACCRED\_IND value of ‘Start During, End After’, and
- a qualification enrolment record (Enrolment E) with a start date after the qualification accreditation period of the ETQE and an end date after the qualification accreditation period of the ETQE is allocated an ETQE\_ACCRED\_IND value of ‘Start After, End After’, and because the end of the active enrolment time period exceeds the latest data submission cycle date the word ‘Predicted’ is appended to the value.

The development of the ETQE\_ACCRED\_IND indicator required the implementation of two additional indicators. These indicators assisted in the development of and further description of the value allocated to ETQE\_ACCRED\_IND. Both of these two additional

indicators were developed as representations of data in relation to a point in time as discussed in Section 3.6.4.4.

The first indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the start date (START\_DATE) of the active enrolment time period and the start date of the ETQE's active accreditation to quality assure the qualification time period (ETQE\_ACCRED\_START\_DATE and ETQE\_ACCRED\_END\_DATE), where;

- a qualification enrolment record with a start date before the start date of the ETQE's accreditation to quality assure the qualification would be given a negative value of the number of months between these two values (for example Enrolment A on Figure E.3.5.1),
- a qualification enrolment record with a start date that falls between the start and end dates of the ETQE's accreditation to quality assure the qualification is given a value of 0 (for example Enrolment C on Figure E.3.5.1), and
- a qualification enrolment record with a start date that is after the end date of the ETQE's accreditation to quality assure the qualification would be given a positive value of the number of months between these two values (for example Enrolment E on Figure E.3.5.1).

The second indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the end date (END\_DATE) of the active enrolment time period and the end date of the ETQE's active accreditation to quality assure the qualification time period (ETQE\_ACCRED\_START\_DATE and ETQE\_ACCRED\_END\_DATE), where;

- a qualification enrolment record with an end date before the start date of the ETQE's accreditation to quality assure the qualification would be given a negative value of the number of months between these two values (for example Enrolment A on Figure E.3.5.1),
- a qualification enrolment record with an end date that falls between the start and end dates of the ETQE's accreditation to quality assure the qualification is given a value of 0 (for example Enrolment C on Figure E.3.5.1), and
- a qualification enrolment record with an end date that is after the end date of the ETQE's accreditation to quality assure the qualification would be given a positive

value of the number of months between these two values (for example Enrolment E on Figure E.3.5.1).

This logic resulted in the addition of the following new indicators on the table DM\_QUAL\_ENROL:

- 54. START\_ETQE\_ACCRED\_IND: Numeric value indicating the distance between the start date of the qualification enrolment record and the ETQE accreditation to quality assure the qualification.
- 55. END\_ETQE\_ACCRED\_IND: Numeric value indicating the distance between the end date of the qualification enrolment record and the ETQE accreditation to quality assure the qualification.

Using the values in these two fields in conjunction with each other it was possible to derive a code and corresponding description for ETQE\_ACCRED\_IND by:

1. Allocating a value of 'OK' to records where START\_ETQE\_ACCRED\_IND is equal to 0 and END\_ETQE\_ACCRED\_IND is equal to 0.

For all remaining records

2. Allocating a value of 'Start Before' to records where START\_ETQE\_ACCRED\_IND was less than 0, 'Start During' to records where START\_ETQE\_ACCRED\_IND is equal to 0 and 'Start After' where START\_ETQE\_ACCRED\_IND was greater than 0.
3. Allocating a value 'End Before' to records where END\_ETQE\_ACCRED\_IND was less than 0, 'End During' to records where END\_ETQE\_ACCRED\_IND is equal to 0 and 'End After' where END\_ETQE\_ACCRED\_IND was greater than 0.

This logic resulted in the addition of the ETQE\_ACCRED\_IND indicator code and corresponding description on the table DM\_QUAL\_ENROL.

- 56. ETQE\_ACCRED\_IND\_ID and ETQE\_ACCRED\_IND\_DESC: Code and corresponding description denoting a record's compliance in regard to whether the ETQE was accredited to quality assure the qualification for the duration of the learner's active enrolment on the qualification.

The above mentioned logic however did not take into consideration the accreditation to quality assure the qualification of the ETQE that was also accredited to quality assure the qualification in situations where ETQEs had been amalgamated (see Section 3.8.3.3). In

order to address this issue, the logic determined whether a different ETQE had in the past been accredited to quality assure the qualification (DM\_ETQE\_ACCRED).

The ETQE identifier, start date and end date of the accreditation to quality assure the qualification of the ETQE that was also accredited to quality assure the qualification was amended to the table DM\_QUAL\_ENROL:

57. OTHR\_ETQE\_ACCRED\_ID: The ETQE identifier of the ETQE.

58. OTHR\_ETQE\_ACCRED\_START\_DATE (START\_DATE on DM\_ETQE\_ACCRED):

The start date of the ETQE accreditation to quality assure the qualification for the ETQE.

59. OTHR\_ETQE\_ACCRED\_END\_DATE (END\_DATE on DM\_ETQE\_ACCRED): The

end date of the accreditation to quality assure the qualification for the ETQE.

Four indicators were developed in the same manner as described for START\_ETQE\_ACCRED\_IND, END\_ETQE\_ACCRED\_IND, ETQE\_ACCRED\_IND\_ID and ETQE\_ACCRED\_IND\_DESC, using the indicators OTHR\_ETQE\_ACCRED\_START\_DATE, OTHR\_ETQE\_ACCRED\_END\_DATE, OTHR\_START\_ETQE\_ACCRED\_IND and OTHR\_END\_ETQE\_ACCRED\_IND in place of the indicators ETQE\_START\_DATE, ETQE\_END\_DATE, START\_ETQE\_ACCRED\_IND and END\_ETQE\_ACCRED\_IND.

This logic resulted in the addition of the following new indicators on the table DM\_QUAL\_ENROL:

60. OTHR\_START\_ETQE\_ACCRED\_IND: Numeric value indicating the distance between the start date of the qualification enrolment record and the other ETQE's accreditation to quality assure the qualification.

61. OTHR\_END\_ETQE\_ACCRED\_IND: Numeric value indicating the distance between the end date of the qualification enrolment record and the other ETQE's accreditation to quality assure the qualification.

62. OTHR\_ETQE\_ACCRED\_IND\_ID and OTHR\_ETQE\_ACCRED\_IND\_DESC: Code and corresponding description denoting a record's compliance in regard to whether the other ETQE was accredited to quality assure the qualification for the duration of the learner's active enrolment on the qualification.

The results of both the ETQE\_ACCRED\_IND\_ID and ETQE\_ACCRED\_IND\_DESC fields and the OTHR\_ETQE\_ACCRED\_IND\_ID and OTHR\_ETQE\_ACCRED\_IND\_DESC fields were then consolidated into the ETQE\_ACCRED\_IND indicators in the following manner:

- Any record that was found to be compliant based on the value stored in ETQE\_ACCRED\_IND\_ID or OTHR\_ETQE\_ACCRED\_IND\_ID was marked as compliant.
- Any record that was found to be non-compliant based on both the value stored in ETQE\_ACCRED\_IND\_ID and OTHR\_ETQE\_ACCRED\_IND\_ID was provided a modified code and corresponding description that show the results of the compliance result of both ETQE\_ACCRED\_IND\_ID and OTHR\_ETQE\_ACCRED\_IND\_ID.

The final derivation step entailed amending the ETQE\_ACCRED\_IND data code and corresponding description to differentiate records with a calculated end date that is greater than the latest data submission cycle date from other records (see Section 3.8.3.2). As a result the data code was amended and the word ‘Predicted’ was appended to the ETQE\_ACCRED\_IND indicator description for any records with an END\_DATE value greater than CYCLE\_DATE (for example Enrolment E on Figure E.3.5.1).

### ***E.3.6 Development of PROV\_IND***

As detailed in Appendix E.2, the development of PROV\_IND required the implementation of an indicator that denotes whether the provider was accredited for the duration of the learner’s active enrolment on the qualification. In order to make the analysis of the data more suitable for data mining purposes the indicator was implemented as a data code and corresponding description.

Figure E.3.6.1 illustrates the manner in which PROV\_IND was developed using five example qualification enrolment records, for a provider that was accredited and is not an ‘ETQE Provider’. The figure shows how:

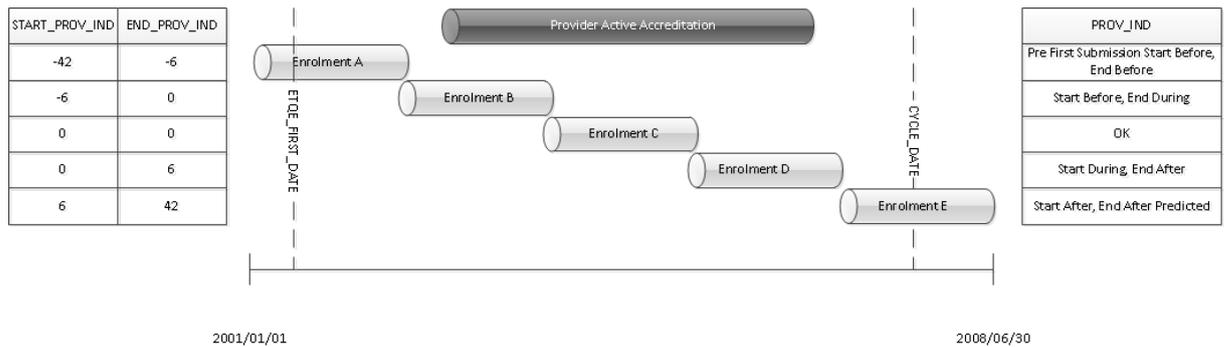


Figure E.3.6.1 Illustrative diagram of PROV\_IND development

- a qualification enrolment record (Enrolment A) with a start date prior to the provider's accreditation period and an end date prior to the provider's accreditation period is allocated a PROV\_IND value of 'Start Before, End Before'. The start of the active enrolment time period precedes the date on which the ETQE submitted its first full data submission to the NLRD, as a result the words 'Pre First Submission' are appended to the value,
- a qualification enrolment record (Enrolment B) with a start date prior to the accreditation period of the provider and an end date during the accreditation period of the provider is allocated a PROV\_IND value of 'Start Before, End During',
- a qualification enrolment record (Enrolment C) with a start date during the accreditation period of the provider and an end date during the accreditation period of the provider is allocated a PROV\_IND value of 'OK',
- a qualification enrolment record (Enrolment D) with a start date during the accreditation period of the provider and an end date after the accreditation period of the provider is allocated a PROV\_IND value of 'Start During, End After', and
- a qualification enrolment record (Enrolment E) with a start date after the accreditation period of the provider and an end date after the accreditation period of the provider is allocated a PROV\_IND value of 'Start After, End After'. The end of the active enrolment time period exceeds the latest data submission cycle date, as a result the word 'Predicted' is appended to the value.

The development of the PROV\_IND indicator required the implementation of two additional indicators. These indicators assisted in the development of and further description of the value allocated to PROV\_IND. Both of these two additional indicators

were developed as representations of data in relation to a point in time as discussed in Section 3.6.4.4.

The first indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the start date (START\_DATE) of the active enrolment time period and the start date of the provider's active accreditation time period (PROV\_START\_DATE and PROV\_END\_DATE), where;

- a qualification enrolment record with a start date before the start date of the provider's accreditation would be given a negative value of the number of months between these two values (for example Enrolment A on Figure E.3.6.1),
- a qualification enrolment record with a start date that falls between the start and end dates of the provider's accreditation is given a value of 0 (for example Enrolment C on Figure E.3.6.1), and
- a qualification enrolment record with a start date that is after the end date of the provider's accreditation would be given a positive value of the number of months between these two values (for example Enrolment E on Figure E.3.6.1).

The second indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the end date (END\_DATE) of the active enrolment time period and the end date of the provider's active accreditation time period (PROV\_START\_DATE and PROV\_END\_DATE)), where;

- a qualification enrolment record with an end date before the start date of the provider's accreditation would be given a negative value of the number of months between these two values (for example Enrolment A on Figure E.3.6.1),
- a qualification enrolment record with an end date that falls between the start and end dates of the provider's accreditation is given a value of 0 (for example Enrolment C on Figure E.3.6.1), and
- a qualification enrolment record with an end date that is after the end date of the provider's accreditation would be given a positive value of the number of months between these two values (for example Enrolment E on Figure E.3.6.1).

This logic resulted in the addition of the following new indicators on the table DM\_QUAL\_ENROL:

63. START\_PROV\_IND: Numeric value indicating the distance between the start date of the qualification enrolment record and the provider accreditation.
64. END\_PROV\_IND: Numeric value indicating the distance between the end date of the qualification enrolment record and the provider accreditation.

Using the values in these two fields in conjunction with each other it was possible to derive a code and corresponding description for PROV\_IND as follows:

- Where a provider is an 'ETQE provider' (see Section 3.8.3.5), allocating a value of 'ETQE Provider'
- Where a provider accreditation did not exist, allocating a value of 'No Accreditation' to the record.
- Allocating a value of 'OK' to records where START\_PROV\_IND is equal to 0 and END\_PROV\_IND is equal to 0.

For all remaining records:

- Allocating a value of 'Start Before' to records where START\_PROV\_IND was less than 0, 'Start During' to records where START\_PROV\_IND is equal to 0 and 'Start After' where START\_PROV\_IND was greater than 0.
- Allocating a value 'End Before' to records where END\_PROV\_IND was less than 0, 'End During' to records where END\_PROV\_IND is equal to 0 and 'End After' where END\_PROV\_IND was greater than 0.

The final derivation steps included:

- Amending the data code and appending the words 'Pre First Submission' to the indicator value for any records with a START\_DATE value less than PROV\_ETQE\_FIRST\_DATE (for example Enrolment A on Figure E.3.6.1). In other words all records where the learner enrolled on the qualification prior to the first full data submission from the primary ETQE of the provider to the NLRD (see Section 3.8.3.1 and Section 3.8.3.5).
- Amending the data code and appending the word 'Predicted' to the indicator value for any records with a END\_DATE value greater than CYCLE\_DATE (for example Enrolment E on Figure E.3.6.1). In other words all records with a calculated end date that is greater than the latest data submission cycle date (see Section 3.8.3.2).

This logic resulted in the addition of the PROV\_IND indicator code and corresponding description on the table DM\_QUAL\_ENROL.

65. PROV\_IND\_ID and PROV\_IND\_DESC: Code and corresponding description denoting a record’s compliance in regard to whether the provider was accredited for the duration of the learner’s active enrolment on the qualification.

### E.3.7 Development of PROV\_ACCRED\_IND

As detailed in Appendix E.2, the development of PROV\_ACCRED\_IND required the implementation of an indicator that denotes whether the provider was accredited to offer the qualification for the duration of the learner’s active enrolment on the qualification. In order to make the analysis of the data more suitable for data mining purposes the indicator was implemented as a data code and corresponding description.

Figure E.3.7.1 illustrates the manner in which PROV\_ACCRED\_IND was developed using five example qualification enrolment records, for a provider that was accredited to offer the qualification and is not an ‘ETQE Provider’. The figure shows how:

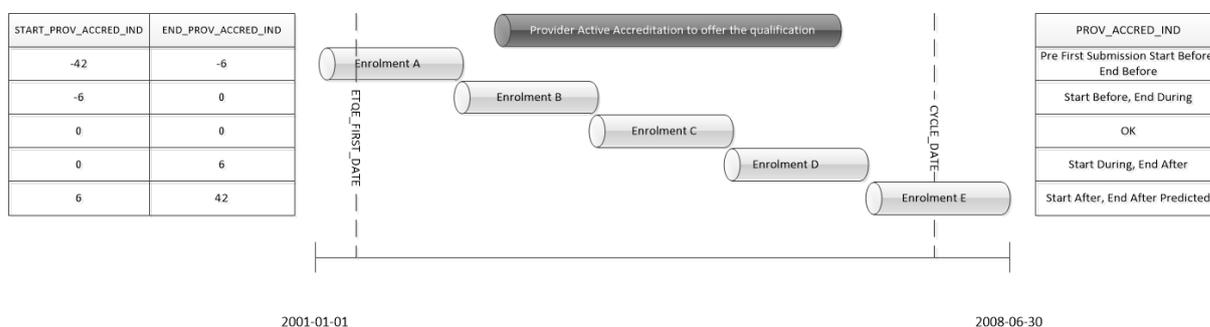


Figure E.3.7.1 Illustrative diagram of PROV\_ACCRED\_IND development

- a qualification enrolment record (Enrolment A) with a start date prior to the qualification accreditation of the provider and an end date prior to the qualification accreditation of the provider is allocated a PROV\_ACCRED\_IND value of ‘Start Before, End Before’. The start of the active enrolment time period precedes the date on which the ETQE submitted its first full data submission to the NLRD, as a result the words ‘Pre First Submission’ are appended to the value,
- a qualification enrolment record (Enrolment B) with a start date prior to the qualification accreditation period of the provider and an end date during the

qualification accreditation period of the provider is allocated a PROV\_ACCRED\_IND value of 'Start Before, End During',

- a qualification enrolment record (Enrolment C) with a start date during the qualification accreditation period of the provider and an end date during the qualification accreditation period of the provider is allocated a PROV\_ACCRED\_IND value of 'OK',
- a qualification enrolment record (Enrolment D) with a start date during the qualification accreditation period of the provider and an end date after the qualification accreditation period of the provider is allocated a PROV\_ACCRED\_IND value of 'Start During, End After', and
- a qualification enrolment record (Enrolment E) with a start date after the qualification accreditation period of the provider and an end date after the qualification accreditation period of the provider is allocated a PROV\_ACCRED\_IND value of 'Start After, End After'. The end of the active enrolment time period exceeds the latest data submission cycle date, as a result the word 'Predicted' is appended to the value.

The development of the PROV\_ACCRED\_IND indicator required the implementation of two additional indicators. These indicators assisted in the development of and further description of the value allocated to PROV\_ACCRED\_IND. Both of these two additional indicators were developed as representations of data in relation to a point in time as discussed in Section 3.6.4.4.

The first indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the start date (START\_DATE) of the active enrolment time period and the start date of the provider's active accreditation to offer the qualification time period (PROV\_ACCRED\_START\_DATE and PROV\_ACCRED\_END\_DATE), where;

- a qualification enrolment record with a start date before the start date of the provider's accreditation to offer the qualification would be given a negative value of the number of months between these two values (for example Enrolment A on Figure E.3.7.1),
- a qualification enrolment record with a start date that falls between the start and end dates of the provider's accreditation to offer the qualification is given a value of 0 (for example Enrolment C on Figure E.3.7.1), and

- a qualification enrolment record with a start date that is after the end date of the provider's accreditation to offer the qualification would be given a positive value of the number of months between these two values (for example Enrolment E on Figure E.3.7.1).

The second indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the end date (END\_DATE) of the active enrolment time period and the end date of the provider's active accreditation to offer the qualification time period (PROV\_ACCRED\_START\_DATE and PROV\_ACCRED\_END\_DATE)), where;

- a qualification enrolment record with an end date before the start date of the provider's accreditation to offer the qualification would be given a negative value of the number of months between these two values (for example Enrolment A on Figure E.3.7.1),
- a qualification enrolment record with an end date that falls between the start and end dates of the provider's accreditation to offer the qualification is given a value of 0 (for example Enrolment C on Figure E.3.7.1), and
- a qualification enrolment record with an end date that is after the end date of the provider's accreditation to offer the qualification would be given a positive value of the number of months between these two values (for example Enrolment E on Figure E.3.7.1).

This logic resulted in the addition of the following new indicators on the table DM\_QUAL\_ENROL:

66. START\_PROV\_ACCRED\_IND: Numeric value indicating the distance between the start date of the qualification enrolment record and the provider accreditation to offer the qualification.
67. END\_PROV\_ACCRED\_IND: Numeric value indicating the distance between the end date of the qualification enrolment record and the provider accreditation to offer the qualification.

Using the values in these two fields in conjunction with each other it was possible to derive a code and corresponding description for PROV\_ACCRED\_IND as follows:

- Where a provider is an 'ETQE provider' (see Section 3.8.3.5), allocating a value of 'ETQE Provider'

- Where a provider accreditation to offer the qualification did not exist, allocating a value of 'No Accreditation' to the record.
- Allocating a value of 'OK' to records where START\_PROV\_ACCRED\_IND is equal to 0 and END\_PROV\_ACCRED\_IND is equal to 0.

For all remaining records:

- Allocating a value of 'Start Before' to records where START\_PROV\_ACCRED\_IND was less than 0, 'Start During' to records where START\_PROV\_ACCRED\_IND is equal to 0 and 'Start After' where START\_PROV\_ACCRED\_IND was greater than 0.
- Allocating a value 'End Before' to records where END\_PROV\_ACCRED\_IND was less than 0, 'End During' to records where END\_PROV\_ACCRED\_IND is equal to 0 and 'End After' where END\_PROV\_ACCRED\_IND was greater than 0.

The final derivation steps included:

- Amending the data code and appending the words 'Pre First Submission' to the indicator value for any records with a START\_DATE value less than PROV\_ACCRED\_ETQE\_FIRST\_DATE (for example Enrolment A on Figure E.3.7.1). In other words all records where the learner enrolled on the qualification prior to the first full data submission from the primary ETQE of the provider to the NLRD (see Section 3.8.3.1 and Section 3.8.3.5).
- Amending the data code and appending the word 'Predicted' to the indicator value for any records with a END\_DATE value greater than CYCLE\_DATE (for example Enrolment E on Figure E.3.7.1). In other words all records with a calculated end date that is greater than the latest data submission cycle date (see Section 3.8.3.2).

This logic resulted in the addition of the PROV\_ACCRED\_IND indicator code and corresponding description on the table DM\_QUAL\_ENROL.

68. PROV\_ACCRED\_IND\_ID and PROV\_ACCRED\_IND\_DESC: Code and corresponding description denoting a record's compliance in regard to whether the provider was accredited to offer the qualification for the duration of the learner's active enrolment on the qualification.

### ***E.3.8 Development of ASOR\_IND***

As detailed in Appendix E.2, the development of ASOR\_IND required the implementation of an indicator that denotes whether the assessor was registered at the time of the

achievement of the qualification. In order to make the analysis of the data more suitable for data mining purposes the indicator was implemented as a data code and corresponding description.

Figure E.3.8.1 illustrates the manner in which ASOR\_IND was developed using four example achieved qualification enrolment records, for an assessor that was registered. The figure shows how:

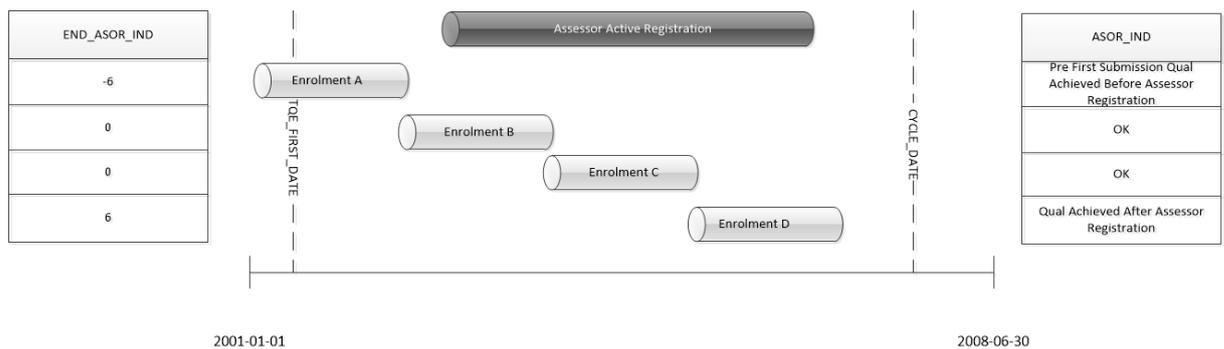


Figure E.3.8.1 Illustrative diagram of ASOR\_IND development

- a qualification enrolment record (Enrolment A) with an end date prior to the registration period of the assessor is allocated an ASOR\_IND value of 'Qual Achieved Before Assessor Registration'. The start of the active enrolment time period precedes the date on which the ETQE submitted its first full data submission to the NLRD, as a result the words 'Pre First Submission' are appended to the value,
- a qualification enrolment record (Enrolment B) with an end date during the registration period of the assessor is allocated an ASOR\_IND value of 'OK',
- a qualification enrolment record (Enrolment C) end date during the registration period of the assessor is allocated an ASOR\_IND value of 'OK', and
- a qualification enrolment record (Enrolment D) with an end date after the registration period of the assessor is allocated an ASOR\_IND value of 'Qual Achieved After Assessor Registration'.

The development of the ASOR\_IND indicator required the implementation of one additional indicator. This indicator assisted in the development of and further description of the value allocated to ASOR\_IND. This additional indicator was developed as a representation of data in relation to a point in time as discussed in Section 3.6.4.4.

The indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the end date (END\_DATE) of the active enrolment time period and assessor's active registration time period (ASOR\_START\_DATE and ASOR\_END\_DATE), where;

- a qualification enrolment record with an end date before the start date of the assessor's registration would be given a negative value of the number of months between these two values (for example Enrolment A on Figure E.3.8.1),
- a qualification enrolment record with an end date that falls between the start and end dates of the assessor's registration is given a value of 0 (for example Enrolment C on Figure E.3.8.1), and
- a qualification enrolment record with an end date that is after the end date of the assessor's registration would be given a positive value of the number of months between these two values (for example Enrolment D on Figure E.3.8.1).

This logic resulted in the addition of the following new indicator on the table DM\_QUAL\_ENROL:

69. END\_ASOR\_IND: Numeric value indicating the distance between the end date of the qualification enrolment record and the assessor registration.

Using the values in this field it was possible to derive a code and corresponding description for ASOR\_IND as follows:

- Where the qualification enrolment had not been achieved, allocating a value of 'Not Achieved'.
- Where the qualification enrolment had been achieved but an assessor identifier had not been provided, allocation a value of 'No Assessor Provided'.
- Where an assessor identifier had been provided but an assessor registration did not exist, allocating a value of 'No Registration' to the record.
- Allocating a value of 'OK' to records where END\_ASOR\_IND is equal to 0.
- Allocating a value of 'Qual Achieved Before Assessor Registration' where END\_ASOR\_IND is less than 0.
- Allocating a value of 'Qual Achieved After Assessor Registration' where END\_ASOR\_IND is greater than 0.

The final derivation steps included:

- Amending the data code and appending the words 'Pre First Submission' to the indicator value for any records with a `START_DATE` value less than `ETQE_FIRST_DATE` (for example Enrolment A on Figure E.3.8.1). In other words all records where the learner enrolled on the qualification prior to the first full data submission from the ETQE to the NLRD (see Section 3.8.3.1).

This logic resulted in the addition of the `ASOR_IND` indicator code and corresponding description on the table `DM_QUAL_ENROL`.

70. `ASOR_IND_ID` and `ASOR_IND_DESC`: Code and corresponding description denoting a record's compliance in regard to whether the assessor was registered at the time of the achievement of the qualification.

### ***E.3.9 Development of ASOR\_REGSTR\_IND***

As detailed in Appendix E.2, the development of `ASOR_REGSTR_IND` required the implementation of an indicator that denotes whether the assessor was registered to assess the qualification at the time of the achievement of the qualification. In order to make the analysis of the data more suitable for data mining purposes the indicator was implemented as a data code and corresponding description.

Figure E.3.9.1 illustrates the manner in which `ASOR_REGSTR_IND` was developed using four example achieved qualification enrolment records, for an assessor that was registered. The figure shows how:

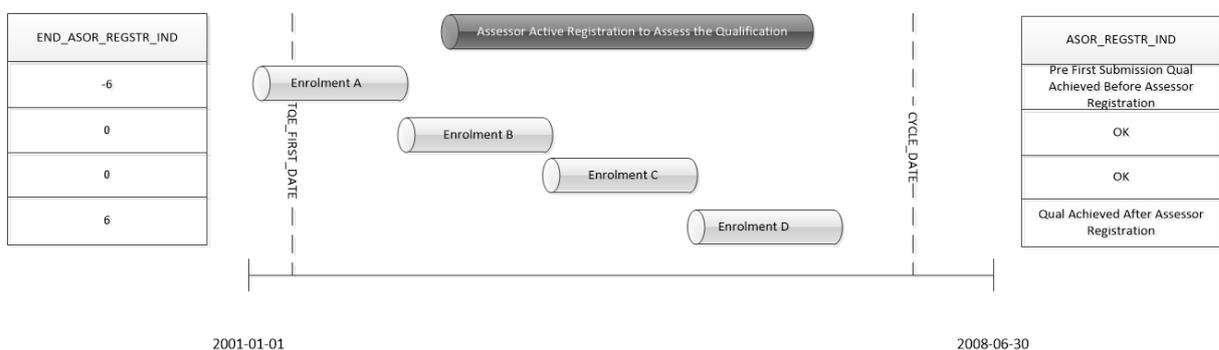


Figure E.3.9.1 Illustrative diagram of `ASOR_REGSTR_IND` development

- a qualification enrolment record (Enrolment A) with an end date prior to the qualification registration period of the assessor is allocated an `ASOR_REGSTR_IND`

value of 'Qual Achieved Before Assessor Registration'. The start of the active enrolment time period precedes the date on which the ETQE submitted its first full data submission to the NLRD, as a result the words 'Pre First Submission' are appended to the value,

- a qualification enrolment record (Enrolment B) with an end date during the qualification registration period of the assessor is allocated an ASOR\_REGSTR\_IND value of 'OK',
- a qualification enrolment record (Enrolment C) end date during the qualification registration period of the assessor is allocated an ASOR\_REGSTR\_IND value of 'OK', and
- a qualification enrolment record (Enrolment D) with an end date after the qualification registration period of the assessor is allocated an ASOR\_REGSTR\_IND value of 'Qual Achieved After Assessor Registration'.

The development of the ASOR\_REGSTR\_IND indicator required the implementation of one additional indicator. This indicator assisted in the development of and further description of the value allocated to ASOR\_REGSTR\_IND. This additional indicator was developed as a representation of data in relation to a point in time as discussed in Section 3.6.4.4.

The indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the end date (END\_DATE) of the active enrolment time period and assessor's active registration to assess the qualification time period (ASOR\_REGSTR\_START\_DATE and ASOR\_REGSTR\_END\_DATE), where;

- a qualification enrolment record with an end date before the start date of the assessor's registration to assess the qualification would be given a negative value of the number of months between these two values (for example Enrolment A on Figure E.3.9.1),
- a qualification enrolment record with an end date that falls between the start and end dates of the assessor's registration to assess the qualification is given a value of 0 (for example Enrolment C on Figure E.3.9.1), and
- a qualification enrolment record with an end date that is after the end date of the assessor's registration to assess the qualification would be given a positive value of the

number of months between these two values (for example Enrolment D on Figure E.3.9.1).

This logic resulted in the addition of the following new indicator on the table DM\_QUAL\_ENROL:

71. END\_ASOR\_REGSTR\_IND: Numeric value indicating the distance between the end date of the qualification enrolment record and the assessor registration to assess the qualification.

Using the values in this field it was possible to derive a code and corresponding description for ASOR\_REGSTR\_IND as follows:

- Where the qualification enrolment had not been achieved, allocating a value of 'Not Achieved'.
- Where the qualification enrolment had been achieved but an assessor identifier had not been provided, allocation a value of 'No Assessor Provided'.
- Where an assessor identifier had been provided but an assessor registration to assess the qualification did not exist, allocating a value of 'No Registration' to the record.
- Allocating a value of 'OK' to records where END\_ASOR\_REGSTR\_IND is equal to 0.
- Allocating a value of 'Qual Achieved Before Assessor Registration' where END\_ASOR\_REGSTR\_IND is less than 0.
- Allocating a value of 'Qual Achieved After Assessor Registration' where END\_ASOR\_REGSTR\_IND is greater than 0.

The final derivation steps included:

- Amending the data code and appending the words 'Pre First Submission' to the indicator value for any records with a START\_DATE value less than ETQE\_FIRST\_DATE (for example Enrolment A on Figure E.3.9.1). In other words all records where the learner enrolled on the qualification prior to the first full data submission from the ETQE to the NLRD (see Section 3.8.3.1).

This logic resulted in the addition of the ASOR\_REGSTR\_IND indicator code and corresponding description on the table DM\_QUAL\_ENROL.

72. ASOR\_REGSTR\_IND\_ID and ASOR\_REGSTR\_IND\_DESC: Code and corresponding description denoting a record's compliance in regard to whether the assessor was registered to assess the qualification at the time of the achievement of the qualification.

### E.3.10 Development of QUAL\_REGSTR\_IND

As detailed in Appendix E.2, the development of QUAL\_REGSTR\_IND required the implementation of an indicator that denotes whether the qualification was registered for the duration of the learner's active enrolment on the qualification. In order to make the analysis of the data more suitable for data mining purposes the indicator was implemented as a data code and corresponding description.

Figure E.3.10.1 illustrates the manner in which QUAL\_REGSTR\_IND was developed using five example qualification enrolment records, for a qualification that was registered. The figure shows how:

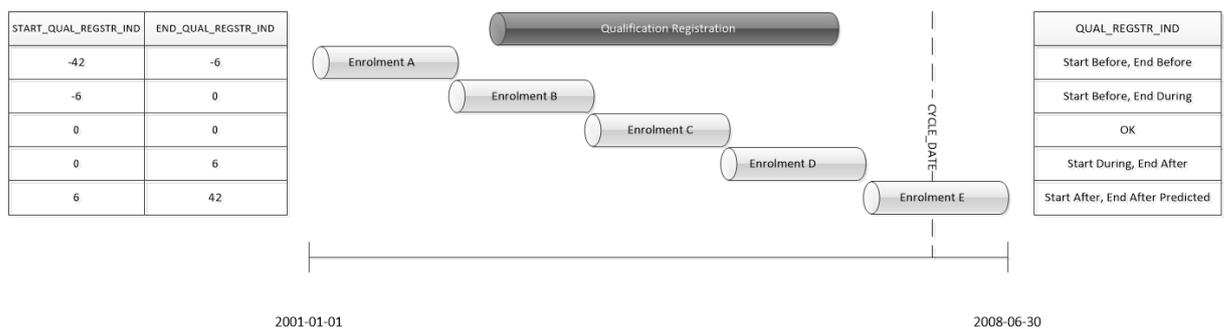


Figure E.3.10.1 Illustrative diagram of QUAL\_REGSTR\_IND development

- a qualification enrolment record (Enrolment A) with a start date prior to the registration of the qualification registration and an end date prior to the registration of the qualification is allocated a QUAL\_REGSTR\_IND value of 'Start Before, End Before',
- a qualification enrolment record (Enrolment B) with a start date prior to the registration of the qualification and an end date that is during the registration of the qualification is allocated a QUAL\_REGSTR\_IND value of 'Start Before, End During',

- a qualification enrolment record (Enrolment C) with a start date that is during the registration of the qualification and an end date that is during the registration of the qualification is allocated a QUAL\_REGSTR\_IND value of ‘OK’,
- a qualification enrolment record (Enrolment D) with a start date that is during the registration of the qualification and an end date that is after the registration of the qualification is allocated a QUAL\_REGSTR\_IND value of ‘Start During, End After’, and
- a qualification enrolment record (Enrolment E) with a start date that is after the registration of the qualification and an end date that is after the registration of the qualification is allocated a QUAL\_REGSTR\_IND value of ‘Start After, End After’, and because the end of the active enrolment time period exceeds the latest data submission cycle date the word ‘Predicted’ is appended to the value.

The development of the QUAL\_REGSTR\_IND indicator required the implementation of two additional indicators. These indicators assisted in the development of and further description of the value allocated to QUAL\_REGSTR\_IND. Both of these two additional indicators were developed as representations of data in relation to a point in time as discussed in Section 3.6.4.4.

The first indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the start date (START\_DATE) of the active enrolment time period and the start date of the qualification’s registration and the last date on which a learner may enrol on a qualification (QUAL\_START\_DATE and QUAL\_MAX\_START\_DATE), where;

- a qualification enrolment record with a start date before the start date of the qualification’s registration (QUAL\_START\_DATE) would be given a negative value of the number of months between these two values (for example Enrolment A on Figure E.3.10.1),
- a qualification enrolment record with a start date that falls between the start date of the qualification’s registration (QUAL\_START\_DATE) and the last date on which a learner may enrol on a qualification (QUAL\_MAX\_START\_DATE) is given a value of 0 (for example Enrolment C on Figure E.3.10.1), and

- a qualification enrolment record with a start date that is after the last date on which a learner may enrol on a qualification would be given a positive value of the number of months between these two values (for example Enrolment E on Figure E.3.10.1).

The second indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the end date (END\_DATE) of the active enrolment time period and the start date of the qualification's registration and the last date on which a learner may achieve a qualification (QUAL\_START\_DATE and QUAL\_MAX\_END\_DATE)), where;

- a qualification enrolment record with an end date before the start date of the qualification's registration would be given a negative value of the number of months between these two values (for example Enrolment A on Figure E.3.10.1),
- a qualification enrolment record with an end date that falls between the start date of the qualification registration and the last date on which a learner may achieve a qualification is given a value of 0 (for example Enrolment C on Figure E.3.10.1), and
- a qualification enrolment record with an end date that is after the last date on which a learner may achieve a qualification would be given a positive value of the number of months between these two values (for example Enrolment E on Figure E.3.10.1).

This logic resulted in the addition of the following new indicators on the table DM\_QUAL\_ENROL:

73. START\_QUAL\_REGSTR\_IND: Numeric value indicating the distance between the start date of the qualification enrolment record and the qualification's active registration.

74. END\_QUAL\_REGSTR\_IND: Numeric value indicating the distance between the end date of the qualification enrolment record and the qualification's active registration.

Using the values in these two fields in conjunction with each other it was possible to derive a code and corresponding description for QUAL\_REGSTR\_IND as follows:

- Where a qualification's registration did not exist, allocating a value of ' No Registration' to the record.
- Allocating a value of 'OK' to records where START\_QUAL\_REGSTR\_IND is equal to 0 and END\_QUAL\_REGSTR\_IND is equal to 0.

For all remaining records:

- Allocating a value of ‘Start Before’ to records where START\_QUAL\_REGSTR\_IND was less than 0, ‘Start During’ to records where START\_QUAL\_REGSTR\_IND is equal to 0 and ‘Start After’ where START\_QUAL\_REGSTR\_IND was greater than 0.
- Allocating a value ‘End Before’ to records where END\_QUAL\_REGSTR\_IND was less than 0, ‘End During’ to records where END\_QUAL\_REGSTR\_IND is equal to 0 and ‘End After’ where END\_QUAL\_REGSTR\_IND was greater than 0.

The final derivation steps included:

- Amending the data code and appending the word ‘Predicted’ to the indicator value for any records with a END\_DATE value greater than CYCLE\_DATE (for example Enrolment E on Figure E.3.10.1). In other words all records with a calculated end date that is greater than the latest data submission cycle date (see Section 3.8.3.2).

This logic resulted in the addition of the QUAL\_REGSTR\_IND indicator code and corresponding description on the table DM\_QUAL\_ENROL.

75. QUAL\_REGSTR\_IND\_ID and QUAL\_REGSTR\_IND\_DESC: Code and corresponding description denoting a record’s compliance in regard to whether the qualification was registered for the duration of the learner’s active enrolment on the qualification.

### ***E.3.11 Development of USTD\_MIX\_IND***

As detailed in Appendix E.2, the development of UNIT\_STD\_MIX\_IND required the implementation of an indicator that denotes whether, in the instance where a learner has achieved a unit standards based qualification, the learner achieved:

- the minimum required number of credits for the qualification, on or before the achievement of the qualification, based on the achievement of unit standards related to the qualification, and
- the correct range of credits for the qualification, achieved on or before the achievement of the qualification, based on the achievement of unit standards that have been defined as core, fundamental and elective unit standards for the qualification.

In order to make the analysis of the data more suitable for data mining purposes the indicator was implemented as a data code and corresponding description.

The development of the UNIT\_STD\_MIX\_IND indicator value required the implementation of a number of additional indicators that assisted in the development of UNIT\_STD\_MIX\_IND and would help to further describe the value assigned to UNIT\_STD\_MIX\_IND. These indicators were developed as follows:

To ensure that the logic had access to records for both the replaced unit standards and the unit standards that replaced them, additional unit standard qualification link records were created using records found in DM\_USTD\_REPL (the table that stores unit standard replacements, see Section 3.8.2.18). Where the unit standard that a qualification was linked to was found as an “old” unit standard (OLD\_USTD\_ID) in DM\_USTD\_REPL, a new unit standard qualification link was created using the “new” unit standard (NEW\_USTD\_ID). These derived unit standard qualification links and the original unit standard qualification links found in DM\_USTD\_QUAL were saved to a new table called DM\_USTD\_QUAL\_FINAL. The contents of this new table were used to determine the link between a qualification and a unit standard.

The actual total number of credits that the learner achieved against the unit standards based qualification on or before the achievement of the qualification was calculated by first linking the table DM\_QUAL\_ENROL to the table DM\_USTD\_QUAL\_FINAL using the unique identifier of the qualification (QUALIFICATION\_ID). In this manner it was possible to determine which unit standards may contribute to the achievement of a unit standards based qualification. These results were then linked to the table DM\_USTD\_ENROL using the combination of the unique identifier of the learner (LEARNER\_ID) and the unique identifier of the unit standard (UNIT\_STANDARD\_ID). In this manner it was possible to calculate all of the credits that the learner had achieved for the qualification on or before the achievement of the qualification.

This logic resulted in the addition of a new indicator on the table DM\_QUAL\_ENROL:

76. ACT\_CREDITS\_TOTAL: Numeric value indicating the total unit standard credits, as found in the table DM\_USTD\_ENROL, achieved against the unit standard based qualification on or before the achievement of the qualification.

By comparing the actual total number of credits that the learner achieved (ACT\_CREDITS\_TOTAL) against the required number of credits the learner should have

achieved (CREDITS on DM\_QUAL\_ENROL) a new indicator that contained both a code and descriptor was developed as follows:

- Allocating a value of ‘Not Unit Standard Based’ if the qualification was not unit standards based.
- Allocating a value of ‘Not achieved’ if the qualification enrolment had not been achieved.
- Allocating a value of ‘Sufficient Credits Achieved’ if the learner had achieved the required number of credits or more.
- Allocating a value of ‘Insufficient Credits Achieved’ if the learner had not achieved the required number of credits.
- Allocating a value of ‘No Unit Standards Achieved’ if the learner had not achieved any credits for a unit standards based qualification.

This logic resulted in the addition of the USTD\_CREDIT\_IND indicator code and corresponding description on the table DM\_QUAL\_ENROL.

77. USTD\_CREDIT\_IND\_ID and USTD\_CREDIT\_IND\_DESC: Code and corresponding description denoting a record’s compliance in regard to whether the learner had achieved the required number of credits for a unit standards based qualification.

The same logic, as described above, was used to determine the actual total number of core/fundamental and elective credits that the learner had achieved.

This logic resulted in the addition of three new indicators on the table DM\_QUAL\_ENROL:

78. ACT\_CREDITS\_CORE: Numeric value indicating the total core unit standard credits, as found in the table DM\_USTD\_ENROL, achieved against the unit standard based qualification on or before the achievement of the qualification.

79. ACT\_CREDITS\_FUND: Numeric value indicating the total fundamental unit standard credits, as found in the table DM\_USTD\_ENROL, achieved against the unit standard based qualification on or before the achievement of the qualification.

80. ACT\_CREDITS\_ELEC: Numeric value indicating the total elective unit standard credits, as found in the table DM\_USTD\_ENROL, achieved against the unit standard based qualification on or before the achievement of the qualification.

By comparing the actual total number of core/fundamental/elective credits that the learner had achieved (ACT\_CREDITS\_CORE, ACT\_CREDITS\_FUND and ACT\_CREDITS\_ELEC) against the required number of core/fundamental/elective credits the learner should have achieved (USTD\_CREDITS\_CORE, USTD\_CREDITS\_FUND and USTD\_CREDITS\_ELECT on DM\_QUAL\_ENROL) three new indicators that contained both a code and descriptor was developed as follows:

- Allocating a value of ‘[Core]/[Fundamental]/[Elective] Credits OK’ if the learner had achieved the required number of core/fundamental/elective credits or more.
- Allocating a value of ‘Insufficient [Core]/[Fundamental]/[Elective] Credits’ if the learner had not achieved the required number of credits.

Additionally the difference between the actual number of core/fundamental/elective credits achieved and the required number of core/fundamental/elective credits were saved as indicators. This logic resulted in the addition of the following indicators on the table DM\_QUAL\_ENROL.

81. USTD\_CORE\_IND\_ID and USTD\_CORE\_IND\_DESC: Code and corresponding description denoting a record’s compliance in regard to whether the learner achieved the required number of core unit standards.
82. USTD\_FUND\_IND\_ID and USTD\_FUND\_IND\_DESC: Code and corresponding description denoting a record’s compliance in regard to whether the learner achieved the required number of fundamental unit standards.
83. USTD\_ELEC\_IND\_ID and USTD\_ELEC\_IND\_DESC: Code and corresponding description denoting a record’s compliance in regard to whether the learner achieved the required number of elective unit standards.
84. USTD\_CORE\_DIFF: the difference between the number of core credits that the learner achieved and the required number of core credits that the learner should have achieved.
85. USTD\_FUND\_DIFF: the difference between the number of fundamental credits that the learner achieved and the required number of fundamental credits that the learner should have achieved.
86. USTD\_ELEC\_DIFF: the difference between the number of elective credits that the learner achieved and the required number of elective credits that the learner should have achieved.

Finally, the indicators described above were utilized to develop a USTD\_MIX\_IND indicator using the following type of logic:

- Allocating a value of ‘Not Unit Standard Based’ if the qualification is not a unit standards based qualification.
- Allocating a value of ‘Not Achieved’ if the qualification enrolment had not been achieved.
- Allocating a value of ‘No Unit Standards Achieved’ if the learner had not achieved any unit standards against the qualification enrolment.
- Allocating a value of ‘Sufficient Credits Achieved’ if the learner achieved sufficient credits in total and sufficient core, fundamental and elective credits for the qualification.
- Allocating a concatenated value of the value for USTD\_CREDIT\_IND\_DESC, USTD\_CORE\_IND\_DESC, USTD\_FUND\_IND\_DESC and USTD\_ELEC\_IND\_DESC if the learner did not achieve the required number of core, fundamental and/or elective credits for the qualification.

This logic resulted in the addition of the USTD\_MIX\_IND indicator code and corresponding description on the table DM\_QUAL\_ENROL.

87. USTD\_MIX\_IND\_ID and USTD\_MIX\_IND\_DESC: Code and corresponding description denoting a record’s compliance in regard to whether the learner achieved:

- the minimum required number of credits for the qualification, on or before the achievement of the qualification, based on the achievement of unit standards related to the qualification, and
- the correct range of credits for the qualification, achieved on or before the achievement of the qualification, based on the achievement of unit standards that have been defined as core, fundamental and elective unit standards for the qualification.

### ***E.3.12 Removal of replaced qualification enrolments***

The evolution of qualifications as described in Section 3.8.3.4 can in some instances result in a data management issue that must be addressed for the purposes of this research.

In some instances, a provider may enrol the learner on a qualification that has been replaced. The enrolment of the qualification that has been replaced may be captured on the operational information system of the ETQE and as a result may be submitted to the NLRD. The NLRD has clearly defined protocol that allows the ETQE to indicate when a record has been incorrectly submitted to the NLRD, unfortunately not all ETQEs complete the protocol in this regard. On discovery of the enrolment on the replaced qualification the ETQE may incorrectly perform the following actions on their operational information system:

- update the existing enrolment record with the replacement qualification ID, or
- create an entirely new enrolment record for the learner against the replacement qualification.

The data loading procedures of the NLRD work on the principle that the combination of learner ID and qualification ID is unique. As a result, in either scenario as described above, when the enrolment against the replacement qualification is loaded on the NLRD, a new qualification enrolment record is created for the learner. Enrolment records against qualifications that have been replaced, that have been incorrectly loaded on the NLRD will invariably generate false positives against a number of the semantic business rules. In consultation with the Director of the NLRD it was decided that such records should be excluded from this research.

As a result, any enrolment records against a qualification that has been replaced, which has a further enrolment for the same learner, against the replacement qualification were allocated a `PROBLEM_ID` of 5 and excluded from the further processing of the data. These records constituted 1.82% of the qualification enrolment records initially extracted from the NLRD.

### ***E.3.13 DM\_QUAL\_ENROL\_FINAL***

The derivation steps described from Appendix E.3.1 to Appendix E.3.11 were saved in a new data table called `DM_QUAL_ENROL_FINAL`. This table included all of the data records initially received from the NLRD in the table `DM_QUAL_ENROL`, including the problem records described in Appendix E.3.1, Appendix E.3.2, Appendix E.3.3 and Appendix E.3.12 (i.e. records that have a value in the data field `PROBLEM_ID`). The

problem records were immediately communicated to SAQA, who in turn implemented processes and procedures in order to address these records.

A technical description of the table DM\_QUAL\_ENROL\_FINAL has been provided in E.1.

#### ***E.4 Appendix summary***

This section detailed the development of the data table DM\_QUAL\_ENROL\_FINAL which will be used in the analysis and data mining of the qualification enrolment records received from the NLRD. The section identified the semantic business rules that are applicable to qualification enrolment records and then described the selection, pre-processing and derivation steps implemented to establish the table DM\_QUAL\_ENROL\_FINAL, which contains the qualification enrolment records in a format that is suited for data mining.

## Appendix F

This appendix provides a detailed specification of the table DM\_QUAL\_ENROL\_FINAL.

Table Name	Column Name	Data Type	Data Length	Allow NULLs	Comment
DM_QUAL_ENROL_FINAL	LEARNER_ENROLMENT_ID	NUMBER	22	N	De-identified
DM_QUAL_ENROL_FINAL	LEARNER_ID	NUMBER	22	N	De-identified
DM_QUAL_ENROL_FINAL	QUALIFICATION_ID	NUMBER	22	Y	De-identified
DM_QUAL_ENROL_FINAL	LEARNERSHIP_ID	NUMBER	22	Y	De-identified
DM_QUAL_ENROL_FINAL	ETQE_ID	NUMBER	22	N	De-identified
DM_QUAL_ENROL_FINAL	PROVIDER_ID	NUMBER	22	Y	De-identified
DM_QUAL_ENROL_FINAL	ASSESSOR_ID	NUMBER	22	Y	De-identified
DM_QUAL_ENROL_FINAL	ENROL_STATUS_ID	NUMBER	22	N	
DM_QUAL_ENROL_FINAL	ENROL_STATUS_DESC	VARCHAR2	26	N	
DM_QUAL_ENROL_FINAL	ENROL_TYPE_ID	NUMBER	22	N	
DM_QUAL_ENROL_FINAL	ENROL_TYPE_DESC	VARCHAR2	50	N	
DM_QUAL_ENROL_FINAL	ENROL_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	ACHIEVE_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	DERIVED_START_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	CREDITS	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	QUAL_START_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	QUAL_END_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	QUAL_MAX_START_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	QUAL_MAX_END_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	TRANSITION_PERIOD	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	TRAIN_OUT_PERIOD	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	NQF_LEVEL_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	NQF_LEVEL_DESC	VARCHAR2	26	Y	
DM_QUAL_ENROL_FINAL	QUALIFICATION_TYPE_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	QUALIFICATION_TYPE_DESC	VARCHAR2	30	Y	
DM_QUAL_ENROL_FINAL	QUALIFICATION_CLASS_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	QUALIFICATION_CLASS_DESC	VARCHAR2	26	Y	
DM_QUAL_ENROL_FINAL	FIELD_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	FIELD_DESC	VARCHAR2	60	Y	
DM_QUAL_ENROL_FINAL	SUBFIELD_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	SUBFIELD_DESC	VARCHAR2	80	Y	
DM_QUAL_ENROL_FINAL	PROBLEM_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	START_DATE_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	START_DATE_DESC	VARCHAR2	26	Y	
DM_QUAL_ENROL_FINAL	START_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	END_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	ETQE_START_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	ETQE_END_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	ETQE_ACCRED_START_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	ETQE_ACCRED_END_DATE	DATE	7	Y	

Table Name	Column Name	Data Type	Data Length	Allow NULLs	Comment
DM_QUAL_ENROL_FINAL	PROV_START_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	PROV_END_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	PROVIDER_TYPE_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	PROVIDER_TYPE_DESC	VARCHAR2	26	Y	
DM_QUAL_ENROL_FINAL	PROVIDER_CLASS_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	PROVIDER_CLASS_DESC	VARCHAR2	50	Y	
DM_QUAL_ENROL_FINAL	PROV_PROVINCE_CODE	VARCHAR2	10	Y	
DM_QUAL_ENROL_FINAL	PROV_PROVINCE_DESC	VARCHAR2	60	Y	
DM_QUAL_ENROL_FINAL	PROV_ACCRED_START_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	PROV_ACCRED_END_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	ASOR_START_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	ASOR_END_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	ASOR_REGSTR_START_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	ASOR_REGSTR_END_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	USTD_CREDITS_TOTAL	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	USTD_CREDITS_CORE	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	USTD_CREDITS_FUND	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	USTD_CREDITS_ELEC	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	ETQE_FIRST_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	CYCLE_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	START_ETQE_IND	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	END_ETQE_IND	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	ETQE_IND_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	ETQE_IND_DESC	VARCHAR2	134	Y	
DM_QUAL_ENROL_FINAL	OTHR_ETQE_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	OTHR_ETQE_START_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	OTHR_ETQE_END_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	OTHR_START_ETQE_IND	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	OTHR_END_ETQE_IND	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	OTHR_ETQE_IND_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	OTHR_ETQE_IND_DESC	VARCHAR2	24	Y	
DM_QUAL_ENROL_FINAL	START_ETQE_ACCRED_IND	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	END_ETQE_ACCRED_IND	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	ETQE_ACCRED_IND_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	ETQE_ACCRED_IND_DESC	VARCHAR2	134	Y	
DM_QUAL_ENROL_FINAL	OTHR_ETQE_ACCRED_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	OTHR_ETQE_ACCRED_START_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	OTHR_ETQE_ACCRED_END_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	OTHR_START_ETQE_ACCRED_IND	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	OTHR_END_ETQE_ACCRED_IND	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	OTHR_ETQE_ACCRED_IND_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	OTHR_ETQE_ACCRED_IND_DESC	VARCHAR2	24	Y	
DM_QUAL_ENROL_FINAL	PROV_ETQE_ID	NUMBER	22	Y	De-identified
DM_QUAL_ENROL_FINAL	PROV_ETQE_FIRST_DATE	DATE	7	Y	
DM_QUAL_ENROL_FINAL	START_PROV_IND	NUMBER	22	Y	

Table Name	Column Name	Data Type	Data Length	Allow NULLs	Comment
DM_QUAL_ENROL_FINAL	END_PROV_IND	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	PROV_IND_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	PROV_IND_DESC	VARCHAR2	45	Y	
DM_QUAL_ENROL_FINAL	START_PROV_ACCRED_IND	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	END_PROV_ACCRED_IND	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	PROV_ACCRED_IND_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	PROV_ACCRED_IND_DESC	VARCHAR2	45	Y	
DM_QUAL_ENROL_FINAL	END_ASOR_IND	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	ASOR_IND_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	ASOR_IND_DESC	VARCHAR2	63	Y	
DM_QUAL_ENROL_FINAL	END_ASOR_REGSTR_IND	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	ASOR_REGSTR_IND_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	ASOR_REGSTR_IND_DESC	VARCHAR2	63	Y	
DM_QUAL_ENROL_FINAL	START_QUAL_REGSTR_IND	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	END_QUAL_REGSTR_IND	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	QUAL_REGSTR_IND_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	QUAL_REGSTR_IND_DESC	VARCHAR2	34	Y	
DM_QUAL_ENROL_FINAL	ACT_CREDITS_TOTAL	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	USTD_CREDIT_IND_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	USTD_CREDIT_IND_DESC	VARCHAR2	29	Y	
DM_QUAL_ENROL_FINAL	ACT_CREDITS_CORE	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	ACT_CREDITS_FUND	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	ACT_CREDITS_ELEC	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	USTD_CORE_DIFF	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	USTD_CORE_IND_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	USTD_CORE_IND_DESC	VARCHAR2	25	Y	
DM_QUAL_ENROL_FINAL	USTD_FUND_DIFF	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	USTD_FUND_IND_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	USTD_FUND_IND_DESC	VARCHAR2	32	Y	
DM_QUAL_ENROL_FINAL	USTD_ELEC_DIFF	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	USTD_ELEC_IND_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	USTD_ELEC_IND_DESC	VARCHAR2	29	Y	
DM_QUAL_ENROL_FINAL	USTD_MIX_IND_ID	NUMBER	22	Y	
DM_QUAL_ENROL_FINAL	USTD_MIX_IND_DESC	VARCHAR2	121	Y	

## Appendix G

### ***G.1 Introduction***

This section details the initial selection, pre-processing and derivation of the unit standard enrolment records, received from the NLRD in the table DM\_USTD\_ENROL, into a format that is suitable for data mining.

The specific semantic business rules that are applicable to unit standard enrolment records are identified in Appendix G.2. The analysis and data mining of these semantic business rules requires the implementation of seven (7) semantic business rule indicators. Appendix G.3 describes the selection, pre-processing and derivation steps required for the implementation of these semantic business rule indicators. Appendix G.3.1 and Appendix G.3.2 describe the type of logic developed for the selection and pre-processing of the data. Whereas Appendix G.3.4 to Appendix G.3.10 describe the type of logic used for the derivation of the data.

The selection, pre-processing and derivation logic resulted in the implementation of a final version of the unit standard enrolment data as a new table called DM\_USTD\_ENROL\_FINAL, described in Appendix G.3.12.

### ***G.2 Applicable semantic business rules and their indicator fields***

A review of the final version of the semantic business rules (see Section 3.6.2) shows that the business rules that are applicable to unit standard enrolment records are as follows:

1. that the ETQE that submitted the record
  - a. was accredited for the duration of the learner's active enrolment on the learnership/qualification/unit standard
  - b. was accredited to quality assure the qualification/unit standard for the duration of the learner's active enrolment on the qualification/unit standard
2. that the provider
  - a. was accredited for the duration of the learner's active enrolment on the learnership/qualification/unit standard
  - b. was accredited to offer the qualification/unit standard for the duration of the learner's active enrolment on the learnership/qualification/unit standard
3. that if the learner has completed the learnership or achieved the qualification/unit standard and the details of the assessor are supplied, that the assessor

- a. was registered at the time of the completion of the learnership or achievement of the qualification/unit standard
  - b. was registered to assess the qualification/unit standard at the time of the completion of the learnership or achievement of the qualification/unit standard
4. that the qualification/unit standard was registered for the duration of the learner's active enrolment on the qualification/unit standard

The main purpose of the derivation of the unit standard enrolment data for analysis and data mining therefore focused on the development of seven (7) semantic business rule indicators (each consisting of a data code and a description) that described the compliance of the record in accordance with these rules:

1. ETQE\_IND

Denotes whether the ETQE was accredited for the duration of the learner's active enrolment on the unit standard.

2. ETQE\_ACCRED\_IND

Denotes whether the ETQE was accredited to quality assure the unit standard for the duration of the learner's active enrolment on the unit standard.

3. PROV\_IND

Denotes whether the provider was accredited for the duration of the learner's active enrolment on the unit standard.

4. PROV\_ACCRED\_IND

Denotes whether the provider was accredited to offer the unit standard for the duration of the learner's active enrolment on the unit standard.

5. ASOR\_IND

Denotes whether the assessor was registered at the time of the achievement of the unit standard.

6. ASOR\_REGSTR\_IND

Denotes whether the assessor was registered to assess the unit standard at the time of the achievement of the unit standard.

7. USTD\_REGSTR\_IND

Denotes whether the unit standard was registered for the duration of the learner's active enrolment on the unit standard.

### ***G.3 Semantic business rule indicator development steps***

### ***G.3.1 Pre derivation data collection***

By their very definition on the NQF, unit standards have start and end dates, denoted by the unit standards registration start and end dates.

Unit standards are seldom enrolled on or achieved as standalone units of learning, rather they are the discrete parts of a qualification. As a result, the start and end date of the qualification that a unit standard enrolment is linked to takes priority over the dates of the unit standard. The end date of a qualification does not indicate the last date on which a learner may enrol on or achieve the unit standards linked to the qualification. Rather, qualifications have transition and train-out time periods that allow for the graceful ending of a qualification in order to allow all stakeholders and learners sufficient time to transition to a qualification's replacement qualification (Section 3.8.2.11).

Consequently the first step of the development of the semantic business rule indicators for the unit standard enrolment records focused on developing data fields that defined the last date on which a learner may have enrolled on a unit standard and the last day on which a learner may have achieved a unit standard if the unit standard enrolment is linked to a qualification enrolment.

In order to address unit standard enrolment records that are linked to qualifications a table was created that described all the permitted combinations of qualifications and unit standards. To ensure that the logic had access to records for both the replaced unit standards and the unit standards that replaced them, additional unit standard qualification link records were created using records found in DM\_USTD\_REPL (the table that stores unit standard replacements, see Section 3.8.2.18). Where the unit standard that a qualification was linked to was found as an "old" unit standard (OLD\_USTD\_ID) in DM\_USTD\_REPL, a new unit standard qualification link was created using the "new" unit standard (NEW\_USTD\_ID). These derived unit standard qualification links and the original unit standard qualification links found in DM\_USTD\_QUAL were saved to a new table called DM\_USTD\_QUAL\_FINAL.

Further, a new table was created called DM\_QUAL\_FINAL which contains the same data fields as the table DM\_QUAL with the addition of two new derived data fields namely:

1. **MAX\_START\_DATE**: A calculated value indicating the last date on which a learner may enrol on a unit standard. The value is calculated as the transition period added to the end date of the unit standard.

$$\text{MAX\_START\_DATE} = \text{TRANSITION\_PERIOD} + \text{END\_DATE}$$

2. **MAX\_END\_DATE**: A calculated value indicating the last date on which a learner may achieve a unit standard. The value is calculated as two years plus the train out period of the unit standard added to the last date on which the learner may enrol on the unit standard.

$$\text{MAX\_END\_DATE} = \text{MAX\_START\_DATE} + 2 + \text{TRAIN\_OUT\_PERIOD}$$

The resulting table **DM\_QUAL\_FINAL** was used instead of the table **DM\_QUAL** to source data values that described the qualification linked unit standard that the learner had enrolled on during the development of the semantic business rule indicators.

Finally, a new table was created called **DM\_USTD\_FINAL** which contains the same data fields as the table **DM\_USTD** with the addition of two new derived data fields namely:

3. **MAX\_START\_DATE**: A calculated value indicating the last date on which a learner may enrol on a unit standard. The value is calculated as the transition period added to the end date of the unit standard.

$$\text{MAX\_START\_DATE} = \text{TRANSITION\_PERIOD} + \text{END\_DATE}$$

4. **MAX\_END\_DATE**: A calculated value indicating the last date on which a learner may achieve a unit standard. The value is calculated as two years plus the train out period of the unit standard added to the last date on which the learner may enrol on the unit standard.

$$\text{MAX\_END\_DATE} = \text{MAX\_START\_DATE} + 2 + \text{TRAIN\_OUT\_PERIOD}$$

The resulting table **DM\_USTD\_FINAL** was used instead of the table **DM\_USTD** to source data values that described the unit standard that the learner had enrolled on during the development of the semantic business rule indicators.

Determining compliance of a data record in regard to all of the semantic business rules that are applicable to unit standard enrolment records required that each unit standard enrolment record have an active enrolment time period. An active enrolment time period needed to be derived for unit standard enrolment records that did not have an enrolment date and/or an achievement date (Section 3.6.4.1). Deriving the active enrolment time period for these

types of enrolment records was accomplished utilizing the calculation of credits to notional hours (Appendix A.2) using the credits of the unit standard.

As a result the next step of the development of the semantic business rule indicators focused on obtaining additional data, including the credits, for the unit standard of the unit standard enrolment record.

The linking of the unit standard enrolment record to its unit standard record in the table DM\_USTD\_FINAL resulted in the addition of the following data fields to the table DM\_USTD\_ENROL:

5. CREDITS: the credits for the unit standard utilized to derive the active enrolment time period of the unit standard enrolment record if required.
6. USTD\_START\_DATE: the active registration start date of the unit standard utilized to derive the active enrolment time period of the unit standard enrolment record if required.
7. USTD\_END\_DATE: the active registration end date of the unit standard utilized to derive the active enrolment time period of the unit standard enrolment record if required.
8. TRANSITION\_PERIOD: the transition period for the unit standard.
9. TRAIN\_OUT\_PERIOD: the train out period for the unit standard.
10. USTD\_MAX\_START\_DATE: the last date on which a learner may enrol on the unit standard.
11. USTD\_MAX\_END\_DATE: the last date on which a learner may achieve the unit standard.
12. NQF\_LEVEL\_ID and NQF\_LEVEL\_DESC: The data code and corresponding description of the NQF Level of the unit standard.
13. UNIT\_STANDARD\_TYPE\_ID and UNIT\_STANDARD\_TYPE\_DESC: The data code and corresponding description of the Unit standard Type of the unit standard.
14. FIELD\_ID and FIELD\_DESC: The data code and corresponding description of the Field of the unit standard.
15. SUBFIELD\_ID and SUBFIELD\_DESC: The data code and corresponding description of the Subfield of the unit standard.

The above-mentioned logic however did not take into consideration unit standard enrolment records that were:

- linked to a qualification directly (i.e. the field QUALIFICATION\_ID on the table DM\_USTD\_ENROL contained a value). For all DM\_USTD\_ENROL records that had a value in QUALIFICATION\_ID, where the unit standard was legitimately linked to the qualification (this was determined using the table DM\_USTD\_QUAL\_FINAL), or
- linked to a qualification indirectly (i.e. the field QUALIFICATION\_ID on the table DM\_USTD\_ENROL was blank or invalid, however the same learner had an enrolment record in DM\_QUAL\_ENROL for a qualification that is linked to the unit standard). This linkage was determined using DM\_USTD\_QUAL\_FINAL and DM\_QUAL\_ENROL.

In both instances the unit standard start date, end date, transition period, train out period, maximum start date and maximum end date was replaced with that of the qualifications values (found in the table DM\_QUAL\_FINAL).

Learners that complete their unit standard via distance learning are given more time in which to complete their unit standard. In this type of scenario, the train out period for the unit standard must be multiplied by 1.5. As a result, the MAX\_END\_DATE value for records where ENROL\_TYPE\_DESC = 'Distance Learning' needed to be recalculated as follows:

$$\text{MAX\_END\_DATE} = \text{MAX\_START\_DATE} + 2 + (\text{TRAIN\_OUT\_PERIOD} * 1.5)$$

Any record that could not be linked to a unit standard record in the table DM\_USTD\_FINAL could not be further processed for the development of the semantic business rule indicators and as a result needed to be excluded from the research. Any records that could not be linked to a unit standard record on the table DM\_USTD\_FINAL were allocated a PROBLEM\_ID of 1 (no such records were found) and excluded from the further processing of the data.

Further, any record that had a 0 or NULL credit value, and was missing a value for the enrolment date or achievement date also needed to be excluded from the research because an active enrolment time period could not be derived for these records. As a result any records that had a CREDITS value of 0 or NULL, and had a NULL value for the

ENROL\_DATE or ACHIEVE\_DATE fields were allocated a PROBLEM\_ID of 2 and excluded from the further processing of the data (no such records were found).

### ***G.3.2 Deriving the active enrolment time period***

Having obtained all the information in regard to the unit standard, the derivation logic focused on deriving the active enrolment period for the unit standard enrolment record.

Two new indicators were created namely; a nominal data value and a corresponding descriptive data value used to record whether the start date of the unit standard enrolment record represented;

- the enrolment date as provided in the unit standard enrolment record,
- the unit standard enrolment record did not have an enrolment date and was as a result derived from the combination of the unit standard achievement date and the unit standard credits (see Section 3.6.4.1.a), or
- that the unit standard enrolment record did not have an enrolment date or an achievement date and was as a result derived from the derived start date of the enrolment (see Section 3.6.4.1.a).

Additionally a new data field was created to store the derived start date of the unit standard enrolment record based on the above.

Once a start date was implemented as described above, an end date for the active enrolment time period was implemented either as:

- the actual achievement date of the unit standard enrolment record, or
- a derived end date calculated using the combination of the start date for the enrolment record and the unit standard credits (see Section 3.6.4.1.b).

This resulted in the addition of the following indicators and data fields on the table DM\_USTD\_ENROL:

16. START\_DATE\_ID: A nominal data code, and

17. START\_DATE\_DESC: a corresponding descriptive data value indicating whether the value in START\_DATE represents

- a) an enrolment date (ENROL\_DATE),

- b) a derived value utilizing the achievement date (ACHIEVE\_DATE) for the enrolment record and the unit standard credits (CREDITS), or
- c) a derived value utilizing the derived start date (DERIVED\_START\_DATE) of the record.

18. START\_DATE: The start date of the active enrolment time period.

19. END\_DATE: The derived end date of the active enrolment time period, representing either the value found in ACHIEVE\_DATE or a value derived from START\_DATE and CREDITS.

A number of the semantic business rules are dependent on the active enrolment period of the record. Additionally an analysis of the unit standard data (DM\_QUAL) shows that the earliest registration for a unit standard occurred on 30 June 2000. As a direct result it could be deduced that any unit standard enrolment record with a start date less than 30 June 2000 was an outlier record. Such records by their very nature were considered erroneous and needed to be excluded from the research. These records were allocated a PROBLEM\_ID code of 3 and excluded from the further processing of the data. These records constituted 0.47% of the unit standard enrolment records initially extracted from the NLRD.

### ***G.3.3 Core data required for the development of the indicator fields and additional data values***

Having established the active enrolment time period of the unit standard enrolment record, the derivation process focused on the:

- collection of the core data required for the development of the semantic business rule indicators described in Appendix G.2, and
- the collection of additional data fields that may prove valuable to analysis of the unit standard enrolment data as described in Section 3.6.5.

The reader should note that the data received from the NLRD was, with the exception of lookup values, provided in a format that closely represents a relational database design. As an example, even though the unit standard enrolment table (DM\_USTD\_ENROL) contained a unique identifier for the unit standard (UNIT\_STANDARD\_ID), the unit standard enrolment table did not contain additional data fields that describe the unit standard, for example the NQF Level of the unit standard.

This section describes which data fields sourced from other tables were added to the unit standard enrolment table (DM\_USTD\_ENROL) and how the linkage between the unit standard enrolment record and the other tables were implemented.

Data that describes the accreditation of the ETQE was obtained from the ETQE accreditation table (DM\_ETQE) using the unique identifier of the ETQE (ETQE\_ID).

20. ETQE\_START\_DATE (START\_DATE on DM\_ETQE): The start date of the accreditation of the ETQE that submitted the enrolment record to the NLRD.

21. ETQE\_END\_DATE (END\_DATE on DM\_ETQE): The end date of the accreditation of the ETQE that submitted the enrolment record to the NLRD.

ETQE's are rarely accredited to quality assure unit standards. Rather they are accredited to quality assure qualifications. The accreditation extends to the quality assurance of unit standards that are linked to the qualification. As a result the qualification accreditations for an ETQE must be mapped to a unit standard level. This information is derived from the table DM\_ETQE\_ACCRED joined to DM\_USTD\_QUAL\_FINAL (linked by QUALIFICATION\_ID) where each possible unit standard that belongs to a qualification that an ETQE has been accredited to offer is allocated the start and end dates of the ETQE qualification accreditation. The results are saved in new table called DM\_ETQE\_ACCRED\_FINAL which has the same data structure as the table DM\_ETQE\_ACCRED.

Data that describes the accreditation of the ETQE to quality assure the unit standard was obtained from the ETQE unit standard accreditation table (DM\_ETQE\_ACCRED\_FINAL) using the unique identifier of the ETQE (ETQE\_ID) and the unique identifier of the unit standard (UNIT\_STANDARD\_ID).

22. ETQE\_ACCRED\_START\_DATE (START\_DATE on DM\_ETQE\_ACCRED): The start date of the accreditation to quality assure the unit standard of the ETQE that submitted the enrolment record to the NLRD.

23. ETQE\_ACCRED\_END\_DATE (END\_DATE on DM\_ETQE\_ACCRED): The end date of the accreditation to quality assure the unit standard of the ETQE that submitted the enrolment record to the NLRD.

Data that describes the provider and its accreditation as obtained from the provider accreditation table (DM\_PROV) using the unique identifier of the provider (PROVIDER\_ID).

24. PROV\_START\_DATE (START\_DATE on DM\_PROV): The start date of the accreditation of the provider that offered the unit standard.
25. PROV\_END\_DATE (END\_DATE on DM\_PROV): The end date of the accreditation of the provider that offered the unit standard.
26. PROV\_ETQE\_ID (ETQE\_ID on DM\_PROV): The primary ETQE of the provider.
27. PROVIDER\_TYPE\_ID and PROVIDER\_TYPE\_DESC: The data code and corresponding description of the provider type.
28. PROVIDER\_CLASS\_ID and PROVIDER\_CLASS\_DESC: The data code and corresponding description of the provider class.
29. PROV\_PROVINCE\_CODE (PROVINCE\_CODE on DM\_PROV) and PROV\_PROVINCE\_DESC (PROVINCE\_DESC on DM\_PROV): The data code and corresponding description of the province that the provider is located in.

The raw data obtained from the NLRD in regard to provider accreditations contains two scenarios in which a provider may have been accredited for the same unit standard. The provider may have been accredited for the unit standard explicitly whereby the provider accreditation record only contains a UNIT\_STANDARD\_ID value. The provider may also have been accredited for the same unit standard implicitly whereby the provider accreditation record contains a QUALIFICATION\_ID and a UNIT\_STANDARD\_ID. As a result a new table called DM\_PROV\_ACCRED\_USTD was created that contains a unique combination of PROVIDER\_ID and UNIT\_STANDARD\_ID, the minimum START\_DATE for the accreditation for the unit standard and the maximum END\_DATE for the accreditation for the unit standard.

Data that describes the accreditation of the provider to offer unit standards as obtained from the provider unit standard accreditation table (DM\_PROV\_ACCRED\_USTD) using the unique identifier of the provider (PROVIDER\_ID) and the unique identifier of the unit standard (UNIT\_STANDARD\_ID).

30. PROV\_ACCRED\_START\_DATE (START\_DATE on DM\_PROV\_ACCRED\_USTD): The start date of the accreditation to offer the unit standard of the provider that offered the unit standard.

31. PROV\_ACCRED\_END\_DATE (END\_DATE on DM\_PROV\_ACCRED\_USTD): The end date of the accreditation to offer the unit standard of the provider that offered the unit standard.

Data that describes the registration of the assessor as obtained from the assessor registration table (DM\_ASOR) using the unique identifier of the assessor (ASSESSOR\_ID).

32. ASOR\_START\_DATE (START\_DATE on DM\_ASOR): The start date of the registration of the assessor that assessed the unit standard achievement.

33. ASOR\_END\_DATE (END\_DATE on DM\_ASOR): The end date of the registration of the assessor that assessed the unit standard achievement.

The raw data obtained from the NLRD in regard to assessor registrations contains two scenarios in which an assessor may have been registered for the same unit standard. The assessor may have been registered for the unit standard explicitly whereby the assessor registration record only contains a UNIT\_STANDARD\_ID value. The assessor may also have been registered for the same unit standard implicitly whereby the assessor registration record contains a QUALIFICATION\_ID and a UNIT\_STANDARD\_ID. As a result a new table called DM\_ASOR\_REGSTR\_USTD was created that contains a unique combination of ASSESSOR\_ID and UNIT\_STANDARD\_ID, the minimum START\_DATE for the assessor's registration for the unit standard and the maximum END\_DATE for the assessor's registration for the unit standard.

Data that describes the registration of the assessor to assess a unit standard as obtained from the assessor unit standard registration table (DM\_ASOR\_REGSTR\_USTD) using the unique identifier of the assessor (ASSESSOR\_ID) and the unique identifier of the unit standard (UNIT\_STANDARD\_ID).

34. ASOR\_REGSTR\_START\_DATE (START\_DATE on DM\_ASOR\_REGSTR\_USTD): The start date of the registration to assess the unit standard of the assessor that assessed the unit standard achievement.

35. ASOR\_REGSTR\_END\_DATE (END\_DATE on DM\_ASOR\_REGSTR\_USTD): The end date of the registration to assess the unit standard of the assessor that assessed the unit standard achievement.

The date on which an ETQE submitted its first full data submission to the NLRD as obtained from the table DM\_ETQE\_START using the ETQE\_ID of the enrolment record (see Section 3.8.3.1).

36. ETQE\_FIRST\_DATE (START\_DATE on DM\_ETQE\_START): The first date on which the ETQE submitted a full submission to the NLRD.

The date of the most recent NLRD data submission cycle as obtained from the Director of the NLRD (see Section 3.8.3.2).

37. CYCLE\_DATE (variable that is set at execution of the script): The date of the most recent NLRD data submission cycle.

### G.3.4 Development of ETQE\_IND

As detailed in Appendix G.2, the development of ETQE\_IND required the implementation of an indicator that denotes whether the ETQE was accredited for the duration of the learner’s active enrolment on the unit standard. In order to make the analysis of the data more suitable for data mining purposes the indicator was implemented as a data code and corresponding description.

Figure G.3.4.1 illustrates the manner in which ETQE\_IND was developed using five example unit standard enrolment records, for an ETQE that has not been amalgamated. The figure shows how:

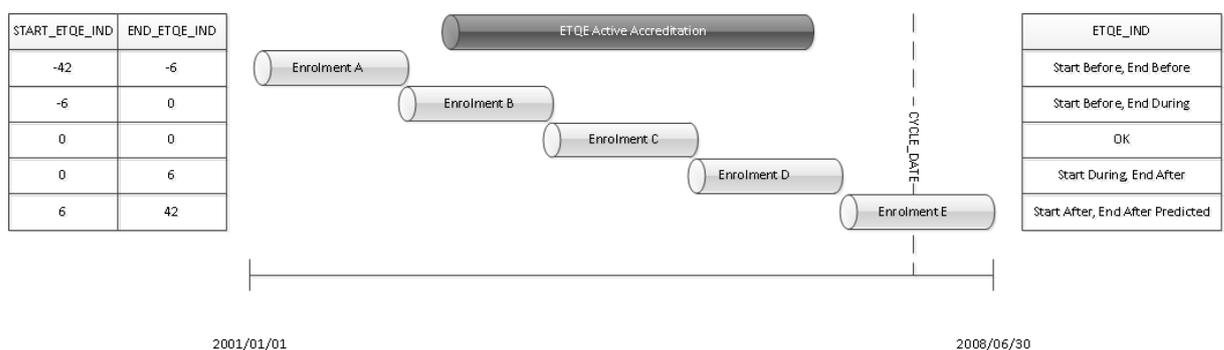


Figure G.3.4.1 Illustrative diagram of ETQE\_IND development

- a unit standard enrolment record (Enrolment A) with a start date prior to the accreditation period of the ETQE and an end date prior to the accreditation period of the ETQE is allocated an ETQE\_IND value of ‘Start Before, End Before’,

- a unit standard enrolment record (Enrolment B) with a start date prior to the accreditation period of the ETQE and an end date during the accreditation period of the ETQE is allocated an ETQE\_IND value of ‘Start Before, End During’,
- a unit standard enrolment record (Enrolment C) with a start date during the accreditation period of the ETQE and an end date during the accreditation period of the ETQE is allocated an ETQE\_IND value of ‘OK’,
- a unit standard enrolment record (Enrolment D) with a start date during the accreditation period of the ETQE and an end date after the accreditation period of the ETQE is allocated an ETQE\_IND value of ‘Start During, End After’, and
- a unit standard enrolment record (Enrolment E) with a start date after the accreditation period of the ETQE and an end date after the accreditation period of the ETQE is allocated an ETQE\_IND value of ‘Start After, End After’, and because the end of the active enrolment time period exceeds the latest data submission cycle date the word ‘Predicted’ is appended to the value.

The development of the ETQE\_IND indicator required the implementation of two additional indicators. These indicators assisted in the development of and further description of the value allocated to ETQE\_IND. Both of these two additional indicators were developed as representations of data in relation to a point in time as discussed in Section 3.6.4.4.

The first indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the start date (START\_DATE) of the active enrolment time period and the start date of the ETQE’s active accreditation time period (ETQE\_START\_DATE and ETQE\_END\_DATE), where;

- a unit standard enrolment record with a start date before the start date of the ETQE’s accreditation would be given a negative value of the number of months between these two values (for example Enrolment A on Figure G.3.4.1),
- a unit standard enrolment record with a start date that falls between the start and end dates of the ETQE’s accreditation is given a value of 0 (for example Enrolment C on Figure G.3.4.1), and

- a unit standard enrolment record with a start date that is after the end date of the ETQE's accreditation would be given a positive value of the number of months between these two values (for example Enrolment E on Figure G.3.4.1).

The second indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the end date (END\_DATE) of the active enrolment time period and the end date of the ETQE's active accreditation time period (ETQE\_START\_DATE and ETQE\_END\_DATE), where;

- a unit standard enrolment record with an end date before the start date of the ETQE's accreditation would be given a negative value of the number of months between these two values (for example Enrolment A on Figure G.3.4.1),
- a unit standard enrolment record with an end date that falls between the start and end dates of the ETQE's accreditation is given a value of 0 (for example Enrolment C on Figure G.3.4.1), and
- a unit standard enrolment record with an end date that is after the end date of the ETQE's accreditation would be given a positive value of the number of months between these two values (for example Enrolment E on Figure G.3.4.1).

This logic resulted in the addition of the following new indicators on the table DM\_USTD\_ENROL:

38. START\_ETQE\_IND: Numeric value indicating the distance between the start date of the unit standard enrolment record and the ETQE accreditation.
39. END\_ETQE\_IND: Numeric value indicating the distance between the end date of the unit standard enrolment record and the ETQE accreditation.

Using the values in these two fields in conjunction with each other it was possible to derive a code and corresponding description for ETQE\_IND by:

- Allocating a value of 'OK' to records where START\_ETQE\_IND is equal to 0 and END\_ETQE\_IND is equal to 0.

For all remaining records

- Allocating a value of 'Start Before' to records where START\_ETQE\_IND was less than 0, 'Start During' to records where START\_ETQE\_IND is equal to 0 and 'Start After' were START\_ETQE\_IND was greater than 0.

- Allocating a value 'End Before' to records where END\_ETQE\_IND was less than 0, 'End During' to records where END\_ETQE\_IND is equal to 0 and 'End After' where END\_ETQE\_IND was greater than 0.

This logic resulted in the addition of the ETQE\_IND indicator code and corresponding description on the table DM\_USTD\_ENROL.

40. ETQE\_IND\_ID and ETQE\_IND\_DESC: Code and corresponding description denoting a record's compliance in regard to whether the ETQE was accredited for the duration of the learner's active enrolment on the unit standard.

The above mentioned logic however did not take into consideration the accreditation of the ETQE that was also accredited to quality assure the unit standard in situations where ETQEs had been amalgamated (see Section 3.8.3.3). In order to address this issue, the logic determined whether a different ETQE had in the past been accredited to quality assure the unit standard (DM\_ETQE\_ACCRED).

The ETQE identifier, start date and end date of the ETQE that was also accredited to quality assure the unit standard was amended to the table DM\_USTD\_ENROL:

41. OTHR\_ETQE\_ID: The ETQE identifier of the ETQE.
42. OTHR\_ETQE\_START\_DATE (START\_DATE on DM\_ETQE): The start date of the accreditation of the ETQE.
43. OTHR\_ETQE\_END\_DATE (END\_DATE on DM\_ETQE): The end date of the accreditation of the ETQE.

Four indicators were developed in the same manner as described for START\_ETQE\_IND, END\_ETQE\_IND, ETQE\_IND\_ID and ETQE\_IND\_DESC, using the indicators OTHR\_ETQE\_START\_DATE, OTHR\_ETQE\_END\_DATE, OTHR\_START\_ETQE\_IND and OTHR\_END\_ETQE\_IND in place of the indicators ETQE\_START\_DATE, ETQE\_END\_DATE, START\_ETQE\_IND and END\_ETQE\_IND.

This logic resulted in the addition of the following new indicators on the table DM\_USTD\_ENROL:

44. OTHR\_START\_ETQE\_IND: Numeric value indicating the distance between the start date of the unit standard enrolment record and the other ETQE's accreditation.

45. OTHR\_END\_ETQE\_IND: Numeric value indicating the distance between the end date of the unit standard enrolment record and the other ETQE's accreditation.
46. OTHR\_ETQE\_IND\_ID and OTHR\_ETQE\_IND\_DESC: Code and corresponding description denoting a record's compliance in regard to whether the other ETQE was accredited for the duration of the learner's active enrolment on the unit standard.

The results of both the ETQE\_IND\_ID and ETQE\_IND\_DESC fields and the OTHR\_ETQE\_IND\_ID and OTHR\_ETQE\_IND\_DESC fields were then consolidated into the ETQE\_IND indicators in the following manner:

- Any record that was found to be compliant based on the value stored in ETQE\_IND\_ID or OTHR\_ETQE\_IND\_ID was marked as compliant.
- Any record that was found to be non-compliant based on both the value stored in ETQE\_IND\_ID and OTHR\_ETQE\_IND\_ID was provided a modified code and corresponding description that show the results of the compliance result of both ETQE\_IND\_ID and OTHR\_ETQE\_IND\_ID.

The final derivation step entailed amending the ETQE\_IND data code and corresponding description to differentiate records with a calculated end date that is greater than the latest data submission cycle date from other records (see Section 3.8.3.2). As a result the data code was amended and the word 'Predicted' was appended to the ETQE\_IND indicator description for any records with an END\_DATE value greater than CYCLE\_DATE (for example Enrolment E on Figure G.3.4.1).

### ***G.3.5 Development of ETQE\_ACCRED\_IND***

As detailed in Appendix G.2, the development of ETQE\_ACCRED\_IND required the implementation of an indicator that denotes whether the ETQE was accredited to quality assure the unit standard for the duration of the learner's active enrolment on the unit standard. In order to make the analysis of the data more suitable for data mining purposes the indicator was implemented as a data code and corresponding description.

Figure G.3.5.1 illustrates the manner in which ETQE\_ACCRED\_IND was developed using five example unit standard enrolment records, for an ETQE that has not been amalgamated. The figure shows how:

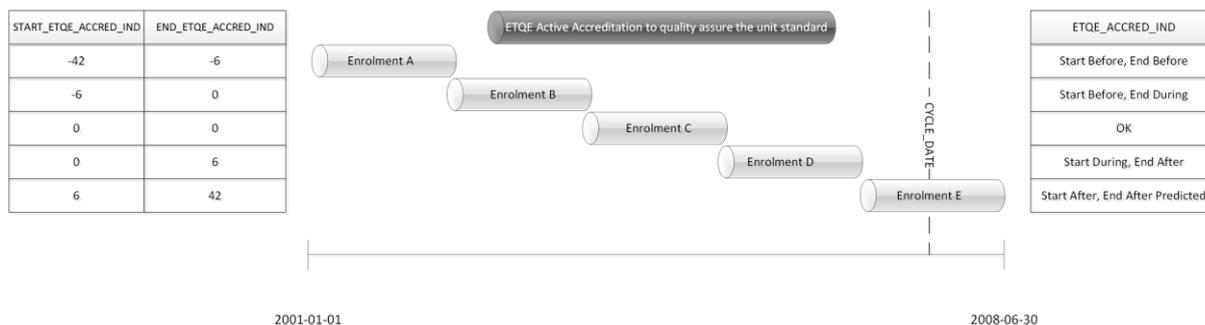


Figure G.3.5.1 Illustrative diagram of ETQE\_ACCRED\_IND development

- a unit standard enrolment record (Enrolment A) with a start date prior to the unit standard accreditation period of the ETQE and an end date prior to the unit standard accreditation period of the ETQE is allocated an ETQE\_ACCRED\_IND value of ‘Start Before, End Before’,
- a unit standard enrolment record (Enrolment B) with a start date prior to the unit standard accreditation period of the ETQE and an end date during the unit standard accreditation period of the ETQE is allocated an ETQE\_ACCRED\_IND value of ‘Start Before, End During’,
- a unit standard enrolment record (Enrolment C) with a start date during the unit standard accreditation period of the ETQE and an end date during the unit standard accreditation period of the ETQE is allocated an ETQE\_ACCRED\_IND value of ‘OK’,
- a unit standard enrolment record (Enrolment D) with a start date during the unit standard accreditation period of the ETQE and an end date after the unit standard accreditation period of the ETQE is allocated an ETQE\_ACCRED\_IND value of ‘Start During, End After’, and
- a unit standard enrolment record (Enrolment E) with a start date after the unit standard accreditation period of the ETQE and an end date after the unit standard accreditation period of the ETQE is allocated an ETQE\_ACCRED\_IND value of ‘Start After, End After’, and because the end of the active enrolment time period exceeds the latest data submission cycle date the word ‘Predicted’ is appended to the value.

The development of the ETQE\_ACCRED\_IND indicator required the implementation of two additional indicators. These indicators assisted in the development of and further description of the value allocated to ETQE\_ACCRED\_IND. Both of these two additional

indicators were developed as representations of data in relation to a point in time as discussed in Section 3.6.4.4.

The first indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the start date (START\_DATE) of the active enrolment time period and the start date of the ETQE's active accreditation to quality assure the unit standard time period (ETQE\_ACCRED\_START\_DATE and ETQE\_ACCRED\_END\_DATE), where;

- a unit standard enrolment record with a start date before the start date of the ETQE's accreditation to quality assure the unit standard would be given a negative value of the number of months between these two values (for example Enrolment A on Figure G.3.5.1),
- a unit standard enrolment record with a start date that falls between the start and end dates of the ETQE's accreditation to quality assure the unit standard is given a value of 0 (for example Enrolment C on Figure G.3.5.1), and
- a unit standard enrolment record with a start date that is after the end date of the ETQE's accreditation to quality assure the unit standard would be given a positive value of the number of months between these two values (for example Enrolment E on Figure G.3.5.1).

The second indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the end date (END\_DATE) of the active enrolment time period and the end date of the ETQE's active accreditation to quality assure the unit standard time period (ETQE\_ACCRED\_START\_DATE and ETQE\_ACCRED\_END\_DATE), where;

- a unit standard enrolment record with an end date before the start date of the ETQE's accreditation to quality assure the unit standard would be given a negative value of the number of months between these two values (for example Enrolment A on Figure G.3.5.1),
- a unit standard enrolment record with an end date that falls between the start and end dates of the ETQE's accreditation to quality assure the unit standard is given a value of 0 (for example Enrolment C on Figure G.3.5.1), and
- a unit standard enrolment record with an end date that is after the end date of the ETQE's accreditation to quality assure the unit standard would be given a positive

value of the number of months between these two values (for example Enrolment E on Figure G.3.5.1).

This logic resulted in the addition of the following new indicators on the table DM\_USTD\_ENROL:

47. START\_ETQE\_ACCRED\_IND: Numeric value indicating the distance between the start date of the unit standard enrolment record and the ETQE accreditation to quality assure the unit standard.
48. END\_ETQE\_ACCRED\_IND: Numeric value indicating the distance between the end date of the unit standard enrolment record and the ETQE accreditation to quality assure the unit standard.

Using the values in these two fields in conjunction with each other it was possible to derive a code and corresponding description for ETQE\_ACCRED\_IND by:

- Allocating a value of 'OK' to records where START\_ETQE\_ACCRED\_IND is equal to 0 and END\_ETQE\_ACCRED\_IND is equal to 0.

For all remaining records

- Allocating a value of 'Start Before' to records where START\_ETQE\_ACCRED\_IND was less than 0, 'Start During' to records where START\_ETQE\_ACCRED\_IND is equal to 0 and 'Start After' were START\_ETQE\_ACCRED\_IND was greater than 0.
- Allocating a value 'End Before' to records where END\_ETQE\_ACCRED\_IND was less than 0, 'End During' to records where END\_ETQE\_ACCRED\_IND is equal to 0 and 'End After' where END\_ETQE\_ACCRED\_IND was greater than 0.

This logic resulted in the addition of the ETQE\_ACCRED\_IND indicator code and corresponding description on the table DM\_USTD\_ENROL.

49. ETQE\_ACCRED\_IND\_ID and ETQE\_ACCRED\_IND\_DESC: Code and corresponding description denoting a record's compliance in regard to whether the ETQE was accredited to quality assure the unit standard for the duration of the learner's active enrolment on the unit standard.

The above mentioned logic however did not take into consideration the accreditation to quality assure the unit standard of the ETQE that was also accredited to quality assure the unit standard in situations where ETQEs had been amalgamated (see Section 3.8.3.3). In

order to address this issue, the logic determined whether a different ETQE had in the past been accredited to quality assure the unit standard (DM\_ETQE\_ACCRED).

The ETQE identifier, start date and end date of the accreditation to quality assure the unit standard of the ETQE that was also accredited to quality assure the unit standard was amended to the table DM\_USTD\_ENROL:

50. OTHR\_ETQE\_ACCRED\_ID: The ETQE identifier of the ETQE.

51. OTHR\_ETQE\_ACCRED\_START\_DATE (START\_DATE on DM\_ETQE\_ACCRED):

The start date of the ETQE accreditation to quality assure the unit standard for the ETQE.

52. OTHR\_ETQE\_ACCRED\_END\_DATE (END\_DATE on DM\_ETQE\_ACCRED): The end date of the accreditation to quality assure the unit standard for the ETQE.

Four indicators were developed in the same manner as described for START\_ETQE\_ACCRED\_IND, END\_ETQE\_ACCRED\_IND, ETQE\_ACCRED\_IND\_ID and ETQE\_ACCRED\_IND\_DESC, using the indicators OTHR\_ETQE\_ACCRED\_START\_DATE, OTHR\_ETQE\_ACCRED\_END\_DATE, OTHR\_START\_ETQE\_ACCRED\_IND and OTHR\_END\_ETQE\_ACCRED\_IND in place of the indicators ETQE\_START\_DATE, ETQE\_END\_DATE, START\_ETQE\_ACCRED\_IND and END\_ETQE\_ACCRED\_IND.

This logic resulted in the addition of the following new indicators on the table DM\_USTD\_ENROL:

53. OTHR\_START\_ETQE\_ACCRED\_IND: Numeric value indicating the distance between the start date of the unit standard enrolment record and the other ETQE's accreditation to quality assure the unit standard.

54. OTHR\_END\_ETQE\_ACCRED\_IND: Numeric value indicating the distance between the end date of the unit standard enrolment record and the other ETQE's accreditation to quality assure the unit standard.

55. OTHR\_ETQE\_ACCRED\_IND\_ID and OTHR\_ETQE\_ACCRED\_IND\_DESC: Code and corresponding description denoting a record's compliance in regard to whether the other ETQE was accredited to quality assure the unit standard for the duration of the learner's active enrolment on the unit standard.

The results of both the ETQE\_ACCRED\_IND\_ID and ETQE\_ACCRED\_IND\_DESC fields and the OTHR\_ETQE\_ACCRED\_IND\_ID and OTHR\_ETQE\_ACCRED\_IND\_DESC fields were then consolidated into the ETQE\_ACCRED\_IND indicators in the following manner:

- Any record that was found to be compliant based on the value stored in ETQE\_ACCRED\_IND\_ID or OTHR\_ETQE\_ACCRED\_IND\_ID was marked as compliant.
- Any record that was found to be non-compliant based on both the value stored in ETQE\_ACCRED\_IND\_ID and OTHR\_ETQE\_ACCRED\_IND\_ID was provided a modified code and corresponding description that show the results of the compliance result of both ETQE\_ACCRED\_IND\_ID and OTHR\_ETQE\_ACCRED\_IND\_ID.

The final derivation step entailed amending the ETQE\_ACCRED\_IND data code and corresponding description to differentiate records with a calculated end date that is greater than the latest data submission cycle date from other records (see Section 3.8.3.2). As a result the data code was amended and the word ‘Predicted’ was appended to the ETQE\_ACCRED\_IND indicator description for any records with an END\_DATE value greater than CYCLE\_DATE (for example Enrolment E on Figure G.3.5.1).

### ***G.3.6 Development of PROV\_IND***

As detailed in Appendix G.2, the development of PROV\_IND required the implementation of an indicator that denotes whether the provider was accredited for the duration of the learner’s active enrolment on the unit standard. In order to make the analysis of the data more suitable for data mining purposes the indicator was implemented as a data code and corresponding description.

Figure G.3.6.1 illustrates the manner in which PROV\_IND was developed using five example unit standard enrolment records, for a provider that was accredited and is not an ‘ETQE Provider’. The figure shows how:

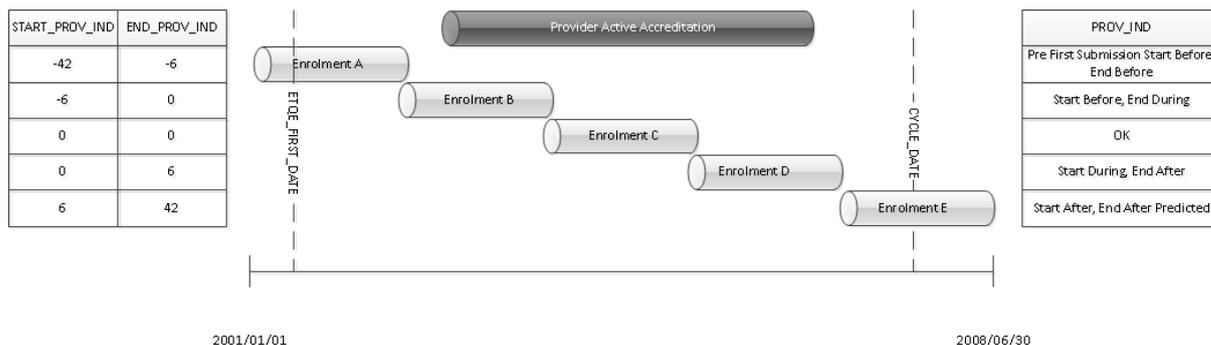


Figure G.3.6.1 Illustrative diagram of PROV\_IND development

- a unit standard enrolment record (Enrolment A) with a start date prior to the provider's accreditation period and an end date prior to the provider's accreditation period is allocated a PROV\_IND value of 'Start Before, End Before'. The start of the active enrolment time period precedes the date on which the ETQE submitted its first full data submission to the NLRD, as a result the words 'Pre First Submission' are appended to the value,
- a unit standard enrolment record (Enrolment B) with a start date prior to the accreditation period of the provider and an end date during the accreditation period of the provider is allocated a PROV\_IND value of 'Start Before, End During',
- a unit standard enrolment record (Enrolment C) with a start date during the accreditation period of the provider and an end date during the accreditation period of the provider is allocated a PROV\_IND value of 'OK',
- a unit standard enrolment record (Enrolment D) with a start date during the accreditation period of the provider and an end date after the accreditation period of the provider is allocated a PROV\_IND value of 'Start During, End After', and
- a unit standard enrolment record (Enrolment E) with a start date after the accreditation period of the provider and an end date after the accreditation period of the provider is allocated a PROV\_IND value of 'Start After, End After'. The end of the active enrolment time period exceeds the latest data submission cycle date, as a result the word 'Predicted' is appended to the value.

The development of the PROV\_IND indicator required the implementation of two additional indicators. These indicators assisted in the development of and further description of the value allocated to PROV\_IND. Both of these two additional indicators

were developed as representations of data in relation to a point in time as discussed in Section 3.6.4.4.

The first indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the start date (START\_DATE) of the active enrolment time period and the start date of the provider's active accreditation time period (PROV\_START\_DATE and PROV\_END\_DATE), where;

- a unit standard enrolment record with a start date before the start date of the provider's accreditation would be given a negative value of the number of months between these two values (for example Enrolment A on Figure G.3.6.1),
- a unit standard enrolment record with a start date that falls between the start and end dates of the provider's accreditation is given a value of 0 (for example Enrolment C on Figure G.3.6.1), and
- a unit standard enrolment record with a start date that is after the end date of the provider's accreditation would be given a positive value of the number of months between these two values (for example Enrolment E on Figure G.3.6.1).

The second indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the end date (END\_DATE) of the active enrolment time period and the end date of the provider's active accreditation time period (PROV\_START\_DATE and PROV\_END\_DATE)), where;

- a unit standard enrolment record with an end date before the start date of the provider's accreditation would be given a negative value of the number of months between these two values (for example Enrolment A on Figure G.3.6.1),
- a unit standard enrolment record with an end date that falls between the start and end dates of the provider's accreditation is given a value of 0 (for example Enrolment C on Figure G.3.6.1), and
- a unit standard enrolment record with an end date that is after the end date of the provider's accreditation would be given a positive value of the number of months between these two values (for example Enrolment E on Figure G.3.6.1).

This logic resulted in the addition of the following new indicators on the table DM\_USTD\_ENROL:

56. START\_PROV\_IND: Numeric value indicating the distance between the start date of the unit standard enrolment record and the provider accreditation.

57. END\_PROV\_IND: Numeric value indicating the distance between the end date of the unit standard enrolment record and the provider accreditation.

Using the values in these two fields in conjunction with each other it was possible to derive a code and corresponding description for PROV\_IND as follows:

- Where a provider is an 'ETQE provider' (see Section 3.8.3.5), allocating a value of 'ETQE Provider'
- Where a provider accreditation did not exist, allocating a value of 'No Accreditation' to the record.
- Allocating a value of 'OK' to records where START\_PROV\_IND is equal to 0 and END\_PROV\_IND is equal to 0.

For all remaining records:

- Allocating a value of 'Start Before' to records where START\_PROV\_IND was less than 0, 'Start During' to records where START\_PROV\_IND is equal to 0 and 'Start After' where START\_PROV\_IND was greater than 0.
- Allocating a value 'End Before' to records where END\_PROV\_IND was less than 0, 'End During' to records where END\_PROV\_IND is equal to 0 and 'End After' where END\_PROV\_IND was greater than 0.

The final derivation steps included:

- Amending the data code and appending the words 'Pre First Submission' to the indicator value for any records with a START\_DATE value less than PROV\_ETQE\_FIRST\_DATE (for example Enrolment A on Figure G.3.6.1). In other words all records where the learner enrolled on the unit standard prior to the first full data submission from the primary ETQE of the provider to the NLRD (see Section 3.8.3.1 and Section 3.8.3.5).
- Amending the data code and appending the word 'Predicted' to the indicator value for any records with a END\_DATE value greater than CYCLE\_DATE (for example Enrolment E on Figure G.3.6.1). In other words all records with a calculated end date that is greater than the latest data submission cycle date (see Section 3.8.3.2).

This logic resulted in the addition of the PROV\_IND indicator code and corresponding description on the table DM\_USTD\_ENROL.

58. PROV\_IND\_ID and PROV\_IND\_DESC: Code and corresponding description denoting a record’s compliance in regard to whether the provider was accredited for the duration of the learner’s active enrolment on the unit standard.

### G.3.7 Development of PROV\_ACCRED\_IND

As detailed in Appendix G.2, the development of PROV\_ACCRED\_IND required the implementation of an indicator that denotes whether the provider was accredited to offer the unit standard for the duration of the learner’s active enrolment on the unit standard. In order to make the analysis of the data more suitable for data mining purposes the indicator was implemented as a data code and corresponding description.

Figure G.3.7.1 illustrates the manner in which PROV\_ACCRED\_IND was developed using five example unit standard enrolment records, for a provider that was accredited to offer the unit standard and is not an ‘ETQE Provider’. The figure shows how:

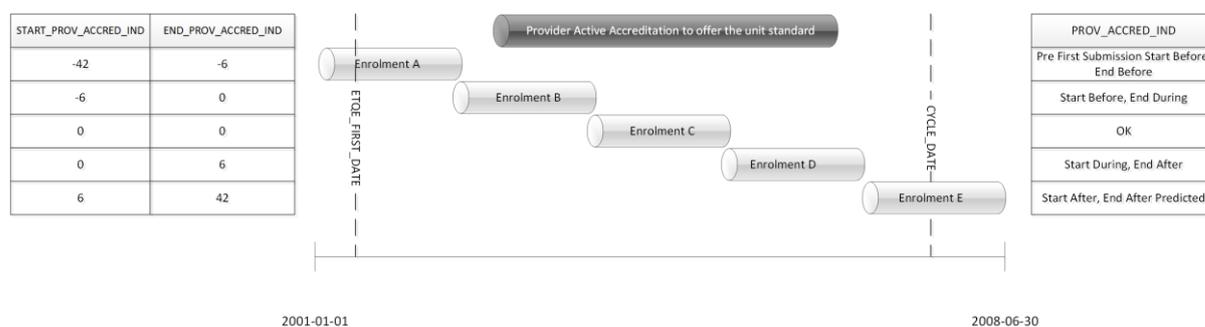


Figure G.3.7.1 Illustrative diagram of PROV\_ACCRED\_IND development

- a unit standard enrolment record (Enrolment A) with a start date prior to the unit standard accreditation of the provider and an end date prior to the unit standard accreditation of the provider is allocated a PROV\_ACCRED\_IND value of ‘Start Before, End Before’. The start of the active enrolment time period precedes the date on which the ETQE submitted its first full data submission to the NLRD, as a result the words ‘Pre First Submission’ are appended to the value,
- a unit standard enrolment record (Enrolment B) with a start date prior to the unit standard accreditation period of the provider and an end date during the unit standard

accreditation period of the provider is allocated a PROV\_ACCRED\_IND value of ‘Start Before, End During’,

- a unit standard enrolment record (Enrolment C) with a start date during the unit standard accreditation period of the provider and an end date during the unit standard accreditation period of the provider is allocated a PROV\_ACCRED\_IND value of ‘OK’,
- a unit standard enrolment record (Enrolment D) with a start date during the unit standard accreditation period of the provider and an end date after the unit standard accreditation period of the provider is allocated a PROV\_ACCRED\_IND value of ‘Start During, End After’, and
- a unit standard enrolment record (Enrolment E) with a start date after the unit standard accreditation period of the provider and an end date after the unit standard accreditation period of the provider is allocated a PROV\_ACCRED\_IND value of ‘Start After, End After’. The end of the active enrolment time period exceeds the latest data submission cycle date, as a result the word ‘Predicted’ is appended to the value.

The development of the PROV\_ACCRED\_IND indicator required the implementation of two additional indicators. These indicators assisted in the development of and further description of the value allocated to PROV\_ACCRED\_IND. Both of these two additional indicators were developed as representations of data in relation to a point in time as discussed in Section 3.6.4.4.

The first indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the start date (START\_DATE) of the active enrolment time period and the start date of the provider’s active accreditation to offer the unit standard time period (PROV\_ACCRED\_START\_DATE and PROV\_ACCRED\_END\_DATE), where;

- a unit standard enrolment record with a start date before the start date of the provider’s accreditation to offer the unit standard would be given a negative value of the number of months between these two values (for example Enrolment A on Figure G.3.7.1),
- a unit standard enrolment record with a start date that falls between the start and end dates of the provider’s accreditation to offer the unit standard is given a value of 0 (for example Enrolment C on Figure G.3.7.1), and

- a unit standard enrolment record with a start date that is after the end date of the provider's accreditation to offer the unit standard would be given a positive value of the number of months between these two values (for example Enrolment E on Figure G.3.7.1).

The second indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the end date (END\_DATE) of the active enrolment time period and the end date of the provider's active accreditation to offer the unit standard time period (PROV\_ACCRED\_START\_DATE and PROV\_ACCRED\_END\_DATE)), where;

- a unit standard enrolment record with an end date before the start date of the provider's accreditation to offer the unit standard would be given a negative value of the number of months between these two values (for example Enrolment A on Figure G.3.7.1),
- a unit standard enrolment record with an end date that falls between the start and end dates of the provider's accreditation to offer the unit standard is given a value of 0 (for example Enrolment C on Figure G.3.7.1), and
- a unit standard enrolment record with an end date that is after the end date of the provider's accreditation to offer the unit standard would be given a positive value of the number of months between these two values (for example Enrolment E on Figure G.3.7.1).

This logic resulted in the addition of the following new indicators on the table DM\_USTD\_ENROL:

59. START\_PROV\_ACCRED\_IND: Numeric value indicating the distance between the start date of the unit standard enrolment record and the provider accreditation to offer the unit standard.
60. END\_PROV\_ACCRED\_IND: Numeric value indicating the distance between the end date of the unit standard enrolment record and the provider accreditation to offer the unit standard.

Using the values in these two fields in conjunction with each other it was possible to derive a code and corresponding description for PROV\_ACCRED\_IND as follows:

- Where a provider is an ‘ETQE provider’ (see Section 3.8.3.5), allocating a value of 'ETQE Provider'
- Where a provider accreditation to offer the unit standard did not exist, allocating a value of 'No Accreditation' to the record.
- Allocating a value of ‘OK’ to records where START\_PROV\_ACCRED\_IND is equal to 0 and END\_PROV\_ACCRED\_IND is equal to 0.

For all remaining records:

- Allocating a value of ‘Start Before’ to records where START\_PROV\_ACCRED\_IND was less than 0, ‘Start During’ to records where START\_PROV\_ACCRED\_IND is equal to 0 and ‘Start After’ where START\_PROV\_ACCRED\_IND was greater than 0.
- Allocating a value ‘End Before’ to records where END\_PROV\_ACCRED\_IND was less than 0, ‘End During’ to records where END\_PROV\_ACCRED\_IND is equal to 0 and ‘End After’ where END\_PROV\_ACCRED\_IND was greater than 0.

The final derivation steps included:

- Amending the data code and appending the words 'Pre First Submission' to the indicator value for any records with a START\_DATE value less than PROV\_ACCRED\_ETQE\_FIRST\_DATE (for example Enrolment A on Figure G.3.7.1). In other words all records where the learner enrolled on the unit standard prior to the first full data submission from the primary ETQE of the provider to the NLRD (see Section 3.8.3.1 and Section 3.8.3.5).
- Amending the data code and appending the word ‘Predicted’ to the indicator value for any records with a END\_DATE value greater than CYCLE\_DATE (for example Enrolment E on Figure G.3.7.1). In other words all records with a calculated end date that is greater than the latest data submission cycle date (see Section 3.8.3.2).

This logic resulted in the addition of the PROV\_ACCRED\_IND indicator code and corresponding description on the table DM\_USTD\_ENROL.

61. PROV\_ACCRED\_IND\_ID and PROV\_ACCRED\_IND\_DESC: Code and corresponding description denoting a record’s compliance in regard to whether the provider was accredited to offer the unit standard for the duration of the learner’s active enrolment on the unit standard.

### ***G.3.8 Development of ASOR\_IND***

As detailed in Appendix G.2, the development of ASOR\_IND required the implementation of an indicator that denotes whether the assessor was registered at the time of the achievement of the unit standard. In order to make the analysis of the data more suitable for data mining purposes the indicator was implemented as a data code and corresponding description.

Figure G.3.8.1 illustrates the manner in which ASOR\_IND was developed using four example achieved unit standard enrolment records, for an assessor that was registered. The figure shows how:

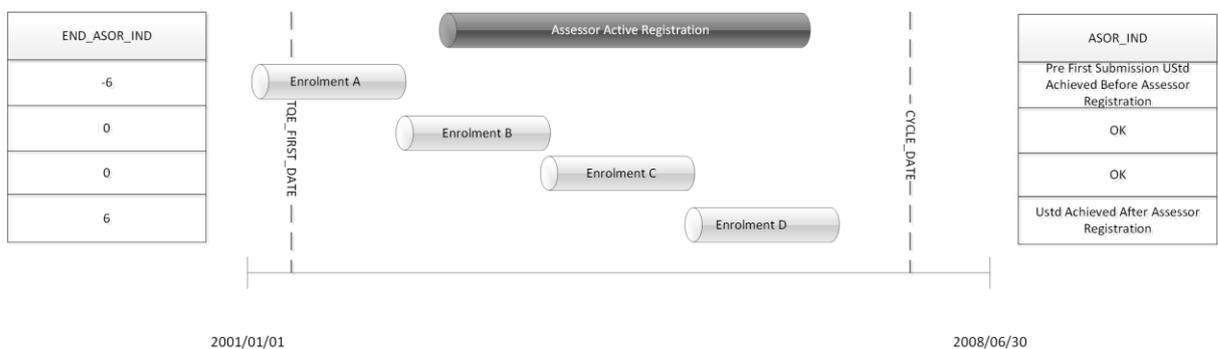


Figure G.3.8.1 Illustrative diagram of ASOR\_IND development

- a unit standard enrolment record (Enrolment A) with an end date prior to the registration period of the assessor is allocated an ASOR\_IND value of 'Qual Achieved Before Assessor Registration'. The start of the active enrolment time period precedes the date on which the ETQE submitted its first full data submission to the NLRD, as a result the words 'Pre First Submission' are appended to the value,
- a unit standard enrolment record (Enrolment B) with an end date during the registration period of the assessor is allocated an ASOR\_IND value of 'OK',
- a unit standard enrolment record (Enrolment C) end date during the registration period of the assessor is allocated an ASOR\_IND value of 'OK', and
- a unit standard enrolment record (Enrolment D) with an end date after the registration period of the assessor is allocated an ASOR\_IND value of 'Qual Achieved After Assessor Registration'.

The development of the ASOR\_IND indicator required the implementation of one additional indicator. This indicator assisted in the development of and further description of

the value allocated to ASOR\_IND. This additional indicator was developed as a representation of data in relation to a point in time as discussed in Section 3.6.4.4.

The indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the end date (END\_DATE) of the active enrolment time period and assessor's active registration time period (ASOR\_START\_DATE and ASOR\_END\_DATE), where;

- a unit standard enrolment record with an end date before the start date of the assessor's registration would be given a negative value of the number of months between these two values (for example Enrolment A on Figure G.3.8.1),
- a unit standard enrolment record with an end date that falls between the start and end dates of the assessor's registration is given a value of 0 (for example Enrolment C on Figure G.3.8.1), and
- a unit standard enrolment record with an end date that is after the end date of the assessor's registration would be given a positive value of the number of months between these two values (for example Enrolment D on Figure G.3.8.1).

This logic resulted in the addition of the following new indicator on the table DM\_USTD\_ENROL:

62. END\_ASOR\_IND: Numeric value indicating the distance between the end date of the unit standard enrolment record and the assessor registration.

Using the values in this field it was possible to derive a code and corresponding description for ASOR\_IND as follows:

- Where the unit standard enrolment had not been achieved, allocating a value of 'Not Achieved'.
- Where the unit standard enrolment had been achieved but an assessor identifier had not been provided, allocation a value of 'No Assessor Provided'.
- Where an assessor identifier had been provided but an assessor registration did not exist, allocating a value of 'No Registration' to the record.
- Allocating a value of 'OK' to records where END\_ASOR\_IND is equal to 0.
- Allocating a value of 'Qual Achieved Before Assessor Registration' where END\_ASOR\_IND is less than 0.

- Allocating a value of 'Qual Achieved After Assessor Registration' where END\_ASOR\_IND is greater than 0.

The final derivation steps included:

- Amending the data code and appending the words 'Pre First Submission' to the indicator value for any records with a START\_DATE value less than ETQE\_FIRST\_DATE (for example Enrolment A on Figure G.3.8.1). In other words all records where the learner enrolled on the unit standard prior to the first full data submission from the ETQE to the NLRD (see Section 3.8.3.1).

This logic resulted in the addition of the ASOR\_IND indicator code and corresponding description on the table DM\_USTD\_ENROL.

63. ASOR\_IND\_ID and ASOR\_IND\_DESC: Code and corresponding description denoting a record's compliance in regard to whether the assessor was registered at the time of the achievement of the unit standard.

### ***G.3.9 Development of ASOR\_REGSTR\_IND***

As detailed in Appendix G.2, the development of ASOR\_REGSTR\_IND required the implementation of an indicator that denotes whether the assessor was registered to assess the unit standard at the time of the achievement of the unit standard. In order to make the analysis of the data more suitable for data mining purposes the indicator was implemented as a data code and corresponding description.

Figure G.3.9.1 illustrates the manner in which ASOR\_REGSTR\_IND was developed using four example achieved unit standard enrolment records, for an assessor that was registered.

The figure shows how:

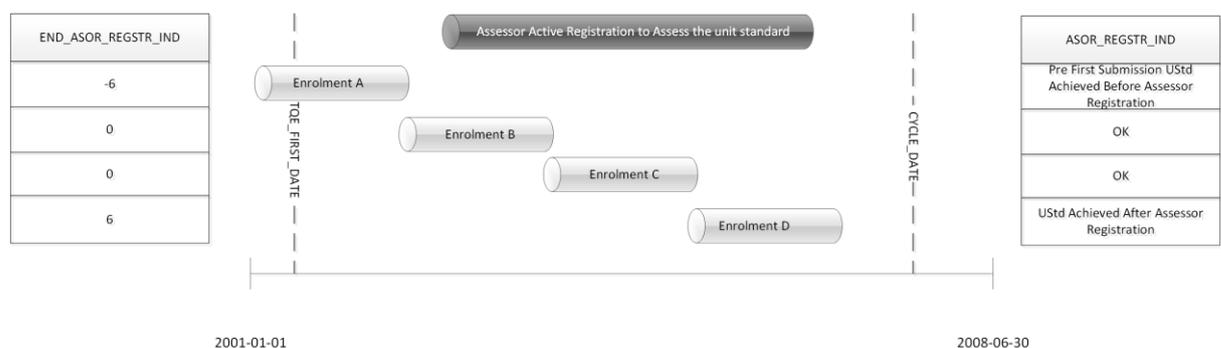


Figure G.3.9.1 Illustrative diagram of ASOR\_REGSTR\_IND development

- a unit standard enrolment record (Enrolment A) with an end date prior to the unit standard registration period of the assessor is allocated an ASOR\_REGSTR\_IND value of 'Qual Achieved Before Assessor Registration'. The start of the active enrolment time period precedes the date on which the ETQE submitted its first full data submission to the NLRD, as a result the words 'Pre First Submission' are appended to the value,
- a unit standard enrolment record (Enrolment B) with an end date during the unit standard registration period of the assessor is allocated an ASOR\_REGSTR\_IND value of 'OK',
- a unit standard enrolment record (Enrolment C) end date during the unit standard registration period of the assessor is allocated an ASOR\_REGSTR\_IND value of 'OK', and
- a unit standard enrolment record (Enrolment D) with an end date after the unit standard registration period of the assessor is allocated an ASOR\_REGSTR\_IND value of 'Qual Achieved After Assessor Registration'.

The development of the ASOR\_REGSTR\_IND indicator required the implementation of one additional indicator. This indicator assisted in the development of and further description of the value allocated to ASOR\_REGSTR\_IND. This additional indicator was developed as a representation of data in relation to a point in time as discussed in Section 3.6.4.4.

The indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the end date (END\_DATE) of the active enrolment time period and assessor's active registration to assess the unit standard time period (ASOR\_REGSTR\_START\_DATE and ASOR\_REGSTR\_END\_DATE), where;

- a unit standard enrolment record with an end date before the start date of the assessor's registration to assess the unit standard would be given a negative value of the number of months between these two values (for example Enrolment A on Figure G.3.9.1),

- a unit standard enrolment record with an end date that falls between the start and end dates of the assessor's registration to assess the unit standard is given a value of 0 (for example Enrolment C on Figure G.3.9.1), and
- a unit standard enrolment record with an end date that is after the end date of the assessor's registration to assess the unit standard would be given a positive value of the number of months between these two values (for example Enrolment D on Figure G.3.9.1).

This logic resulted in the addition of the following new indicator on the table DM\_USTD\_ENROL:

64. END\_ASOR\_REGSTR\_IND: Numeric value indicating the distance between the end date of the unit standard enrolment record and the assessor registration to assess the unit standard.

Using the values in this field it was possible to derive a code and corresponding description for ASOR\_REGSTR\_IND as follows:

- Where the unit standard enrolment had not been achieved, allocating a value of 'Not Achieved'.
- Where the unit standard enrolment had been achieved but an assessor identifier had not been provided, allocation a value of 'No Assessor Provided'.
- Where an assessor identifier had been provided but an assessor registration to assess the unit standard did not exist, allocating a value of 'No Registration' to the record.
- Allocating a value of 'OK' to records where END\_ASOR\_REGSTR\_IND is equal to 0.
- Allocating a value of 'Qual Achieved Before Assessor Registration' where END\_ASOR\_REGSTR\_IND is less than 0.
- Allocating a value of 'Qual Achieved After Assessor Registration' where END\_ASOR\_REGSTR\_IND is greater than 0.

The final derivation steps included:

- Amending the data code and appending the words 'Pre First Submission' to the indicator value for any records with a START\_DATE value less than ETQE\_FIRST\_DATE (for example Enrolment A on Figure G.3.9.1). In other words

all records where the learner enrolled on the unit standard prior to the first full data submission from the ETQE to the NLRD (see Section 3.8.3.1).

This logic resulted in the addition of the ASOR\_REGSTR\_IND indicator code and corresponding description on the table DM\_USTD\_ENROL.

65. ASOR\_REGSTR\_IND\_ID and ASOR\_REGSTR\_IND\_DESC: Code and corresponding description denoting a record’s compliance in regard to whether the assessor was registered to assess the unit standard at the time of the achievement of the unit standard.

### G.3.10 Development of USTD\_REGSTR\_IND

As detailed in Appendix G.2, the development of USTD\_REGSTR\_IND required the implementation of an indicator that denotes whether the unit standard was registered for the duration of the learner’s active enrolment on the unit standard. In order to make the analysis of the data more suitable for data mining purposes the indicator was implemented as a data code and corresponding description.

Figure G.3.10.1 illustrates the manner in which USTD\_REGSTR\_IND was developed using five example unit standard enrolment records, for a unit standard that was registered. The figure shows how:



Figure G.3.10.1 Illustrative diagram of USTD\_REGSTR\_IND development

- a unit standard enrolment record (Enrolment A) with a start date prior to the registration of the unit standard registration and an end date prior to the registration of the unit standard is allocated a USTD\_REGSTR\_IND value of ‘Start Before, End Before’,

- a unit standard enrolment record (Enrolment B) with a start date prior to the registration of the unit standard and an end date that is during the registration of the unit standard is allocated a USTD\_REGSTR\_IND value of ‘Start Before, End During’,
- a unit standard enrolment record (Enrolment C) with a start date that is during the registration of the unit standard and an end date that is during the registration of the unit standard is allocated a USTD\_REGSTR\_IND value of ‘OK’,
- a unit standard enrolment record (Enrolment D) with a start date that is during the registration of the unit standard and an end date that is after the registration of the unit standard is allocated a USTD\_REGSTR\_IND value of ‘Start During, End After’, and
- a unit standard enrolment record (Enrolment E) with a start date that is after the registration of the unit standard and an end date that is after the registration of the unit standard is allocated a USTD\_REGSTR\_IND value of ‘Start After, End After’, and because the end of the active enrolment time period exceeds the latest data submission cycle date the word ‘Predicted’ is appended to the value.

The development of the USTD\_REGSTR\_IND indicator required the implementation of two additional indicators. These indicators assisted in the development of and further description of the value allocated to USTD\_REGSTR\_IND. Both of these two additional indicators were developed as representations of data in relation to a point in time as discussed in Section 3.6.4.4.

The first indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the start date (START\_DATE) of the active enrolment time period and the start date of the unit standard’s registration and the last date on which a learner may enrol on a unit standard (USTD\_START\_DATE and USTD\_MAX\_START\_DATE), where;

- a unit standard enrolment record with a start date before the start date of the unit standard’s registration (USTD\_START\_DATE) would be given a negative value of the number of months between these two values (for example Enrolment A on Figure G.3.10.1),
- a unit standard enrolment record with a start date that falls between the start date of the unit standard’s registration (USTD\_START\_DATE) and the last date on which a

learner may enrol on a unit standard (USTD\_MAX\_START\_DATE) is given a value of 0 (for example Enrolment C on Figure G.3.10.1), and

- a unit standard enrolment record with a start date that is after the last date on which a learner may enrol on a unit standard would be given a positive value of the number of months between these two values (for example Enrolment E on Figure G.3.10.1).

The second indicator shows the difference, as a rounded numeric value (see Section 3.8.3.6), between the end date (END\_DATE) of the active enrolment time period and the start date of the unit standard's registration and the last date on which a learner may achieve a unit standard (USTD\_START\_DATE and USTD\_MAX\_END\_DATE)), where;

- a unit standard enrolment record with an end date before the start date of the unit standard's registration would be given a negative value of the number of months between these two values (for example Enrolment A on Figure G.3.10.1),
- a unit standard enrolment record with an end date that falls between the start date of the unit standard registration and the last date on which a learner may achieve a unit standard is given a value of 0 (for example Enrolment C on Figure G.3.10.1), and
- a unit standard enrolment record with an end date that is after the last date on which a learner may achieve a unit standard would be given a positive value of the number of months between these two values (for example Enrolment E on Figure G.3.10.1).

This logic resulted in the addition of the following new indicators on the table DM\_USTD\_ENROL:

66. START\_USTD\_REGSTR\_IND: Numeric value indicating the distance between the start date of the unit standard enrolment record and the unit standard's active registration.
67. END\_USTD\_REGSTR\_IND: Numeric value indicating the distance between the end date of the unit standard enrolment record and the unit standard's active registration.

Using the values in these two fields in conjunction with each other it was possible to derive a code and corresponding description for USTD\_REGSTR\_IND as follows:

- Where a unit standard's registration did not exist, allocating a value of ' No Registration' to the record.

- Allocating a value of ‘OK’ to records where START\_USTD\_REGSTR\_IND is equal to 0 and END\_USTD\_REGSTR\_IND is equal to 0.

For all remaining records:

- Allocating a value of ‘Start Before’ to records where START\_USTD\_REGSTR\_IND was less than 0, ‘Start During’ to records where START\_USTD\_REGSTR\_IND is equal to 0 and ‘Start After’ where START\_USTD\_REGSTR\_IND was greater than 0.
- Allocating a value ‘End Before’ to records where END\_USTD\_REGSTR\_IND was less than 0, ‘End During’ to records where END\_USTD\_REGSTR\_IND is equal to 0 and ‘End After’ where END\_USTD\_REGSTR\_IND was greater than 0.

The final derivation steps included:

- Amending the data code and appending the word ‘Predicted’ to the indicator value for any records with a END\_DATE value greater than CYCLE\_DATE (for example Enrolment E on Figure G.3.10.1). In other words all records with a calculated end date that is greater than the latest data submission cycle date (see Section 3.8.3.2).

This logic resulted in the addition of the USTD\_REGSTR\_IND indicator code and corresponding description on the table DM\_USTD\_ENROL.

68. USTD\_REGSTR\_IND\_ID and USTD\_REGSTR\_IND\_DESC: Code and corresponding description denoting a record’s compliance in regard to whether the unit standard was registered for the duration of the learner’s active enrolment on the unit standard.

### ***G.3.11 Removal of replaced unit standard enrolments***

The evolution of unit standards as described in Section 3.8.3.4 can in some instances result in a data management issue that must be addressed for the purposes of this research.

In some instances, a provider may enrol the learner on a unit standard that has been replaced. The enrolment of the unit standard that has been replaced may be captured on the operational information system of the ETQE and as a result may be submitted to the NLRD. The NLRD has clearly defined protocol that allows the ETQE to indicate when a record has been incorrectly submitted to the NLRD, unfortunately not all ETQEs complete the protocol in this regard. On discovery of the enrolment on the replaced unit standard the ETQE may incorrectly perform the following actions on their operational information system:

- update the existing enrolment record with the replacement unit standard ID, or
- create an entirely new enrolment record for the learner against the replacement unit standard.

The data loading procedures of the NLRD work on the principle that the combination of learner ID and unit standard ID is unique. As a result, in either scenario as described above, when the enrolment against the replacement unit standard is loaded on the NLRD, a new unit standard enrolment record is created for the learner. Enrolment records against unit standards that have been replaced, that have been incorrectly loaded on the NLRD will invariably generate false positives against a number of the semantic business rules. In consultation with the Director of the NLRD it was decided that such records should be excluded from this research.

As a result, any enrolment records against a unit standard that has been replaced, which has a further enrolment for the same learner, against the replacement unit standard were allocated a PROBLEM\_ID of 4 and excluded from the further processing of the data. These records constituted 1.26% of the unit standard enrolment records initially extracted from the NLRD.

#### ***G.3.12 DM\_USTD\_ENROL\_FINAL***

The derivation steps described from Appendix G.3.1 to Appendix G.3.10 were saved in a new data table called DM\_USTD\_ENROL\_FINAL. This table included all of the data records initially received from the NLRD in the table DM\_USTD\_ENROL, including the problem records described in Appendix G.3.1, Appendix G.3.2 and Appendix G.3.11 (i.e. records that have a value in the data field PROBLEM\_ID). The problem records were immediately communicated to SAQA, who in turn implemented processes and procedures in order to address these records.

A technical description of the table DM\_USTD\_ENROL\_FINAL has been provided in G.1.

#### ***G.4 Appendix summary***

This section detailed the development of the data table DM\_USTD\_ENROL\_FINAL which will be used in the analysis and data mining of the unit standard enrolment records received from the NLRD. The section identified the semantic business rules that are applicable to unit

standard enrolment records and then described the selection, pre-processing and derivation steps implemented to establish the table DM\_USTD\_ENROL\_FINAL, which contains the unit standard enrolment records in a format that is suited for data mining.

## Appendix H

This appendix provides a detailed specification of the table DM\_USTD\_ENROL\_FINAL.

Table Name	Column Name	Data Type	Data Length	Allow NULLs	Comment
DM_USTD_ENROL_FINAL	LEARNER_ENROLMENT_ID	NUMBER	22	Y	De-identified
DM_USTD_ENROL_FINAL	LEARNER_ID	NUMBER	22	Y	De-identified
DM_USTD_ENROL_FINAL	UNIT_STANDARD_ID	NUMBER	22	Y	De-identified
DM_USTD_ENROL_FINAL	QUALIFICATION_ID	NUMBER	22	Y	De-identified
DM_USTD_ENROL_FINAL	LEARNERSHIP_ID	NUMBER	22	Y	De-identified
DM_USTD_ENROL_FINAL	ETQE_ID	NUMBER	22	Y	De-identified
DM_USTD_ENROL_FINAL	PROVIDER_ID	NUMBER	22	Y	De-identified
DM_USTD_ENROL_FINAL	ASSESSOR_ID	NUMBER	22	Y	De-identified
DM_USTD_ENROL_FINAL	ENROL_STATUS_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	ENROL_STATUS_DESC	VARCHAR2	50	Y	
DM_USTD_ENROL_FINAL	ENROL_TYPE_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	ENROL_TYPE_DESC	VARCHAR2	50	Y	
DM_USTD_ENROL_FINAL	ENROL_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	ACHIEVE_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	DERIVED_START_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	CREDITS	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	USTD_START_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	USTD_END_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	USTD_MAX_START_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	USTD_MAX_END_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	TRANSITION_PERIOD	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	TRAIN_OUT_PERIOD	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	NQF_LEVEL_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	NQF_LEVEL_DESC	VARCHAR2	26	Y	
DM_USTD_ENROL_FINAL	UNIT_STD_TYPE_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	UNIT_STD_TYPE_DESC	VARCHAR2	26	Y	
DM_USTD_ENROL_FINAL	FIELD_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	FIELD_DESC	VARCHAR2	60	Y	
DM_USTD_ENROL_FINAL	SUBFIELD_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	SUBFIELD_DESC	VARCHAR2	80	Y	
DM_USTD_ENROL_FINAL	PROBLEM_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	START_DATE_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	START_DATE_DESC	VARCHAR2	26	Y	
DM_USTD_ENROL_FINAL	START_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	END_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	ETQE_START_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	ETQE_END_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	ETQE_ACCRED_START_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	ETQE_ACCRED_END_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	PROV_START_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	PROV_END_DATE	DATE	7	Y	

Table Name	Column Name	Data Type	Data Length	Allow NULLs	Comment
DM_USTD_ENROL_FINAL	PROVIDER_TYPE_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	PROVIDER_TYPE_DESC	VARCHAR2	26	Y	
DM_USTD_ENROL_FINAL	PROVIDER_CLASS_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	PROVIDER_CLASS_DESC	VARCHAR2	50	Y	
DM_USTD_ENROL_FINAL	PROV_PROVINCE_CODE	VARCHAR2	10	Y	
DM_USTD_ENROL_FINAL	PROV_PROVINCE_DESC	VARCHAR2	60	Y	
DM_USTD_ENROL_FINAL	PROV_ACCRED_START_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	PROV_ACCRED_END_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	ASOR_START_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	ASOR_END_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	ASOR_REGSTR_START_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	ASOR_REGSTR_END_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	ETQE_FIRST_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	CYCLE_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	START_ETQE_IND	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	END_ETQE_IND	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	ETQE_IND_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	ETQE_IND_DESC	VARCHAR2	156	Y	
DM_USTD_ENROL_FINAL	OTHR_ETQE_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	OTHR_ETQE_START_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	OTHR_ETQE_END_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	OTHR_START_ETQE_IND	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	OTHR_END_ETQE_IND	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	OTHR_ETQE_IND_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	OTHR_ETQE_IND_DESC	VARCHAR2	24	Y	
DM_USTD_ENROL_FINAL	START_ETQE_ACCRED_IND	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	END_ETQE_ACCRED_IND	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	ETQE_ACCRED_IND_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	ETQE_ACCRED_IND_DESC	VARCHAR2	156	Y	
DM_USTD_ENROL_FINAL	OTHR_ETQE_ACCRED_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	OTHR_ETQE_ACCRED_START_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	OTHR_ETQE_ACCRED_END_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	OTHR_START_ETQE_ACCRED_IND	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	OTHR_END_ETQE_ACCRED_IND	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	OTHR_ETQE_ACCRED_IND_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	OTHR_ETQE_ACCRED_IND_DESC	VARCHAR2	24	Y	
DM_USTD_ENROL_FINAL	PROV_ETQE_ID	NUMBER	22	Y	De-identified
DM_USTD_ENROL_FINAL	PROV_ETQE_FIRST_DATE	DATE	7	Y	
DM_USTD_ENROL_FINAL	START_PROV_IND	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	END_PROV_IND	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	PROV_IND_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	PROV_IND_DESC	VARCHAR2	67	Y	
DM_USTD_ENROL_FINAL	START_PROV_ACCRED_IND	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	END_PROV_ACCRED_IND	NUMBER	22	Y	

Table Name	Column Name	Data Type	Data Length	Allow NULLs	Comment
DM_USTD_ENROL_FINAL	PROV_ACCRED_IND_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	PROV_ACCRED_IND_DESC	VARCHAR2	67	Y	
DM_USTD_ENROL_FINAL	END_ASOR_IND	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	ASOR_IND_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	ASOR_IND_DESC	VARCHAR2	85	Y	
DM_USTD_ENROL_FINAL	END_ASOR_REGSTR_IND	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	ASOR_REGSTR_IND_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	ASOR_REGSTR_IND_DESC	VARCHAR2	85	Y	
DM_USTD_ENROL_FINAL	START_USTD_REGSTR_IND	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	END_USTD_REGSTR_IND	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	USTD_REGSTR_IND_ID	NUMBER	22	Y	
DM_USTD_ENROL_FINAL	USTD_REGSTR_IND_DESC	VARCHAR2	56	Y	

## Appendix I

### *1.1 Introduction*

The literature review conducted for this research identified three data mining techniques that lend themselves to the identification, measurement and description of data quality deficiencies (see Section 2.4). The three data mining techniques identified were EDM techniques, cluster data mining and association rule data mining.

- EDM techniques allows for the summarization of the data and identification of hidden relationships.
- Cluster data mining was utilized if a relationship was identified using EDM techniques, but the relationship could not be properly understood because the patterns in the data were too diverse.
- In contrast association rule data mining techniques were implemented across all data records that contravened one or more semantic business rules in order to determine whether there were relationships between these records that could not be easily identified by EDM and cluster data mining techniques.

The following sections provide an overview of each of these techniques and how they were implemented in this study.

### *1.2 Exploratory data mining*

The most frequently used data mining techniques for this study were EDM techniques. As already indicated in Section 2.2, due to the descriptive nature of the research, the techniques utilized focused on summarizing the data and finding hidden relationships.

The analysis of any data indicator was always preceded with an exploratory view of the data generated by the data mining tool. The tool automatically generated a large amount of statistics about the data being analysed, including aggregated values per data field such as (see Figure I.2.1):

1. The percentage records that had a NULL value for the data field.
2. The number and percentage of distinct values for the data field.
3. The mode, average, median, minimum value, maximum value, standard deviation and variance for the data field.

Name	Histogram	Data Type	Percent NULLs	Distinct Values	Distinct Per...	Mode	Average	Median	Min Value	Max Value	Standard Devi...	Variance
START_DATE_IND		NUMBER	0	86	4.3566		98.7097	99	33	156	14.9803	224.4089
START_DATE_ID		VARCHAR2	0	2	0.1013	1						
START_DATE_DESC		VARCHAR2	0	2	0.1013	Actual						
START_DATE		VARCHAR2	0	87	4.4073	<Other>						
QENROL_IND_ID		VARCHAR2	0	1	0.0507	4						
QENROL_IND_DESC		VARCHAR2	0	1	0.0507	Lshp Enrolled, Qual Ac...						
QENROL_ENROL_STATUS_DESC		VARCHAR2	0	1	0.0507	Achieved						
PROVIDER_ID		VARCHAR2	0	32	1.6211	11,091						
NQF_LEVEL_ID		VARCHAR2	0	4	0.2026	4						
NQF_LEVEL_DESC		VARCHAR2	0	4	0.2026	Level 3						
LSHP_ETQE_ID		VARCHAR2	0	1	0.0507	1,115						
LEARNERSHIP_ID		VARCHAR2	0	19	0.9625	<Other>						
LEARNER_ID		VARCHAR2	0	1,694	85.8156	<Other>						
LEARNER_ENROLMENT_ID		VARCHAR2	0	1,974	100	<Other>						
ETQE_ID		VARCHAR2	0	1	0.0507	1,115						
ENROL_TYPE_ID		VARCHAR2	0	1	0.0507	6						
ENROL_TYPE_DESC		VARCHAR2	0	1	0.0507	Other						
ENROL_STATUS_ID		VARCHAR2	0	1	0.0507	3						
ENROL_STATUS_DESC		VARCHAR2	0	1	0.0507	Enrolled						
END_QENROL_IND		NUMBER	100	0	0							
END_DATE_IND		NUMBER	0	86	4.3566		110.7097	111	45	168	14.9803	224.4089
END_DATE		VARCHAR2	0	87	4.4073	<Other>						

Figure I.2.1 Screenshot of exploratory data mining results generated the data mining tool

The tool also automatically generated a graph, per data field, that shows the distribution of values in the data field (see Figure I.2.2).

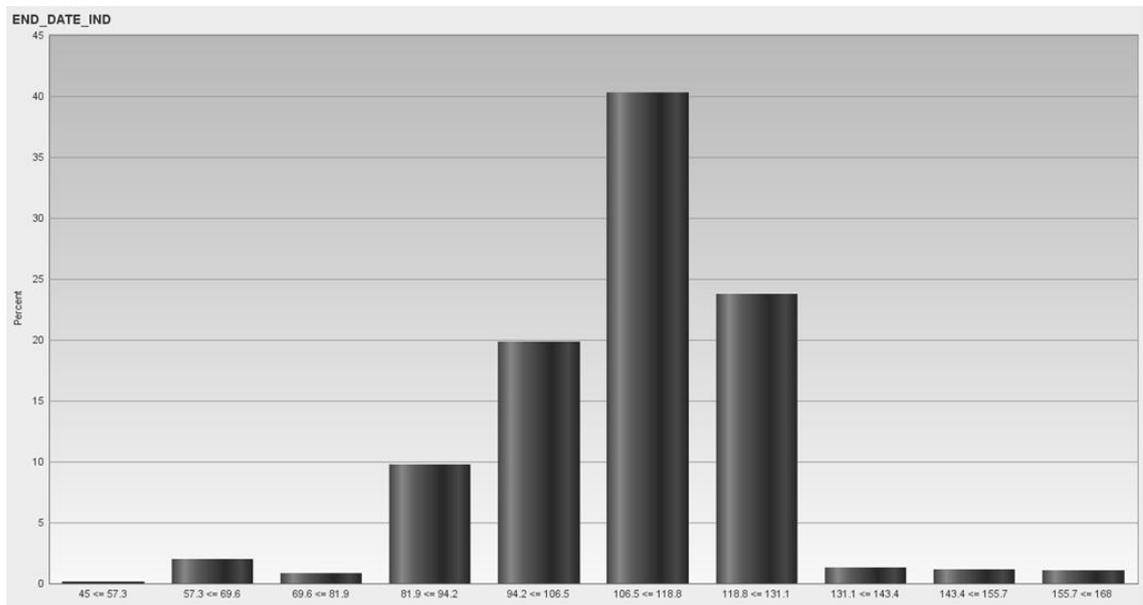


Figure I.2.2 Screenshot of automatically generated graph showing the distribution of value for a data field

Although these automatically generated statistics and graphs were extremely useful when trying to gain an understanding of the data, the expansiveness of the results prevented the utilization of these types of outputs for the analysis of the data.

Specific queries were developed to provide outputs for instances where standardized statistics needed to be generated for the research. These types of outputs have been used throughout the analysis of the data in this document. For example a specific query was developed to generate standard results such as the overview of the categories for an indicator (see Table 4.2.1.1 for an example). The results of specific queries were also used to generate standard graphs for an indicator (see Figure 4.2.1.1 for an example).

Finally, the results of specific queries also generated the statistics that were elaborated on in the analysis. For example the following type of output was developed and utilized for the analysis of the ‘Start Before, End During’ category of the analysis of whether the ETQE was accredited for the duration of the learner’s active enrolment on the learnership.

Table I.2.3 Example of statistics generated for records that have ETQE\_IND\_DESC ‘Start Before, End During’

ETQE Identifier	% Records of Rule	% Records submitted
1105	95.93%	5.37%
1116	1.90%	0.54%
1122	0.98%	0.35%
1112	0.98%	0.09%
1119	0.11%	0.02%
1127	0.05%	0.01%
1115	0.05%	0.01%
<b>Grand Total</b>	<b>100.00%</b>	<b>6.39%</b>

### ***I.3 Clustering***

Cluster analysis is one of the most frequently used data mining techniques. The technique involves separating groups of data into collections that include consistent patterns. K-means clustering is a clustering method used to automatically partition data into  $k$  groups (MacQueen, 1967, p. 281). The clustering process starts by assigning a value to  $k$  for the number of cluster centres which are randomly placed as centroids in the data. Each item is then assigned to its nearest cluster centre using Euclidean distance (1). The mean of all of the vectors in the group is then used to recompute the cluster centre (2). Steps (1) and (2) are repeated until convergence is achieved (the cluster centres do not move much) or for a fixed number of iterations (MacQueen, 1967, p. 282). The overall processing required for clustering is computationally economical however the data mining technique has two disadvantages in that it requires the assignment of a value to  $k$  and the initial placement of the

cluster centres is random which means that the results may not be repeatable (MacQueen, 1967, p. 281).

This section describes the approach that was implemented when cluster data mining was conducted for this research. The source data set used in the cluster data mining was the data set that was developed as a result of the data selection, pre-processing and derivation activity for the specific type of data (see Appendix C, Appendix E and Appendix G).

The data set was then filtered according to the requirements of the analysis. For example, the data set mined for the learnership enrolment provider accreditation category 'Start Before, End Before or End During' (see Appendix J.1.7) was filtered from the data set DM\_LSHP\_ENROL\_FINAL (see Appendix C.3.8) where PROV\_IND\_DESC contained 'Start Before, End Before' or 'Start Before, End During'.

The resultant data set was then further transformed to exclude any data values that were irrelevant or redundant in order to reduce the risk of over fitting (Berthold, Borgelt, Höppner, & Klawonn, 2010, p. 117). For example, in the data set mined for the learnership enrolment provider accreditation category 'Start Before, End Before or End During' all lookup value ID fields were removed and their corresponding descriptive data fields retained.

The resultant data set was then divided into two data sets, in order to ensure a realistic performance of the model generated. The first of these data sets was a build data set (sometimes referred to as a training data set) which contained 60% of the data. The second of these data sets was a test data set which contained 40% of the data. The build data set is defined with more records than the test data set as recommended by Berthold, Borgelt, Höppner, et al. (Berthold, Borgelt, Höppner, & Klawonn, 2010, p. 102). The records for each data set were selected randomly using LEARNER\_ENROLMENT\_ID (the unique identifier for each enrolment record) as the case identifier. The implementation of a case identifier ensured reproducibility.

The cluster model was fitted using the build data set. The cluster algorithm utilized the k-means clustering algorithm, with parameters requiring that the model generate 8 clusters, using the Euclidean distance function with a variance split criterion.

Once the trained model was generated the model was tested against the test data set. The test data set in turn produced a result per LEARNER\_ENROLMENT\_ID indicating the cluster that the record was allocated to and the probability of the record belonging to the cluster. The overall accuracy of the model was then calculated as the average of all of the probabilities allocated to each data record.

The combination of the LEARNER\_ENROLMENT\_ID and its cluster probability provided a mechanism with which to determine whether a data record was anomalous. In the same manner as illustrated by Thiprungsri & Vasarhelyi (Thiprungsri & Vasarhelyi, 2011, p. 76), any records that have a cluster probability lower than .6 were considered anomalous records. Further, also as illustrated by Thiprungsri & Vasarhelyi (Thiprungsri & Vasarhelyi, 2011, p. 76), data records that belong to any cluster that was generated by the model, with a population of less than 1% of the total records, were also considered anomalous records.

A technical description of the model is provided for each clustering activity. For example the technical description of the model generated for the learnership enrolment provider accreditation category 'Start Before, End Before or End During' (see Appendix J.1.7) is provided in Appendix K.1.1. The technical description includes each cluster, the percentage of records in the cluster and the average probability of the cluster. The top ten attributes of the rule for the cluster is also provided.

A description of the model is also provided for each clustering activity. For example the description of the model generated for the learnership enrolment provider accreditation category 'Start Before, End Before or End During' (see Appendix J.1.7) is provided in Appendix J.1.7.

#### ***1.4 Association Rule***

Association rule mining searches for similarities or events that occur together within data records and tries to infer rules that express those relationships (Agrawal, Imieliński, & Swami, Mining association rules between sets of items in large databases, 1993, p. 208). Each rule is comprised of two different sets of items namely X and Y, where X is the antecedent and Y is the consequent, together they are referred to as an itemset (Agrawal, Imieliński, & Swami, Mining association rules between sets of items in large databases, 1993, p. 208). The

specific relationships are then applied in order to measure the support, confidence and lift for the rule where:

- Support is a measure of the proportion of records in which the itemset appear
- Confidence is a measure of the proportion of records with item X in which item Y also appear
- Lift is a measure of how likely items Y is to exist for item X

The Apriori principle states that if an itemset is uncommon then all item sets that contain the itemset in combination with additional items will also be uncommon (Agrawal & Srikant, Fast algorithms for mining association rules, 1994, p. 489). The Apriori principle is applied to association rule mining in order to reduce the number of item sets that need to be investigated by the algorithm (Agrawal & Srikant, Fast algorithms for mining association rules, 1994, p. 488). Notwithstanding the implementation of the Apriori principle association rule mining has the disadvantage of being computationally expensive and can generate uninteresting rules.

This section describes the approach that was implemented when association rule data mining was conducted for this research. The source data set used in the association rule data mining was the data set that was developed as a result of the data selection, pre-processing and derivation activity for the specific type of data (Appendix C, Appendix E and Appendix G).

The association rule data mining was conducted in order to determine whether there are any associations and connections between the contraventions of semantic business rules in learner enrolment records that contravene one or more of the semantic business rules. As a result the data set was filtered to include all records that contravened one or more of the semantic business rules for the type of data record as follows:

- For learnership enrolment records, all records that contravened the semantic business rules described in Sections 4.2.1, 4.4.1, 4.6.1 and 0.
- For qualification enrolment records, all records that contravened the semantic business rules described in Sections 4.2.2, 4.3.1, 4.4.2, 4.5.1, 4.6.2, 4.7.1, 4.9.2 and 4.10.1.
- For unit standard enrolment records, all records that contravened the semantic business rules described in Section 4.2.3, 4.3.2, 4.4.3, 4.5.2, 4.6.3, 4.7.2 and 4.9.2.

Association rules do not have a predefined testing metric, as a result the complete data set was data mined. The association rules were generated using the Apriori algorithm, and were evaluated with two measures, namely:

- The minimum support of the rule, i.e. the fraction of the cases in which the rule is correct (Berthold, Borgelt, Höppner, & Klawonn, 2010, p. 186), was set at 0.3% (Hipp, Muller, Hohendorff, & Naumann, 2007, p. 9), and
- The minimum confidence of the rule, i.e. the number of cases in which the rule was correct in relation to the number of cases in which the rule was applicable (Berthold, Borgelt, Höppner, & Klawonn, 2010, p. 186), was set at 85% (Hipp, Muller, Hohendorff, & Naumann, 2007, p. 9).

A technical description of the association rules generated for each association rule data mining activity is provided in each analysis.

### ***1.5 Appendix summary***

This section provided an overview of the three data mining techniques used for this research. The data mining techniques include EDM, which allows for the summarization and identification of hidden relationship in the data. Cluster data mining which was used to further describe a relationship that had been identified by EDM but could not be properly understood. Finally, association rule mining was utilized in order to infer relationships in the data. Further, this section also details the parameters implemented during the utilization of cluster data mining and association rule data mining.

## Appendix J

This appendix provides a detailed review of learner enrolment records in relation to whether the provider was accredited for the duration of the learner's active enrolment on the learnership, qualification or unit standard. The review focuses on gaining a better understanding of data records that fall into specific categories of the data field PROV\_IND (see Appendix C.3.5, Appendix E.3.6 and Appendix G.3.6).

The appendix was necessitated as a result of the scope and volume of records that infringe on this semantic business rule for learnership, qualification and unit standard enrolment records. As a result the structure of this appendix has sub sections that focus on the each of these types of enrolment records.

### *J.1 Learnership enrolments*

#### *J.1.1 Start Before, End Before*

This category indicates that the learnership enrolment both started before and either was completed or expired before the provider was accredited. This category contains 41.64% of all of the records that infringe on this semantic business rule.

Of the 27 discrete ETQEs in the dataset, 19 ETQEs are linked to this category. Of these records, 63.81% were submitted to the NLRD by 3 ETQEs.

Of the 814 discrete learnerships in the dataset, 211 learnerships are linked to this category. Of these 211 learnerships, 10 learnerships contribute to 62.65% of records in this category. Most notably, although 1 of the 211 learnerships only constitute 0.03% of the records; the records for this learnership represent 100% of the learnership enrolment records submitted to the NLRD for the learnership.

Of the 3038 discrete providers in the dataset, 203 providers are linked to this category. Of these 203 providers, 10 providers contribute to 55.51% of the records. Most notably, although 30 of the 203 providers only constitute 4.29% of the records; the records for these providers represent 100% of the learnership enrolment records submitted to the NLRD for the providers.

The high densities of records that infringe on this rule, in combination with the high percentage of records in this category for some providers, hints at possible systemic issues in regard to provider accreditations. The patterns in the data are however not clearly delineated and a further review of learnership enrolment records that fall into this category was conducted (refer to Appendix J.1.7).

### ***J.1.2 No Accreditation***

This category indicates that the provider that is linked to the learnership has never had an active accreditation. This category contains 23.24% of all of the records that infringe on this semantic business rule and this category is of greatest concern to SAQA.

Of the 27 discrete ETQEs in the dataset, 20 ETQEs are linked to this category. Of these records, 74.64% were submitted to the NLRD by 3 ETQEs.

Of the 814 discrete learnerships in the dataset, 158 learnerships are linked to this category. Of these 158 learnerships, 10 learnerships contribute to 49.92% of records in this category. Most notably, although 2 of the 158 learnerships only constitute 0.54% of the records; the records for these learnerships represent 100% of the learnership enrolment records submitted to the NLRD for the learnerships.

Of the 3038 discrete providers in the dataset, 159 providers are linked to this category. Of these 159 providers, 10 providers contribute to 62.64% of the records. Most notably, 149 of the 159 providers constitute 91.26% of the records; the records for these providers represent 100% of the learnership enrolment records submitted to the NLRD for the providers.

As already noted, this category indicates that the provider has never had an active accreditation. As a result, providers in this category cannot be reported on in any of the other categories that form part of this research. However, another category that these providers can exist in is the 'No Accreditation Predicted' category that is excluded from this research.

The reader should therefore note that the above statement *"...149 of the 159 providers constitute 91.26% of the records; the records for these providers represent 100% of the*

*learnership enrolment records submitted to the NLRD for the providers.*” must further be interpreted to mean that for the remaining 10 providers, 100% of the records for the specific provider fall into the categories ‘No Accreditation’ or ‘No Accreditation Predicted’.

Unlike any of the other categories for this semantic business rule, it is unlikely that learnership enrolment records that appear in the ‘No Accreditation’ category do so as a result of data capturing issues related to the enrolment record. Rather the data capturing or data quality issues reside in the provider record. As a result, the analysis of this category focuses on the provider records.

Table J.1.2.1 provides an overview of the records found in this category grouped by submitting ETQE. The table differentiates the number of providers that have the submitting ETQE as their primary ETQE (“Primary ETQE of provider” on the table) and the number of providers where the submitting ETQE is not the primary ETQE of the provider (“Not Primary ETQE of provider”). A submitting ETQE may utilize another ETQE’s providers for the offering of a learnership (Section 3.8.3.5). It is found that the same providers that are not accredited have been utilized by more than one ETQE and as a result the count of provider by submitting ETQE is 171, whereas there are only 159 discrete providers in this category.

Table J.1.2.1 ‘No Accreditation’ records by submitting ETQE identifier, count of Not Primary ETQE providers, count of Primary ETQE of provider and % learnership enrolment records in the category

Submitting ETQE Identifier	Not Primary ETQE of Provider		Primary ETQE of Provider		Total	
	Count of Provider	% Records	Count of Provider	% Records	Count of Provider	% Records
1102	2	0.43%	1	0.01%	3	0.44%
1103	7	2.31%	5	0.25%	12	2.57%
1105			12	19.05%	12	19.05%
1108	3	0.69%	1	0.35%	4	1.04%
1109	4	3.91%			4	3.91%
1111			10	1.88%	10	1.88%
1112	3	0.90%			3	0.90%
1114	4	6.88%			4	6.88%
1115			19	41.02%	19	41.02%
1116	1	0.03%	14	0.57%	15	0.60%
1117			1	0.10%	1	0.10%
1119	1	0.93%	43	13.63%	44	14.56%
1122	2	0.05%	12	0.78%	14	0.83%
1123			6	0.95%	6	0.95%
1125	2	0.03%	4	0.68%	6	0.72%
1126	2	0.05%			2	0.05%
1127	5	3.91%			5	3.91%
1113			5	0.57%	5	0.57%
1104			1	0.01%	1	0.01%
1107	1	0.01%			1	0.01%
<b>Grand Total</b>	<b>37</b>	<b>20.14%</b>	<b>134</b>	<b>79.86%</b>	<b>171</b>	<b>100.00%</b>

Analysis of Table J.1.2.1 shows the following notable trends:

1. ETQE identifier 1119 has the highest incidence of the number of providers (43) where the primary ETQE of the provider is the same as the ETQE that submitted the learnership enrolment records to the NLRD. ETQE identifier 1119 contributes to 13.63% of the records. Further, it is found that these 43 providers constitute 40.57% of the overall number of providers that this ETQE references in learnership enrolment records.
2. ETQE identifier 1115 has the second highest incidence of the number of providers (19) where the primary ETQE of the provider is the same as the ETQE that submitted the learnership enrolment records to the NLRD. ETQE identifier 1115 also has the highest percentage (41.02%) of learnership enrolment records that fall into this category. Further, it is found that these 19 providers constitute 32.20% of the overall number of providers that this ETQE references in learnership enrolment records.

Further analysis also highlighted that two of these providers have provider names like Private, No Provider and Other.

3. 20.14% of the records are as a result of an ETQE utilizing a provider, which has another ETQE as its primary ETQE, despite the provider not having been accredited by the primary ETQE of the provider.

### ***J.1.3 Start Before, End During***

This category indicates that the learnership enrolment started before the provider was accredited and either was completed or expired whilst the provider was accredited. This category contains 21.33% of all of the records that infringe on this semantic business rule.

Of the 27 discrete ETQEs in the dataset, 17 ETQEs are linked to this category. Of these records, 65.68% were submitted to the NLRD by 3 ETQEs.

Of the 814 discrete learnerships in the dataset, 214 learnerships are linked to this category. Of these 214 learnerships, 10 learnerships contribute to 53.10% of records in this category. Most notably, although 3 of the 214 learnerships only constitute 0.17% of the records; the records for these learnerships represent 100% of the learnership enrolment records submitted to the NLRD for the learnerships.

Of the 3038 discrete providers in the dataset, 251 providers are linked to this category. Of these 251 providers, 10 providers contribute to 54.32% of the records. Most notably, although 40 of the 251 providers only constitute 4.25% of the records; the records for these providers represent 100% of the learnership enrolment records submitted to the NLRD for the providers.

The high densities of records that infringe on this rule, in combination with the high percentage of records in this category for some providers, hints at possible systemic issues in regard to provider accreditations. The patterns in the data are however not clearly delineated and a further review of learnership enrolment that fall into this category needs to be conducted (refer to Appendix J.1.7).

Cross checking the results in this category with the results in the ‘Start Before, End Before’ (Appendix J.1.1) category shows some remarkable similarities. The first of which was that

two of the ETQEs identified as top contributors to the ‘Start Before, End Before’ category are top contributors to this category. A precursory review also revealed that 60% of the top 10 providers that contributed to the ‘Start Before, End Before’ category are top 10 contributors in this category. These similarities suggested that a further review conducted on these records should be conducted in conjunction with the records in the ‘Start Before, End Before’ category.

#### ***J.1.4 Start After, End After***

This category indicates that the learnership enrolment both started after and either was completed or expired after the provider was accredited. This category contains 9.22% of all of the records that infringe on this semantic business rule.

Of the 27 discrete ETQEs in the dataset, 19 ETQEs are linked to this category. Of these records, 71.78% were submitted to the NLRD by 3 ETQEs.

Of the 814 discrete learnerships in the dataset, 126 learnerships are linked to this category. Of these 126 learnerships, 10 learnerships contribute to 77.56% of records in this category. Most notably, although 3 of the 126 learnerships only constitute 0.30% of the records; the records for these learnerships represent 100% of the learnership enrolment records submitted to the NLRD for the learnerships.

Of the 3038 discrete providers in the dataset, 257 providers are linked to this category. Of these 257 providers, 10 providers contribute to 67.88% of the records. Most notably, although 49 of the 257 providers only contribute 7.96% of the records; the records for these providers represent 100% of the learnership enrolment records submitted to the NLRD for the providers.

The high densities of records that infringe on this rule in relation to the number of learnership enrolment records submitted to the NLRD for the provider hints at possible systemic issues in regard to provider accreditations. The patterns in the data are however not clearly delineated and a further review of learnership enrolment records submitted to the NLRD for these providers needed to be conducted (refer to Appendix J.1.8).

#### ***J.1.5 Start During, End After***

This category indicates that the learnership enrolment started whilst the provider was accredited, and either was completed or expired after the provider was no longer accredited. This category contains 4.56% of all of the records that infringe on this semantic business rule.

Of the 27 discrete ETQEs in the dataset, 21 ETQEs are linked to this category. Of these records, 38.56% were submitted to the NLRD by 3 ETQEs.

Of the 814 discrete learnerships in the dataset, 124 learnerships are linked to this category. Of these 124 learnerships, 10 learnerships contribute to 42.27% of records in this category. Most notably, although 1 of the 124 learnerships only constitutes 2.15% of the records; the records for this learnership represent 100% of the learnership enrolment records submitted to the NLRD for the learnership.

Of the 3038 discrete providers in the dataset, 152 providers are linked to this category. Of these 152 providers, 10 providers contribute to 49.33% of the records. Most notably, although 26 of the 152 providers only contribute 10% of the records; the records for these providers represent 100% of the learnership enrolment records submitted to the NLRD for the providers.

The high densities of records that infringe on this rule in relation to the number of learnership enrolment records submitted to the NLRD for the provider hints at possible systemic issues in regard to provider accreditations. The patterns in the data are however not clearly delineated and a further review of learnership enrolment records submitted to the NLRD for these providers needed to be conducted (refer to Appendix J.1.8).

Cross checking the results in this category with the results in the 'Start After, End After' (Appendix J.1.4) category shows some similarities. The first of which is that one of the ETQEs identified as top contributors to the 'Start After, End After' category is a top contributor to this category. A precursory review also revealed that 20% of the top 10 providers that contributed to the 'Start After, End After' category are top 10 contributors in this category. These similarities suggested that a further review conducted on these records should be conducted in conjunction with the records in the 'Start After, End After' category.

### ***J.1.6 Start Before, End After***

This category indicates that the learnership enrolment both started before the provider was accredited and either was completed or expired after the provider was no longer accredited. This category contains 0.00% of all of the records that infringe on this semantic business rule.

The records that fall into this category all belong to 1 ETQE (ETQE identifier 1116), 1 learnership (learnership identifier 24) and 1 provider (provider identifier 49419).

Further investigation revealed that the specific provider was only accredited from 14 January 2013 to 31 January 2013. In consideration to the fact that provider accreditations are generally allocated 1, 3 or 5 year time periods, it seems most likely that the accreditation details captured in the provider record are incorrect.

### ***J.1.7 Start Before, End Before or End During***

As stated in Appendix J.1.1 and J.1.3, the high density of records submitted to the NLRD for the providers in the categories ‘Start Before, End Before’ and ‘Start Before, End During’, in conjunction with intersections in the top 3 ranked ETQEs and top 10 ranked providers in these categories, suggested that further analysis of these records as a single data set should be conducted.

Further analysis found that these two initial categories shared 162 learnerships and 120 providers. As a result, the categories ‘Start Before, End Before’ and ‘Start Before, End During’ were grouped into a category called ‘Start Before, End Before or End During’ for this analysis. This category indicates that the learnership enrolment started before the provider was accredited and either was completed or expired before or whilst the provider was accredited. As a result of this consolidation the ‘Start Before, End Before or End During’ category contains 62.97% of all the records that infringe on this semantic business rule.

Of the 27 discrete ETQEs in the dataset, 20 ETQEs are linked to this category. More than 60% of these records were submitted to the NLRD by 3 ETQEs.

Of the 814 discrete learnerships in the dataset, 263 learnerships are linked to this category. Of these 263 learnerships, 10 learnerships contribute to 55.94% of records in this category. Most notably, although 6 of the 263 learnerships only constitute 0.11% of the records; the records for these learnerships represent 100% of the learnership enrolment records submitted to the NLRD for the learnerships.

Of the 3038 discrete providers in the dataset, 334 providers are linked to this category. Of these 334 providers, 10 providers contribute to 52.19% of the records. Most notably, although 92 of the 334 providers only constitute 8.51% of the records; the records for these providers represent 100% of the learnership enrolment records submitted to the NLRD for the providers.

The volume of records found in this consolidated category exceeded 8% of the total learnership enrolment records that form part of the research. The implementation of data mining techniques, beyond exploratory data mining techniques, is needed to provide SAQA with both a succinct description of these records and to attempt to identify any data records that exist in this category as a result of possible incorrect data capturing at the source of the data.

The clustering data mining technique, as described in Appendix I.3, is applied to this data set in an effort to achieve both these aims.

An initial review of the data in this category, in combination with the understanding that the provider linked to a learnership enrolment record may have a primary ETQE that differs from the ETQE that submitted the learnership enrolment record to the NLRD, prompted the implementation of a new data field on the data set prior to data mining. The indicator PRIMARY\_ETQE\_DESC was developed as a nominal data value that contains the value 'Primary ETQE of provider' if the ETQE identifier of the ETQE that submitted the learnership enrolment record to the NLRD (ETQE\_ID) was the same as the ETQE identifier of the primary ETQE of the provider (PROV\_ETQE\_ID). The same indicator would have the value 'Not Primary ETQE of provider' if the primary ETQE identifier of the provider differed from the ETQE identifier of the ETQE that submitted the learnership record to the NLRD. It was hoped that the implementation of the PRIMARY\_ETQE\_DESC indicator would allow the data mining algorithm to find discrete patterns in the data related to the

utilization of providers that are accredited by an ETQE other than the ETQE that submitted the learnership enrolment record to the NLRD.

The data mining effort results in the development of 8 data clusters (a technical description of each cluster is provided in Appendix K.1.1) that have the following most dominant characteristics:

1. Cluster 1

This cluster describes more than 18% of the records. The cluster is predominantly described as containing records where the provider associated to the record does not have the same primary ETQE as the ETQE that submitted the records to the NLRD. Further, this cluster indicates that the majority of these providers belong to ETQE identifier 1033.

2. Cluster 2

The cluster describes nearly 17% of the records as being submitted to the NLRD by ETQE identifier 1111. These records encompass 32 learnerships, which constitute slightly more than 35% of the learnerships referred to in the learnership enrolment records for this ETQE. Further, these records encompass 20 providers, which in turn represent nearly 16% of the providers referred to in the learnership enrolment records for this ETQE.

3. Cluster 3

This cluster describes slightly more than 16% of the records as belonging to a single learnership being offered by 2 providers. The records were submitted to the NLRD by ETQE identifier 1105 and the single learnership represents nearly 6% of the learnerships referred to in the learnership enrolment records for this ETQE. The 2 providers represent only slightly more than 1% of the providers referred to in the learnership enrolment records of this ETQE.

4. Cluster 4

The cluster describes slightly more than 12% of the records as belonging to a single provider ranging over 7 different learnerships. These records were submitted to the NLRD by ETQE identifier 1126 and although the single provider represents less than 1% of the providers referred to in the ETQEs learnership enrolment records, the 7 learnerships represents nearly 15% of the learnerships referred to in the ETQE's learner enrolment records.

5. Cluster 5

This cluster describes 11.5% of the records as being submitted by 4 different ETQEs. The providers in these records predominantly have a provider class description of 'Unknown' and have ETQE identifier 1126 as the provider's primary ETQE. The records encompass 8 providers and 12 learnerships.

6. Cluster 6

The cluster describes nearly 11.5% of the records as being submitted by 3 different ETQEs. The providers in these records predominantly have ETQE identifier 1105 and 1103 as the provider's primary ETQE. The records encompass 21 providers and 9 learnerships

7. Cluster 7

This cluster describes nearly 8% of the records as being submitted to the NLRD by 3 different ETQEs. These records encompass 25 learnerships and 24 providers. The providers in these records predominantly have ETQE identifier 1116 and 1115 as their primary ETQE.

8. Cluster 8

The cluster describes nearly 6% of the records as having been submitted to the NLRD by ETQE identifier 1114. These records encompass 7 learnerships, which constitute 17.50% of the learnerships referred to in the learnership enrolment records for this ETQE. Further, these records encompass 7 providers, which in turn represent nearly 6% of the providers referred to in the learnership enrolment records for this ETQE.

As stated before, this category contains records from 20 different ETQEs. The above description of the 8 clusters generated by the clustering algorithm shows that 4 of these clusters (clusters 2, 3, 4 and 5) each describe records that were submitted to the NLRD by a specific ETQE. Cluster 1 in turn shows remarkable affects in terms of the utilization of providers whose primary ETQE is other than that of the submitting ETQE.

The clustering algorithm does not generate any clusters with a population of less than 1%. A review of the probabilities allocated to each record in a cluster does however allow for the identification of records that are low enough to be considered anomalous as per the criteria defined in Appendix I.3. These records constitute 0.30% of the records found in this category, and possibly exist in this category as a result of data capturing problems at the source of the data.

### ***J.1.8 Start During, Start After and End After***

As stated in Appendix J.1.4 and J.1.5, the high density of records submitted to the NLRD for the providers in the categories ‘Start After, End After’ and ‘Start During, End After’, in conjunction with intersections in the top 3 ranked ETQEs and top 10 ranked providers in these categories, suggested that further analysis of these records as a single data set should be conducted.

Further analysis found that these two initial categories shared 72 learnerships and 54 providers. As a result, the categories ‘Start After, End After’ and ‘Start During, End After’ were grouped into a category called ‘Start During, Start After and End After’ for this analysis. This category indicates that the learnership enrolment started during or after the provider was accredited and either was completed or expired after the provider was no longer accredited. As a result of this consolidation the ‘Start During, Start After and End After’ category contains 13.78% of all the records that infringe on this semantic business rule.

Of the 27 discrete ETQEs in the dataset, 21 ETQEs are linked to this category. More than 53% of these records were submitted to the NLRD by 3 ETQEs.

Of the 814 discrete learnerships in the dataset, 178 learnerships are linked to this category. Of these 178 learnerships, 10 learnerships contribute to 59.07% of records in this category. Most notably, although 4 of the 178 learnerships only constitute 0.91% of the records; the records for these learnerships represent 100% of the learnership enrolment records submitted to the NLRD for the learnerships.

Of the 3038 discrete providers in the dataset, 355 providers are linked to this category. Of these 355 providers, 10 providers contribute to 56.43% of the records. Most notably, although 89 of the 355 providers only constitute 13.92% of the records; the records for these providers represent 100% of the learnership enrolment records submitted to the NLRD for the providers.

The volume of records found in this consolidated category constitutes 1.79% of the total learnership enrolment records that form part of the research. The implementation of data mining techniques, beyond exploratory data mining techniques, is needed to provide SAQA

with both a succinct description of these records and to attempt to identify any data records that exist in this category as a result of possible incorrect data capturing at the source of the data.

The clustering data mining technique, as described in Appendix I.3, is applied to this data set in an effort to achieve both these aims.

An initial review of the data in this category, in combination with the understanding that the provider linked to a learnership enrolment record may have a primary ETQE that differs from the ETQE that submitted the learnership enrolment record to the NLRD, prompted the implementation of a new data field on the data set prior to data mining. The indicator PRIMARY\_ETQE\_DESC was developed as a nominal data value that contains the value 'Primary ETQE of provider' if the ETQE identifier of the ETQE that submitted the learnership enrolment record to the NLRD (ETQE\_ID) was the same as the ETQE identifier of the primary ETQE of the provider (PROV\_ETQE\_ID). The same indicator would have the value 'Not Primary ETQE of provider' if the primary ETQE identifier of the provider differed from the ETQE identifier of the ETQE that submitted the learnership record to the NLRD. It was hoped that the implementation of the PRIMARY\_ETQE\_DESC indicator would allow the data mining algorithm to find discrete patterns in the data related to whether or not the ETQE submitted the data was primary ETQE of the provider that offered the learnership.

The data mining effort results in the development of 8 data clusters (a technical description of each cluster is provided in Appendix K.1.2) that have the following most dominant characteristics:

1. Cluster 1

This cluster describes slightly more than 27% of the records. The cluster is predominantly described as containing records with learnership identifiers 53 and 24, which are offered by 24 different providers that are accredited by ETQE identifier 1120 and 1116. For learnership identifier 53 the records in this category represent more than 15% of all learnership enrolments for this learnership. For learnership identifier 24 the records in this category represent more than 18% of all the learnership enrolments for this learnership.

2. Cluster 2

The cluster describes more than 22% of the records as belonging to 2 providers (provider identifiers 49723 and 37631). Both providers are accredited by ETQE identifier 1127 and the records in the cluster predominantly belong to learnership identifier 460. For provider identifier 49723 the records in this category represent nearly 82% of all learnership enrolments for this provider. For provider identifier 37631 the records in this category represent nearly 90% of all the learnership enrolments for this provider. Further, the records in this category represent nearly 40% of all the learnership enrolment records for the learnership with learnership identifier 460.

3. Cluster 3

This cluster describes nearly 18% of the records. The cluster is diverse in that it describes records submitted by 9 ETQEs, covering 35 different learnerships offered by 42 different providers.

4. Cluster 4

The cluster describes nearly 9% of the records as learnerships that are offered by providers where the submitting ETQE is not the primary ETQE of the provider. The primary ETQE of these providers have ETQE identifiers 1126, 1125 and 1031.

5. Cluster 5

This cluster describes more than 7% of the records. The cluster is diverse in that it describes records submitted by 4 ETQEs, covering 22 learnerships offered by 22 different providers.

6. Cluster 6

The cluster describes 6.5% of the records as learnerships that are offered by providers where the submitting ETQE is not the primary ETQE of the provider. The primary ETQE of the providers have ETQE identifiers 1115 and 1103.

7. Cluster 7

This cluster describes slightly more than 6% of the records as belonging to 5 learnerships, offered by 10 providers. The learnership was generally started whilst the provider was accredited and ended within 12 months of the end of the provider's accreditation.

8. Cluster 8

The cluster describes nearly 4% of the records as learnerships that are offered by providers where the submitting ETQE is not the primary ETQE of the provider. The submitting ETQEs of these records have ETQE identifiers 1106 and 1103.

The most notable clusters that are generated for this category are clusters 1, 2, 4, 6 and 8. Cluster 1 seems to describe a specific problem with the implementation of a learnership identifiers 53 and 24. Cluster 2 in turn seems to imply specific issues in the offering of a single learnership, by provider identifiers 49723 and 37631. Clusters 4, 6 and 8 show remarkable affects in terms of the utilization of providers whose primary ETQE is other than that of the submitting ETQE.

The clustering algorithm does not generate any clusters with a population of less than 1%. A review of the probabilities allocated to each record in a cluster does however allow for the identification of records that are low enough to be considered anomalous as per the criteria defined in Appendix I.3. These records constitute 3.86% of the records found in this category, and possibly exist in this category as a result of data capturing problems at the source of the data.

#### ***J.1.9 Summary of semantic infringements by ETQE***

The preceding sections provide the results of records that infringe on this semantic business rule from the granular perspective of the learnership enrolment record in relation to the complete dataset. This approach supports the determination of patterns within the data that point to systemic and anomalous problems within the overall dataset, which in turn lends itself to assessing the quality of the data in the data set.

The approach however ignores the diverse nature of ETQEs, and in particular the volume of the records that each ETQE submits to the NLRD. The final step in the analysis of this semantic business rule provides an overview of the percentage of records, calculated as a percentage of the number of records submitted by the ETQE, which infringe on this semantic business rule.

The results are presented as the percentage of records submitted by the ETQE that fall into a category that describes a semantic business rule issue (see Table J.1.9.1):

Table J.1.9.1 % of records submitted by an ETQE that have a category that describes a semantic business rule issue

ETQE Identifier	% Semantic Rule Issue
1115	52.16%
1105	30.65%
1121	29.82%
1106	22.91%
1116	22.45%
1117	19.62%
1119	16.72%
1111	15.75%
1127	12.88%
1126	12.32%
1120	10.33%
1103	6.64%
1114	6.29%
1125	4.75%
1102	4.33%
1109	4.14%
1108	4.13%
1118	3.68%
1110	3.41%
1123	2.63%
1112	2.43%
1122	2.11%
1107	1.64%
1113	1.07%
1104	0.10%

The results clearly illustrate that the infringement of this semantic business rule could be considered systemic at a number of the ETQEs.

## ***J.2 Qualification enrolments***

### ***J.2.1 Start Before, End Before***

This category indicates that the qualification enrolment both started before and either was achieved or expired before the provider was accredited. This category contains 38.20% of all of the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 20 ETQEs are linked to this category. Of these records, 63.23% were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 201 qualifications are linked to this category. Of these 201 qualifications, 10 qualifications contribute to 66.61% of records in this category.

Of the 5669 discrete providers in the dataset, 269 providers are linked to this category. Of these 269 providers, 10 providers contribute to 67.65% of the records. Most notably, although 49 of the 269 providers only constitute 2.09% of the records; the records for these providers represent 100% of the qualification enrolment records submitted to the NLRD for the providers.

The high densities of records that infringe on this rule, in combination with the high percentage of records in this category for some providers, hints at possible systemic issues in regard to provider accreditations. The patterns in the data are however not clearly delineated and a further review of qualification enrolment records that fall into this category was conducted (refer to Appendix J.2.7).

### ***J.2.2 Start Before, End During***

This category indicates that the qualification enrolment started before the provider was accredited and either was achieved or expired whilst the provider was accredited. This category contains 26.93% of all of the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 21 ETQEs are linked to this category. Of these records, 57.02% were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 235 qualifications are linked to this category. Of these 235 qualifications, 10 qualifications contribute to 58.47% of records in this category.

Of the 5669 discrete providers in the dataset, 388 providers are linked to this category. Of these 388 providers, 10 providers contribute to 53.10% of the records. Most notably, although 57 of the 388 providers only constitute 2.09% of the records; the records for these providers represent 100% of the qualification enrolment records submitted to the NLRD for the providers.

The high densities of records that infringe on this rule, in combination with the high percentage of records in this category for some providers, hints at possible systemic issues in regard to provider accreditations. The patterns in the data are however not clearly delineated and a further review of qualification enrolment that fall into this category needs to be conducted (refer to Appendix J.2.7).

Cross checking the results in this category with the results in the ‘Start Before, End Before’ (Appendix J.2.1) category shows some remarkable similarities. The first of which was that two of the ETQEs identified as top contributors to the ‘Start Before, End Before’ category are top contributors to this category. A precursory review also revealed that 50% of the top 10 providers that contributed to the ‘Start Before, End Before’ category are top 10 contributors in this category. These similarities suggested that a further review conducted on these records should be conducted in conjunction with the records in the ‘Start Before, End Before’ category.

### ***J.2.3 No Accreditation***

This category indicates that the provider that is linked to the qualification has never had an active accreditation. This category contains 21.25% of all of the records that infringe on this semantic business rule and this category is of greatest concern to SAQA.

Of the 29 discrete ETQEs in the dataset, 24 ETQEs are linked to this category. Of these records, 54.87% were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 180 qualifications are linked to this category. Of these 180 qualifications, 10 qualifications contribute to 60.96% of records in this category. Most notably, although 2 of the 180 qualifications only constitute 0.08% of the records; the records for these qualifications represent 100% of the qualification enrolment records submitted to the NLRD for the qualifications.

Of the 5669 discrete providers in the dataset, 303 providers are linked to this category. Of these 303 providers, 10 providers contribute to 57.18% of the records. Most notably, 284 of the 303 providers constitute 88.00% of the records; the records for these providers represent 100% of the qualification enrolment records submitted to the NLRD for the providers.

As already noted, this category indicates that the provider has never had an active accreditation. As a result, providers in this category cannot be reported on in any of the other categories that form part of this research. However, another category that these providers can exist in is the ‘No Accreditation (Qual Linked to Lshp)’ category that is excluded from this research.

The reader should therefore note that the above statement “...284 of the 303 providers constitute 88.00% of the records; the records for these providers represent 100% of the qualification enrolment records submitted to the NLRD for the providers.” must further be interpreted to mean that for the remaining 19 providers, 100% of the records for the specific provider fall into the categories ‘No Accreditation’ or ‘No Accreditation (Qual Linked to Lshp)’.

Unlike any of the other categories for this semantic business rule, it is unlikely that qualification enrolment records that appear in the ‘No Accreditation’ category do so as a result of data capturing issues related to the enrolment record. Rather the data capturing or data quality issues reside in the lack of a provider record with an active accreditation.

Table J.2.3.1 provides an overview of the records found in this category grouped by submitting ETQE. The table differentiates the number of providers that have the submitting ETQE as their primary ETQE (“Primary ETQE of provider” on the table) and the number of providers where the submitting ETQE is not the primary ETQE of the provider (“Not Primary ETQE of provider”). A submitting ETQE may utilize another ETQE’s providers for the offering of a qualification (Section 3.8.3.5). It is found that the same providers that are not accredited have been utilized by more than one ETQE and as a result the count of provider by submitting ETQE is 319, whereas there are only 303 discrete providers in this category.

Table J.2.3.1 ‘No Accreditation’ records by submitting ETQE identifier, count of Not Primary ETQE providers, count of Primary ETQE of provider and % qualification enrolment records in the category

Submitting ETQE Identifier	Not Primary ETQE of Provider		Primary ETQE of Provider		Total	
	Count of Provider	% Records	Count of Provider	% Records	Count of Provider	% Records
1034			2	0.11%	2	0.11%
1079			25	0.20%	25	0.20%
1102	2	0.47%	2	0.02%	4	0.50%
1103	6	1.18%	7	0.14%	13	1.31%
1104			3	0.07%	3	0.07%
1105			13	11.13%	13	11.13%
1106	3	2.40%			3	2.40%
1107	2	0.01%	1	0.02%	3	0.03%
1108	3	0.68%	1	0.08%	4	0.76%
1109	4	3.82%			4	3.82%
1110	1	12.56%			1	12.56%
1111			17	1.53%	17	1.53%
1112	3	0.47%			3	0.47%
1113			63	9.52%	63	9.52%
1114	5	3.89%	4	2.89%	9	6.78%
1115	1	0.06%	20	31.12%	21	31.18%
1116	3	0.20%	14	0.35%	17	0.55%
1119	1	0.49%	58	6.55%	59	7.04%
1122	2	0.27%	14	0.49%	16	0.75%
1123			4	0.37%	4	0.37%
1124			2	0.33%	2	0.33%
1125	2	0.02%	6	0.50%	8	0.52%
1126	4	3.04%	16	1.38%	20	4.43%
1127	5	3.66%			5	3.66%
<b>Grand Total</b>	<b>47</b>	<b>33.19%</b>	<b>272</b>	<b>66.81%</b>	<b>319</b>	<b>100.00%</b>

Analysis of Table J.2.3.1 shows the following notable trends:

1. ETQE identifier 1113 has the highest incidence of the number of providers (63) where the primary ETQE of the provider is the same as the ETQE that submitted the qualification enrolment records to the NLRD. ETQE identifier 1113 contributes to 9.52% of the records. Further, it is found that these 63 providers constitute 26.81% of the overall number of providers that this ETQE references in qualification enrolment records.
2. ETQE identifier 1119 has the second highest incidence of the number of providers (58) where the primary ETQE of the provider is the same as the ETQE that submitted the qualification enrolment records to the NLRD. ETQE identifier 1119 contributes to 7.04% of the records. Further, it is found that these 58 providers constitute 41.73% of the overall number of providers that this ETQE references in qualification enrolment records. This specific ETQE shows similar trends in the analysis of the No

Accreditation category for learnership enrolments in regard to provider accreditations (see Appendix J.1.2).

3. ETQE identifier 1115 has the highest percentage (31.18%) of qualification enrolment records that fall into this category. Further analysis also highlighted that two of these providers have provider names like Private, No Provider and Other. This specific ETQE shows similar trends in the analysis of the No Accreditation category for learnership enrolments in regard to provider accreditations (see Appendix J.1.2).
4. 33.19% of the records are as a result of an ETQE utilizing a provider, which has another ETQE as its primary ETQE, despite the provider not having been accredited by the primary ETQE of the provider.

#### ***J.2.4 Start After, End After***

This category indicates that the qualification enrolment both started after and either was achieved or expired after the provider was accredited. This category contains 7.42% of all of the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 21 ETQEs are linked to this category. Of these records, 63.45% were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 127 qualifications are linked to this category. Of these 127 qualifications, 10 qualifications contribute to 63.80% of records in this category. Most notably, although 4 of the 127 qualifications only constitute 0.09% of the records; the records for these qualifications represent 100% of the qualification enrolment records submitted to the NLRD for the qualifications.

Of the 5669 discrete providers in the dataset, 264 providers are linked to this category. Of these 264 providers, 10 providers contribute to 68.47% of the records. Most notably, although 53 of the 264 providers only contribute 4.69% of the records; the records for these providers represent 100% of the qualification enrolment records submitted to the NLRD for the providers.

The high densities of records that infringe on this rule in relation to the number of qualification enrolment records submitted to the NLRD for the provider hints at possible systemic issues in regard to provider accreditations. The patterns in the data are however

not clearly delineated and a further review of qualification enrolment records submitted to the NLRD for these providers needed to be conducted (refer to Appendix J.2.8).

### ***J.2.5 Start During, End After***

This category indicates that the qualification enrolment started whilst the provider was accredited, and either was achieved or expired after the provider was no longer accredited. This category contains 5.99% of all of the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 20 ETQEs are linked to this category. Of these records, 44.97% were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 142 qualifications are linked to this category. Of these 142 qualifications, 10 qualifications contribute to 50.38% of records in this category. Most notably, although 3 of the 142 qualifications only constitutes 0.08% of the records; the records for these qualifications represent 100% of the qualification enrolment records submitted to the NLRD for the qualification.

Of the 5669 discrete providers in the dataset, 256 providers are linked to this category. Of these 256 providers, 10 providers contribute to 42.51% of the records. Most notably, although 54 of the 256 providers only contribute 5.60% of the records; the records for these providers represent 100% of the qualification enrolment records submitted to the NLRD for the providers.

The high densities of records that infringe on this rule in relation to the number of qualification enrolment records submitted to the NLRD for the provider hints at possible systemic issues in regard to provider accreditations. The patterns in the data are however not clearly delineated and a further review of qualification enrolment records submitted to the NLRD for these providers needed to be conducted (refer to Appendix J.2.8).

Cross checking the results in this category with the results in the ‘Start After, End After’ (Appendix J.2.7) category shows some similarities. The first of which is that one of the ETQEs identified as top contributors to the ‘Start After, End After’ category is a top contributor to this category. A precursory review also revealed that 30% of the top 10

providers that contributed to the 'Start After, End After' category are top 10 contributors in this category. These similarities suggested that a further review conducted on these records should be conducted in conjunction with the records in the 'Start After, End After' category.

### ***J.2.6 Start Before, End After***

This category indicates that the qualification enrolment started before the provider was accredited and either was achieved or expired after the provider was no longer accredited. This category contains 0.21% of all of the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 5 ETQEs are linked to this category. Of these records, 95.56% were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 10 qualifications are linked to this category.

Of the 5669 discrete providers in the dataset, 33 providers are linked to this category. Of these 33 providers, 10 providers contribute to 84.00% of the records. Most notably, although 10 of the 33 providers only contribute 8.00% of the records; the records for these providers represent 100% of the qualification enrolment records submitted to the NLRD for the providers.

The low incidence of records that fall into this category suggests that these records are found in this category as a result of incorrect data capturing at the source of the data.

### ***J.2.7 Start Before, End Before or End During***

As stated in Appendix J.2.1 and J.2.2, the high density of records submitted to the NLRD for the providers in the categories 'Start Before, End Before' and 'Start Before, End During', in conjunction with intersections in the top 3 ranked ETQEs and top 10 ranked providers in these categories, suggested that further analysis of these records as a single data set should be conducted.

Further analysis found that these two initial categories shared 150 qualifications and 170 providers. As a result, the categories 'Start Before, End Before' and 'Start Before, End

During' were grouped into a category called 'Start Before, End Before or End During' for this analysis. This category indicates that the qualification enrolment started before the provider was accredited and either was achieved or expired before or whilst the provider was accredited. As a result of this consolidation the 'Start Before, End Before or End During' category contains 65.13% of all the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 22 ETQEs are linked to this category. Nearly 59% of these records were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 286 qualifications are linked to this category. Of these 286 qualifications, 10 qualifications contribute to 60.72% of records in this category. Most notably, although 1 of the 286 qualifications only constitute 0.03% of the records; the records for this qualification represent 100% of the qualification enrolment records submitted to the NLRD for the qualifications.

Of the 5669 discrete providers in the dataset, 487 providers are linked to this category. Of these 487 providers, 10 providers contribute to 57.25% of the records. Most notably, although 130 of the 487 providers only constitute 3.68% of the records; the records for these providers represent 100% of the qualification enrolment records submitted to the NLRD for the providers.

The volume of records found in this consolidated category constituted nearly 7% of the total qualification enrolment records that form part of the research. The implementation of data mining techniques, beyond exploratory data mining techniques, is needed to provide SAQA with both a succinct description of these records and to attempt to identify any data records that exist in this category as a result of possible incorrect data capturing at the source of the data.

The clustering data mining technique, as described in Appendix I.3, is applied to this data set in an effort to achieve both these aims.

An initial review of the data in this category, in combination with the understanding that the provider linked to a qualification enrolment record may have a primary ETQE that differs

from the ETQE that submitted the qualification enrolment record to the NLRD, prompted the implementation of a new data field on the data set prior to data mining. The indicator PRIMARY\_ETQE\_DESC was developed as a nominal data value that contains the value 'Primary ETQE of provider' if the ETQE identifier of the ETQE that submitted the qualification enrolment record to the NLRD (ETQE\_ID) was the same as the ETQE identifier of the primary ETQE of the provider (PROV\_ETQE\_ID). The same indicator would have the value 'Not Primary ETQE of provider' if the primary ETQE identifier of the provider differed from the ETQE identifier of the ETQE that submitted the qualification record to the NLRD. It was hoped that the implementation of the PRIMARY\_ETQE\_DESC indicator would allow the data mining algorithm to find discrete patterns in the data related to the utilization of providers that are accredited by an ETQE other than the ETQE that submitted the qualification enrolment record to the NLRD.

The data mining effort results in the development of 8 data clusters (a technical description of each cluster is provided in Appendix K.2.1) that have the following most dominant characteristics:

1. Cluster 1

This cluster describes nearly 21% of the records. The cluster is predominantly described as containing qualification enrolment records for 4 qualifications as offered by 16 providers. The records were submitted to the NLRD by ETQE identifier 1105 and have a subfield of 'Safety in Society'.

2. Cluster 2

The cluster describes nearly 19% of the records as being submitted to the NLRD by ETQE identifier 1106. These records encompass 5 qualifications offered by 7 providers. In all instances ETQE identifier 1106 is not the primary ETQE of the provider.

3. Cluster 3

This cluster describes slightly more than 13% of the records. The cluster is diverse in that it describes records submitted by 9 ETQEs, covering 41 different qualifications offered by 53 different providers.

4. Cluster 4

The cluster describes slightly more than 10% of the records. These records were submitted to the NLRD by ETQE identifier 1126 and encompass 16 different qualifications. The qualifications were offered by 21 providers all of which have ETQE identifier 1126 as their primary ETQE.

#### 5. Cluster 5

This cluster describes slightly more than 10% of the records as belonging to a single provider and encompassing 2 qualifications. The records were submitted to the NLRD by ETQE identifier 1116 and the same ETQE is the primary ETQE of the provider that offered these qualifications.

#### 6. Cluster 6

The cluster describes slightly more than 9.5% of the records as being submitted to the NLRD by ETQE identifier 1111. The records encompass 9 qualifications offered by 14 providers. The providers in these records predominantly have ETQE identifier 1111 as the provider's primary ETQE.

#### 7. Cluster 7

This cluster describes nearly 9% of the records as being submitted to the NLRD by ETQE identifier 1116. These records encompass 5 qualifications offered by 3 providers. All of these providers have ETQE identifier 1116 as their primary ETQE.

#### 8. Cluster 8

The cluster describes slightly more than 8.5% of the records as having been submitted to the NLRD by ETQE identifier 1126. These records encompass 9 qualifications offered by 2 providers, both of which have ETQE identifier 1126 as their primary ETQE.

As stated before, this category contains records from 22 different ETQEs. The above description of the 8 clusters generated by the clustering algorithm shows that 7 of these clusters (clusters 1, 2, 4, 5, 6, 7 and 8) each describe records that were submitted to the NLRD by a specific ETQE. Cluster 2 in turn shows remarkable affects in terms of the utilization of providers whose primary ETQE is other than that of the submitting ETQE.

The clustering algorithm does not generate any clusters with a population of less than 1%. A review of the probabilities allocated to each record in a cluster does however allow for the identification of records that are low enough to be considered anomalous as per the criteria defined in Appendix I.3. These records constitute 0.79% of the records found in this category, and possibly exist in this category as a result of data capturing problems at the source of the data.

### ***J.2.8 Start During, Start After and End After***

As stated in Appendix J.2.4 and J.2.5, the high density of records submitted to the NLRD for the providers in the categories ‘Start After, End After’ and ‘Start During, End After’, in conjunction with intersections in the top 3 ranked ETQEs and top 10 ranked providers in these categories, suggested that further analysis of these records as a single data set should be conducted.

Further analysis found that these two initial categories shared 77 qualifications and 67 providers. As a result, the categories ‘Start After, End After’ and ‘Start During, End After’ were grouped into a category called ‘Start During, Start After and End After’ for this analysis. This category indicates that the qualification enrolment started during or after the provider was accredited and either was achieved or expired after the provider was no longer accredited. As a result of this consolidation the ‘Start During, Start After and End After’ category contains 13.41% of all the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 22 ETQEs are linked to this category. More than 44% of these records were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 192 qualifications are linked to this category. Of these 192 qualifications, 10 qualifications contribute to 48.96% of records in this category. Most notably, although 7 of the 192 qualifications only constitute 0.08% of the records; the records for these qualifications represent 100% of the qualification enrolment records submitted to the NLRD for the qualifications.

Of the 5569 discrete providers in the dataset, 453 providers are linked to this category. Of these 453 providers, 10 providers contribute to 49.26% of the records. Most notably, although 122 of the 453 providers only constitute 8.44% of the records; the records for these providers represent 100% of the qualification enrolment records submitted to the NLRD for the providers.

The volume of records found in this consolidated category constitutes 1.43% of the total qualification enrolment records that form part of the research. The implementation of data mining techniques, beyond exploratory data mining techniques, is needed to provide SAQA with both a succinct description of these records and to attempt to identify any data records

that exist in this category as a result of possible incorrect data capturing at the source of the data.

The clustering data mining technique, as described in Appendix I.3, is applied to this data set in an effort to achieve both these aims.

An initial review of the data in this category, in combination with the understanding that the provider linked to a qualification enrolment record may have a primary ETQE that differs from the ETQE that submitted the qualification enrolment record to the NLRD, prompted the implementation of a new data field on the data set prior to data mining. The indicator PRIMARY\_ETQE\_DESC was developed as a nominal data value that contains the value 'Primary ETQE of provider' if the ETQE identifier of the ETQE that submitted the qualification enrolment record to the NLRD (ETQE\_ID) was the same as the ETQE identifier of the primary ETQE of the provider (PROV\_ETQE\_ID). The same indicator would have the value 'Not Primary ETQE of provider' if the primary ETQE identifier of the provider differed from the ETQE identifier of the ETQE that submitted the qualification record to the NLRD. It was hoped that the implementation of the PRIMARY\_ETQE\_DESC indicator would allow the data mining algorithm to find discrete patterns in the data related to whether or not the ETQE submitted the data was primary ETQE of the provider that offered the qualification.

The data mining effort results in the development of 8 data clusters (a technical description of each cluster is provided in Appendix K.2.2) that have the following most dominant characteristics:

1. Cluster 1

This cluster describes slightly more than 33.5% of the records. The cluster predominantly describes records that encompass 13 qualifications that are provided by 10 providers. These records were submitted to the NLRD by ETQE identifiers 1126 and 1106. In some instances, the providers in this cluster do not have ETQE identifiers 1126 and 1106 as their primary ETQE. The qualifications in this cluster generally have a Field description of 'Education, Training and Development' or 'Business, Commerce and Management Studies'.

2. Cluster 2

The cluster describes nearly 20% of the records as belonging to 9 qualifications offered by 14 providers. The records in this cluster were submitted to the NLRD by 3 different ETQEs (ETQE identifiers 1075, 1126 and 1127). The qualifications in this cluster generally have a Field description of 'Business, Commerce and Management Studies'.

3. Cluster 3

This cluster describes nearly 10% of the records. The cluster describes records submitted to the NLRD by 4 ETQEs (ETQE identifiers 1102, 1103, 1109 and 1111) covering 25 different qualifications, the majority of which were offered as part of a learnership as offered by 16 different providers. The qualifications in this cluster generally have a Field description of 'Physical Planning and Construction' or 'Manufacturing, Engineering and Technology'.

4. Cluster 4

This cluster has a probability of 1 and describes nearly 9.5% of the records as being submitted to the NLRD by ETQE identifier 1105. The cluster comprises of 3 qualifications offered by 9 providers. The ETQE in this instance is the primary ETQE of the 9 providers.

5. Cluster 5

The cluster is very diverse and describes nearly 9.5% of the records as having been submitted to the NLRD by 6 different ETQEs (ETQE identifiers 1107, 1110, 1111, 1122, 1125 and 1126). The cluster encompasses 20 qualifications offered by 35 providers.

6. Cluster 6

This cluster also has a probability of 1 and describes slightly more than 8.5% of the records as belonging to qualification identifier 48550. The qualification is offered by provider identifier 1905 as part of learnership identifier 53. All of these records were submitted to the NLRD by ETQE identifier 1120.

7. Cluster 7

This cluster describes slightly more than 5.60% of the records as belonging to 3 'Post Graduate Diploma' qualifications with qualification identifiers 20408, 20409 and 73729. These qualifications were offered by 98 different providers. All of these records were submitted to the NLRD by ETQE identifier 1116.

8. Cluster 8

The cluster describes nearly 4% of the records as belonging to 2 qualifications, with qualification identifiers 24010 and 49623 that have a Field description of 'Health

Sciences and Social Services'. The qualifications were offered by 4 providers (provider identifiers 2159, 37747, 38989 and 39001), all of which have ETQE identifier 1117 as their primary ETQE. All of these enrolment records were submitted to the NLRD by ETQE identifier 1117.

Of the 8 clusters generated 7 provide a very discrete description of the characteristics of the records found in the cluster. The most notable clusters that are generated for this category are clusters 4, 6, 7 and 8. Each of these clusters points to problems related either to specific qualifications, providers and ETQEs. None of the clusters seemed to indicate a trend in regard to the utilization of providers whose primary ETQE is other than that of the submitting ETQE.

The clustering algorithm does not generate any clusters with a population of less than 1%. A review of the probabilities allocated to each record in a cluster does however allow for the identification of records that are low enough to be considered anomalous as per the criteria defined in Appendix I.3. These records constitute 0.15% of the records found in this category, and possibly exist in this category as a result of data capturing problems at the source of the data.

### ***J.2.9 Summary of semantic infringements by ETQE***

The preceding sections provide the results of records that infringe on this semantic business rule from the granular perspective of the qualification enrolment record in relation to the complete dataset. This approach supports the determination of patterns within the data that point to systemic and anomalous problems within the overall dataset, which in turn lends itself to assessing the quality of the data in the data set.

The approach however ignores the diverse nature of ETQEs, and in particular the volume of the records that each ETQE submits to the NLRD. The final step in the analysis of this semantic business rule provides an overview of the percentage of records, calculated as a percentage of the number of records submitted by the ETQE, which infringe on this semantic business rule.

The results are presented as the percentage of records submitted by the ETQE that fall into a category that describes a semantic business rule issue (see Table J.2.9.1):

Table J.2.9.1 % of records submitted by an ETQE that have a category that describes a semantic business rule issue

ETQE Identifier	% Semantic Rule Issue
1116	49.81%
1110	25.25%
1075	19.62%
1105	18.62%
1115	15.59%
1111	13.48%
1106	13.29%
1113	11.50%
1124	11.35%
1126	8.92%
1118	8.91%
1120	8.75%
1119	8.44%
1114	7.74%
1127	7.04%
1109	5.60%
1125	4.96%
1123	4.77%
1117	4.09%
1103	3.13%
1102	2.99%
1108	2.27%
1122	2.13%
1112	1.87%
1079	1.87%
1107	0.88%
1034	0.52%
1104	0.31%

The results clearly illustrate that the infringement of this semantic business rule could be considered systemic at a number of the ETQEs.

### ***J.3 Unit Standard enrolments***

#### ***J.3.1 Start Before, End Before***

This category indicates that the unit standard enrolment both started before and either was achieved or expired before the provider was accredited. This category contains 45.75% of all of the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 23 ETQEs are linked to this category. Of these records, 57.59% were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 4497 are linked to this category. Of these 4497 unit standards, 10 unit standards contribute to 11.79% of records in this category. Most notably, although 5 of the 4497 unit standards only constitute 0.01% of the records; the records for these unit standards represent 100% of the unit standard enrolment records submitted to the NLRD.

Of the 6254 discrete providers in the dataset, 908 providers are linked to this category. Of these 908 providers, 10 providers contribute to 49.35% of the records.

The high densities of records that infringe on this rule, in combination with the high percentage of records in this category for some providers, hints at possible systemic issues in regard to provider accreditations. The patterns in the data are however not clearly delineated and a further review of unit standard enrolment records that fall into this category was conducted (refer to Appendix J.3.7).

### ***J.3.2 Start Before, End During***

This category indicates that the unit standard enrolment started before the provider was accredited and either was achieved or expired whilst the provider was accredited. This category contains 22.67% of all of the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 22 ETQEs are linked to this category. Of these records, 70.46% were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 4075 are linked to this category. Of these 4075 unit standards, 10 unit standards contribute to 21.10% of records in this category.

Of the 6254 discrete providers in the dataset, 791 providers are linked to this category. Of these 791 providers, 10 providers contribute to 49.52% of the records. Most notably, although 18 of the 791 providers only constitute 0.07% of the records; the records for these providers represent 100% of the unit standard enrolment records submitted to the NLRD for the providers.

The high densities of records that infringe on this rule, in combination with the high percentage of records in this category for some providers, hints at possible systemic issues in regard to provider accreditations. The patterns in the data are however not clearly delineated and a further review of unit standard enrolment that fall into this category needs to be conducted (refer to Appendix J.3.7).

Cross checking the results in this category with the results in the ‘Start Before, End Before’ (Appendix J.3.1) category shows some remarkable similarities. The first of which was that two of the ETQEs identified as top contributors to the ‘Start Before, End Before’ category are top contributors to this category. A precursory review also revealed that 60% of the top 10 providers that contributed to the ‘Start Before, End Before’ category are top 10 contributors in this category. These similarities suggested that a further review conducted on these records should be conducted in conjunction with the records in the ‘Start Before, End Before’ category.

### ***J.3.3 Start After, End After***

This category indicates that the unit standard enrolment both started after and either was achieved or expired after the provider was accredited. This category contains 12.63% of all of the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 24 ETQEs are linked to this category. Of these records, 47.20% were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 3970 are linked to this category. Of these 3970 unit standards, 10 unit standards contribute to 11.99% of records in this category. Most notably, although 40 of the 3970 unit standards only constitute 0.13% of the records; the records for these unit standards represent 100% of the unit standard enrolment records submitted to the NLRD.

Of the 6254 discrete providers in the dataset, 735 providers are linked to this category. Of these 735 providers, 10 providers contribute to 56.13% of the records. Most notably, although 94 of the 735 providers only contribute 4.75% of the records; the records for these

providers represent 100% of the unit standard enrolment records submitted to the NLRD for the providers.

The high densities of records that infringe on this rule in relation to the number of unit standard enrolment records submitted to the NLRD for the provider hints at possible systemic issues in regard to provider accreditations. The patterns in the data are however not clearly delineated and a further review of unit standard enrolment records submitted to the NLRD for these providers needed to be conducted (refer to Appendix J.3.8).

#### ***J.3.4 No Accreditation***

This category indicates that the provider that is linked to the unit standard has never had an active accreditation. This category contains 12.08% of all of the records that infringe on this semantic business rule and this category is of greatest concern to SAQA.

Of the 29 discrete ETQEs in the dataset, 26 ETQEs are linked to this category. Of these records, 50.50% were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 2973 are linked to this category. Of these 2973 unit standards, 10 unit standards contribute to 9.31% of records in this category. Most notably, although 23 of the 2973 unit standards only constitute 0.36% of the records; the records for these unit standards represent 100% of the unit standard enrolment records submitted to the NLRD.

Of the 6254 discrete providers in the dataset, 379 providers are linked to this category. Of these 379 providers, 10 providers contribute to 59.30% of the records. Most notably, 341 of the 379 providers constitute 48.37% of the records; the records for these providers represent 100% of the unit standard enrolment records submitted to the NLRD for the providers.

As already noted, this category indicates that the provider has never had an active accreditation. As a result, providers in this category cannot be reported on in any of the other categories that form part of this research. However, another category that these providers can exist in is the 'No Accreditation (Ustd Linked to Lshp)' category that is excluded from this research.

The reader should therefore note that the above statement “...341 of the 379 providers constitute 48.37% of the records; the records for these providers represent 100% of the unit standard enrolment records submitted to the NLRD for the providers” must further be interpreted to mean that for the remaining 38 providers, 100% of the records for the specific provider fall into the categories ‘No Accreditation’ or ‘No Accreditation (UStd Linked to Lshp)’.

Unlike any of the other categories for this semantic business rule, it is unlikely that unit standard enrolment records that appear in the ‘No Accreditation’ category do so as a result of data capturing issues related to the enrolment record. Rather the data capturing or data quality issues reside in the lack of a provider record with an active accreditation.

Table J.3.4.1 provides an overview of the records found in this category grouped by submitting ETQE. The table differentiates the number of providers that have the submitting ETQE as their primary ETQE (“Primary ETQE of provider” on the table) and the number of providers where the submitting ETQE is not the primary ETQE of the provider (“Not Primary ETQE of provider”). A submitting ETQE may utilize another ETQE’s providers for the offering of a unit standard (Section 3.8.3.5). It is found that the same providers that are not accredited have been utilized by more than one ETQE and as a result the count of provider by submitting ETQE is 455, whereas there are only 379 discrete providers in this category.

Table J.3.4.1 ‘No Accreditation’ records by submitting ETQE identifier, count of Not Primary ETQE providers, count of Primary ETQE of provider and % unit standard enrolment records in the category

Submitting ETQE Identifier	Not Primary ETQE of Provider		Primary ETQE of Provider		Total	
	Count of Provider	% Records	Count of Provider	% Records	Count of Provider	% Records
1100			2	2.39%	2	2.39%
1102	3	0.99%	7	0.12%	10	1.11%
1103	11	1.55%	17	0.24%	28	1.80%
1104	2	0.05%			2	0.05%
1105			53	21.45%	53	21.45%
1106	8	1.90%			8	1.90%
1107	2	0.11%	1	0.01%	3	0.12%
1108	2	0.08%	1	0.00%	3	0.08%
1109	5	4.35%			5	4.35%
1110	4	19.03%			4	19.03%
1111	2	0.08%	34	9.94%	36	10.02%
1112	2	0.37%			2	0.37%
1113	3	0.45%	45	3.31%	48	3.76%
1114	13	4.23%	4	1.31%	17	5.54%
1115	1	0.01%	31	9.29%	32	9.30%
1116	4	0.13%	2	0.06%	6	0.19%
1117			1	0.00%	1	0.00%
1118	1	0.04%	1	0.16%	2	0.21%
1119	71	2.21%	25	0.90%	96	3.11%
1120	1	0.76%			1	0.76%
1122	2	0.04%	52	0.53%	54	0.57%
1123	1	0.01%	4	0.90%	5	0.91%
1124	7	0.02%			7	0.02%
1125	2	0.21%	6	0.27%	8	0.48%
1126	9	4.19%	7	0.16%	16	4.35%
1127	6	8.13%			6	8.13%
<b>Grand Total</b>	<b>162</b>	<b>48.96%</b>	<b>293</b>	<b>51.04%</b>	<b>455</b>	<b>100.00%</b>

Analysis of Table J.3.4.1 shows the following notable trends:

1. ETQE identifier 1119 has the highest incidence of the number of providers (71) where the primary ETQE of the provider is not the same as the ETQE that submitted the unit standard enrolment records to the NLRD. ETQE identifier 1119 contributes to 2.21% of the records. Further, it is found that these 71 providers constitute 41.04% of the overall number of providers that this ETQE references in unit standard enrolment records.
2. ETQE identifier 1105 has the highest incidence of the number of providers (53) where the primary ETQE of the provider is the same as the ETQE that submitted the unit standard enrolment records to the NLRD. ETQE identifier 1105 contributes to 21.45%

of the records. Further, it is found that these 53 providers constitute 4.91% of the overall number of providers that this ETQE references in unit standard enrolment records.

3. ETQE identifier 1105 has the highest percentage (21.45%) of unit standard enrolment records that fall into this category. This specific ETQE shows similar trends in the analysis of the No Accreditation category for qualification and learnership enrolments in regard to provider accreditations (see Appendix J.1.2 and Appendix J.2.3).
4. 51.04% of the records are as a result of an ETQE utilizing a provider, which has another ETQE as its primary ETQE, despite the provider not having been accredited by the primary ETQE of the provider.

### ***J.3.5 Start During, End After***

This category indicates that the unit standard enrolment started whilst the provider was accredited, and either was achieved or expired after the provider was no longer accredited. This category contains 6.72% of all of the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 21 ETQEs are linked to this category. Of these records, 66.08% were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 3223 are linked to this category. Of these 3223 unit standards, 10 unit standards contribute to 13.53% of records in this category. Most notably, although 10 of the 3223 unit standards constitutes less than 0.01% of the records; the records for these unit standards represent 100% of the unit standard enrolment records submitted to the NLRD for the unit standard.

Of the 6254 discrete providers in the dataset, 578 providers are linked to this category. Of these 578 providers, 10 providers contribute to 46.13% of the records. Most notably, although 10 of the 578 providers only contribute 0.02% of the records; the records for these providers represent 100% of the unit standard enrolment records submitted to the NLRD for the providers.

The high densities of records that infringe on this rule in relation to the number of unit standard enrolment records submitted to the NLRD for the provider hints at possible systemic issues in regard to provider accreditations. The patterns in the data are however

not clearly delineated and a further review of unit standard enrolment records submitted to the NLRD for these providers needed to be conducted (refer to Appendix J.3.8).

Cross checking the results in this category with the results in the ‘Start After, End After’ (Appendix J.3.3) category shows some similarities. A precursory review also revealed that 10% of the top 10 providers that contributed to the ‘Start After, End After’ category are top 10 contributors in this category. These similarities suggested that a further review conducted on these records should be conducted in conjunction with the records in the ‘Start After, End After’ category.

### ***J.3.6 Start Before, End After***

This category indicates that the unit standard enrolment started before the provider was accredited and either was achieved or expired after the provider was no longer accredited. This category contains 0.15% of all of the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 12 ETQEs are linked to this category. Of these records, 73.77% were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 676 are linked to this category. Of these 682 unit standards, 10 unit standards contribute to 33.54% of records in this category.

Of the 6254 discrete providers in the dataset, 34 providers are linked to this category. Of these 34 providers, 10 providers contribute to 90.12% of the records.

The low incidence of records that fall into this category suggests that these records are found in this category as a result of incorrect data capturing at the source of the data.

### ***J.3.7 Start Before, End Before or End During***

As stated in Appendix J.3.1 and J.3.2, the high density of records submitted to the NLRD for the providers in the categories ‘Start Before, End Before’ and ‘Start Before, End During’, in conjunction with intersections in the top 3 ranked ETQEs and top 10 ranked providers in these categories, suggested that further analysis of these records as a single data set should be conducted.

Further analysis found that these two initial categories shared 3552 unit standards and 595 providers. As a result, the categories ‘Start Before, End Before’ and ‘Start Before, End During’ were grouped into a category called ‘Start Before, End Before or End During’ for this analysis. This category indicates that the unit standard enrolment started before the provider was accredited and either was achieved or expired before or whilst the provider was accredited. As a result of this consolidation the ‘Start Before, End Before or End During’ category contains 68.42% of all the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 23 ETQEs are linked to this category. Nearly 60% of these records were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 5039 are linked to this category. Of these 5039 unit standards, 10 unit standards contribute to 14.13% of records in this category. Most notably, although 11 of the 5039 unit standards only constitute 0.01% of the records; the records for this unit standard represent 100% of the unit standard enrolment records submitted to the NLRD.

Of the 6254 discrete providers in the dataset, 1107 providers are linked to this category. Of these 1107 providers, 10 providers contribute to 46.98% of the records. Most notably, although 91 of the 1107 providers only constitute 1.09% of the records; the records for these providers represent 100% of the unit standard enrolment records submitted to the NLRD for the providers.

The volume of records found in this consolidated category constituted nearly 9% of the total unit standard enrolment records that form part of the research. The implementation of data mining techniques, beyond exploratory data mining techniques, is needed to provide SAQA with both a succinct description of these records and to attempt to identify any data records that exist in this category as a result of possible incorrect data capturing at the source of the data.

The clustering data mining technique, as described in Appendix I.3, is applied to this data set in an effort to achieve both these aims.

An initial review of the data in this category, in combination with the understanding that the provider linked to a unit standard enrolment record may have a primary ETQE that differs from the ETQE that submitted the unit standard enrolment record to the NLRD, prompted the implementation of a new data field on the data set prior to data mining. The indicator PRIMARY\_ETQE\_DESC was developed as a nominal data value that contains the value 'Primary ETQE of provider' if the ETQE identifier of the ETQE that submitted the unit standard enrolment record to the NLRD (ETQE\_ID) was the same as the ETQE identifier of the primary ETQE of the provider (PROV\_ETQE\_ID). The same indicator would have the value 'Not Primary ETQE of provider' if the primary ETQE identifier of the provider differed from the ETQE identifier of the ETQE that submitted the unit standard record to the NLRD. It was hoped that the implementation of the PRIMARY\_ETQE\_DESC indicator would allow the data mining algorithm to find discrete patterns in the data related to the utilization of providers that are accredited by an ETQE other than the ETQE that submitted the unit standard enrolment record to the NLRD.

The data mining effort results in the development of 8 data clusters (a technical description of each cluster is provided in Appendix K.3.1) that have the following most dominant characteristics:

1. Cluster 1

This cluster describes more than 33% of the records. The cluster is predominantly described as containing unit standard enrolment records for 226 unit standards as offered by 28 providers. The records were submitted to the NLRD by ETQE identifier 1111 and have a field of 'Manufacturing, Engineering and Technology'.

2. Cluster 2

The cluster describes more than 19% of the records as being submitted to the NLRD by the ETQE identifier 1116 and 1127. These records encompass 247 unit standards offered by 15 providers. In some instances, ETQE identifier 1116 and 1127 is not the primary ETQE of the provider.

3. Cluster 3

The cluster describes more than 16% of the records as being submitted to the NLRD by the ETQE identifier 1105 and 1127. These records encompass 287 unit standards offered by 68 providers. In all instances the primary ETQE of the provider is 1105.

4. Cluster 4

The cluster describes more than 11% of the records. These records were submitted to the NLRD by ETQE identifier 1126 and 1127 and encompass 376 different unit standards. The unit standards were offered by 20 providers all of which have ETQE identifier 1126 as their primary ETQE. The majority of these provider are located in Gauteng.

5. Cluster 5

This cluster describes more than 6% of the records. The cluster is diverse in that it describes records submitted by 7 ETQEs, covering 450 unit standards offered by 61 different providers. In all instances the primary ETQE of the provider is the same as the submitting ETQE.

6. Cluster 6

The cluster describes slightly more than 6% of the records as being submitted to the NLRD by an ETQE other than the primary ETQE of the provider. The records encompass 96 unit standards offered by 11 providers. The providers in these records predominantly have ETQE identifier 1033 or 1106 as the provider's primary ETQE. These records were primarily submitted to the NLRD by ETQE 1106.

7. Cluster 7

This cluster describes more than 4% of the records as being submitted to the NLRD by an ETQE that is not the primary ETQE of the provider that offered the unit standard. These records encompass 610 unit standards offered by 62 providers.

8. Cluster 8

The cluster describes nearly 3% of the records as being Regular-Fundamental unit standards. These records encompass 146 unit standards offered by 56 providers.

As stated before, this category contains records from 23 different ETQEs. The above description of the 8 clusters generated by the clustering algorithm shows that 4 of these clusters (clusters 2, 3, 4 and 7) each describe records that were submitted to the NLRD by a specific ETQE. Clusters 6 and 7 in turn show remarkable affects in terms of the utilization of providers whose primary ETQE is other than that of the submitting ETQE.

The clustering algorithm does not generate any clusters with a population of less than 1%. A review of the probabilities allocated to each record in a cluster does however allow for the identification of records that are low enough to be considered anomalous as per the criteria defined in Appendix I.3. These records constitute 1.25% of the records found in this

category, and possibly exist in this category as a result of data capturing problems at the source of the data.

### ***J.3.8 Start During, Start After and End After***

As stated in Appendix J.3.3 and J.3.5, the high density of records submitted to the NLRD for the providers in the categories ‘Start After, End After’ and ‘Start During, End After’, in conjunction with intersections in the top 3 ranked ETQEs and top 10 ranked providers in these categories, suggested that further analysis of these records as a single data set should be conducted.

Further analysis found that these two initial categories shared 2701 unit standards and 325 providers. As a result, the categories ‘Start After, End After’ and ‘Start During, End After’ were grouped into a category called ‘Start During, Start After and End After’ for this analysis. This category indicates that the unit standard enrolment started during or after the provider was accredited and either was achieved or expired after the provider was no longer accredited. As a result of this consolidation the ‘Start During, Start After and End After’ category contains 19.35% of all the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 24 ETQEs are linked to this category. Nearly 46% of these records were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 4544 are linked to this category. Of these 4544 unit standards, 10 unit standards contribute to 10.88% of records in this category. Most notably, although 58 of the 4544 unit standards constitute less than 0.01% of the records; the records for these unit standards represent 100% of the unit standard enrolment records submitted to the NLRD.

Of the 6254 discrete providers in the dataset, 988 providers are linked to this category. Of these 988 providers, 10 providers contribute to 43.00% of the records. Most notably, although 112 of the 988 providers only constitute 0.04% of the records; the records for these providers represent 100% of the unit standard enrolment records submitted to the NLRD for the providers.

The volume of records found in this consolidated category constitutes 2.43% of the total unit standard enrolment records that form part of the research. The implementation of data mining techniques, beyond exploratory data mining techniques, is needed to provide SAQA with both a succinct description of these records and to attempt to identify any data records that exist in this category as a result of possible incorrect data capturing at the source of the data.

The clustering data mining technique, as described in Appendix I.3, is applied to this data set in an effort to achieve both these aims.

An initial review of the data in this category, in combination with the understanding that the provider linked to a unit standard enrolment record may have a primary ETQE that differs from the ETQE that submitted the unit standard enrolment record to the NLRD, prompted the implementation of a new data field on the data set prior to data mining. The indicator PRIMARY\_ETQE\_DESC was developed as a nominal data value that contains the value 'Primary ETQE of provider' if the ETQE identifier of the ETQE that submitted the unit standard enrolment record to the NLRD (ETQE\_ID) was the same as the ETQE identifier of the primary ETQE of the provider (PROV\_ETQE\_ID). The same indicator would have the value 'Not Primary ETQE of provider' if the primary ETQE identifier of the provider differed from the ETQE identifier of the ETQE that submitted the unit standard record to the NLRD. It was hoped that the implementation of the PRIMARY\_ETQE\_DESC indicator would allow the data mining algorithm to find discrete patterns in the data related to whether or not the ETQE submitted the data was primary ETQE of the provider that offered the unit standard.

The data mining effort results in the development of 8 data clusters (a technical description of each cluster is provided in Appendix K.3.2) that have the following most dominant characteristics:

1. Cluster 1

This cluster describes nearly 22% of the records. The cluster predominantly describes records that encompass 758 unit standards that are provided by 51 providers of which most have a provider class of "Private". These records were submitted to the NLRD by 12 ETQEs.

2. Cluster 2

The cluster describes nearly 21% of the records as belonging to 89 unit standards offered by 55 providers. The records in this cluster were submitted to the NLRD by ETQE identifier 1105. The unit standards in this cluster generally have a Field description of “Business, Commerce and Management Studies”.

3. Cluster 3

This cluster describes nearly 16% of the records. The cluster describes records submitted to the NLRD by 6 ETQEs (ETQE identifiers 1075, 1100, 1116, 1123, 1125 and 1126) covering 68 different unit standards as offered by 17 different providers of which most have a provider class of “Private”.

4. Cluster 4

This cluster describes nearly 13% of the records as being submitted to the NLRD by ETQE identifiers 1107 and 1115. The cluster comprises of 476 unit standards offered by 29 providers. The ETQE in this instance is the primary ETQE of the 29 providers.

5. Cluster 5

The cluster describes slightly more than 12% of the records as having been submitted to the NLRD by 6 different ETQEs (ETQE identifiers 1075, 1103, 1110, 1114, 1118 and 1126). The cluster encompasses 399 unit standards offered by 30 providers.

6. Cluster 6

This cluster describes slightly more than 7% of the records. The cluster describes records submitted to the NLRD by 6 ETQEs (ETQE identifiers 1102, 1103, 1109, 1122, 1126 and 1127) covering 387 different unit standards as offered by 55 different providers.

7. Cluster 7

This cluster describes slightly more than 6.5% of the records as being offered by 5 providers and encompass 206 unit standards. The majority of these unit standards have a subfield description of “Finance, Economics and Accounting”. The majority of these records were submitted to the NLRD by ETQE identifier 1127.

8. Cluster 8

The cluster describes slightly more than 3% of the records as belonging to 37 unit standards offered by 17 providers. All of these enrolment records were submitted to the NLRD by ETQE identifier 1106.

Of the 8 clusters generated 5 provide a very discrete description of the characteristics of the records found in the cluster. The most notable clusters that are generated for this category

are clusters 1, 2, 3, 7 and 8. Each of these clusters points to problems related either to specific unit standards, providers and ETQEs. None of the clusters seemed to indicate a trend in regard to the utilization of providers whose primary ETQE is other than that of the submitting ETQE.

The clustering algorithm does not generate any clusters with a population of less than 1%. A review of the probabilities allocated to each record in a cluster does however allow for the identification of records that are low enough to be considered anomalous as per the criteria defined in Appendix I.3. These records constitute 3.22% of the records found in this category, and possibly exist in this category as a result of data capturing problems at the source of the data.

### ***J.3.9 Summary of semantic infringements by ETQE***

The preceding sections provide the results of records that infringe on this semantic business rule from the granular perspective of the unit standard enrolment record in relation to the complete dataset. This approach supports the determination of patterns within the data that point to systemic and anomalous problems within the overall dataset, which in turn lends itself to assessing the quality of the data in the data set.

The approach however ignores the diverse nature of ETQEs, and in particular the volume of the records that each ETQE submits to the NLRD. The final step in the analysis of this semantic business rule provides an overview of the percentage of records, calculated as a percentage of the number of records submitted by the ETQE, which infringe on this semantic business rule.

The results are presented as the percentage of records submitted by the ETQE that fall into a category that describes a semantic business rule issue (see Table J.3.9.1):

Table J.3.9.1 % of records submitted by an ETQE that have a category that describes a semantic business rule issue

ETQE Identifier	% Semantic Rule Issue
1100	95.19%
1116	50.26%
1115	39.64%
1111	23.49%
1075	22.27%
1110	19.97%
1105	16.07%
1118	16.03%
1127	12.51%
1106	9.63%
1103	9.43%
1126	8.58%
1113	8.31%
1114	6.70%
1109	6.58%
1125	6.17%
1122	5.90%
1119	5.36%
1102	5.28%
1107	4.86%
1123	3.73%
1108	3.71%
1112	3.70%
1120	3.07%
1117	2.60%
1124	1.35%
1104	0.15%

The results clearly illustrate that the infringement of this semantic business rule could be considered systemic at a number of the ETQEs.

## Appendix K

This appendix provides a technical description of the outputs of data mining activities that were conducted when analysing whether the provider was accredited for the duration of the learner's active enrolment on the learnership, qualification or unit standard. The data mining activities focuses on gaining a better understanding of data records that fall into specific categories of the data field PROV\_IND (see Appendix C.3.5, Appendix E.3.6 and Appendix G.3.6) and the possible identification of anomalous data records in the respective data sets.

This semantic business rule defines that a provider must be accredited for the duration of the learner's active enrolment and is applicable to learnership, qualification and unit standard enrolments. As a result the structure of this appendix has sub sections that focus on the specific data mining activities per specific categories for each of these types of enrolment records.

### ***K.1 Learnership enrolment***

#### ***K.1.1 Start Before, End Before or End During cluster data mining***

This section provides a technical description of the clusters that were generated by cluster data mining the consolidated data category 'Start Before, End Before or End During' (see Appendix J.1.7) for learnership enrolment records. A detailed description of the manner in which cluster data mining was conducted can be found in Appendix I.3. The results of the generated clustering model were significant because the model was measured as being 99.25% accurate.

The generated clusters show a tight coupling between data fields that describe the ETQEs, learnerships and providers. This is as a result of the organic relationship between learnerships and ETQEs (learnerships are generally implemented by one ETQE only) and providers and ETQEs (providers generally offer learnerships that are implemented by their primary ETQE).

The model did not generate any clusters that contained less than 1% of the records. The detection of anomalous records based on cluster specific probabilities of less than .6 being allocated to the record was conducted. In this manner it was determined that 0.29% of the records possibly exist in this category as a result of data capturing problems (see Appendix I.3).

The technical description of each of the 8 clusters is provided below. The description is limited to the top 10 attributes of the cluster rule, where each indent of text represents the importance of the attribute. For example:

A line formatted like this represents an importance of 100%

*A line formatted like this represents an importance greater than 75% and less than 100%*

*A line formatted like this represents an importance greater than 50% and less than or equal to 75%*

*A line formatted like this represents an importance less than or equal to 50%*

#### 1. Cluster 1

% of records: 18.38%

Average probability: 0.9993

Rule:

PRIMARY\_ETQE\_DESC IN ('Not Primary ETQE of provider')

*AND PROVIDER\_ID IN ('48591', '48475', '46423', '21780', '21778', '21389',  
'21375', '21337', '21328', '21318')*

*AND LEARNERSHIP\_ID IN ('932', '744', '714', '692', '291', '285', '284', '240',  
'1374', '1303', '1274', '1269', '1242')*

*AND PROV\_ETQE\_ID IN ('1033')*

*AND ASSESSOR\_ID IN ('NULL', '8145591', '8145095', '4632260', '4282957',  
'4282924', '4282714', '4282614', '3557298', '3040050', '3027229', '3008406')*

*AND LSHP\_ETQE\_ID IN ('1106', '1103')*

*AND ETQE\_ID IN ('1106', '1103')*

*AND PROVIDER\_TYPE\_DESC IN ('Education')*

*AND PROVIDER\_CLASS\_DESC IN ('Public')*

*AND PROV\_PROVINCE\_DESC IN ('Western Cape', 'North West', 'Eastern Cape')*

#### 2. Cluster 2

% of records: 16.90%

Average probability: 0.9938

Rule:

*PROVIDER\_ID IN ('50456', '46459', '44143', '38615', '38601', '38596', '38591',  
'38579', '38574', '38570', '38569', '38564', '38563', '38560', '38555', '38554',  
'38552', '38551', '38549', '38542')*

*AND ASSESSOR\_ID IN ('NULL')*

*AND LSHP\_ETQE\_ID IN ('1111')*  
*AND LEARNERSHIP\_ID IN ('794', '676', '668', '656', '655', '653', '651', '650',*  
*'645', '643', '642', '641', '640', '637', '631', '628', '627', '626', '623', '619', '618',*  
*'615', '614', '611', '610', '1537', '1483', '1476', '1466', '1465', '1463', '1460')*  
*AND ETQE\_ID IN ('1111')*  
*AND PROV\_ETQE\_ID IN ('1111')*  
 AND PROVIDER\_TYPE\_DESC IN ('Training')  
 AND PROVIDER\_CLASS\_DESC IN ('Private')  
 AND -12.8 <= END\_PROV\_IND <= 0  
 AND PROV\_PROVINCE\_DESC IN ('North West', 'Mpumalanga', 'Gauteng')

### 3. Cluster 3

% of records: 16.08%

Average probability: 0.9970

Rule:

*PROVIDER\_ID IN ('47993', '47992')*  
*AND LEARNERSHIP\_ID IN ('1554')*  
*AND ASSESSOR\_ID IN ('NULL')*  
*AND LSHP\_ETQE\_ID IN ('1105')*  
*AND ETQE\_ID IN ('1105')*  
*AND PROV\_ETQE\_ID IN ('1105')*  
*AND NQF\_LEVEL\_DESC IN ('Level 4')*  
 AND PROV\_PROVINCE\_DESC IN ('South Africa National')  
 AND PROVIDER\_TYPE\_DESC IN ('Education and Training')  
 AND PROVIDER\_CLASS\_DESC IN ('Mixed: Public and Private')

### 4. Cluster 4

% of records: 12.08%

Average probability: 0.9998

Rule:

*PROVIDER\_ID IN ('49153')*  
*AND LEARNERSHIP\_ID IN ('899', '895', '894', '892', '888', '884', '877')*  
*AND ASSESSOR\_ID IN ('NULL')*  
*AND LSHP\_ETQE\_ID IN ('1126')*  
*AND ETQE\_ID IN ('1126')*

*AND PROV\_ETQE\_ID IN ('1126')*  
*AND PROV\_PROVINCE\_DESC IN ('Gauteng')*  
*AND PROVIDER\_TYPE\_DESC IN ('Training')*  
*AND PROVIDER\_CLASS\_DESC IN ('Mixed: Public and Private')*  
*AND NQF\_LEVEL\_DESC IN ('Level 4', 'Level 2')*

5. Cluster 5

% of records: 11.50%

Average probability: 0.9814

Rule:

*PROVIDER\_CLASS\_DESC IN ('Unknown')*  
*AND PROVIDER\_ID IN ('50114', '46460', '42963', '42946', '39919', '39897', '39890', '37405')*  
*AND LEARNERSHIP\_ID IN ('932', '899', '895', '894', '893', '889', '888', '884', '523', '285', '1374', '1031')*  
*AND PROV\_ETQE\_ID IN ('1126')*  
*AND PROV\_PROVINCE\_DESC IN ('Mpumalanga', 'Gauteng', 'Free State')*  
*AND ASSESSOR\_ID IN ('NULL', '3018856')*  
*AND ETQE\_ID IN ('1127', '1126', '1120', '1106')*  
*AND LSHP\_ETQE\_ID IN ('1127', '1126', '1106')*  
*AND PROVIDER\_TYPE\_DESC IN ('Education and Training')*  
*AND NQF\_LEVEL\_DESC IN ('Level 4', 'Level 2')*

6. Cluster 6

% of records: 11.48%

Average probability: 0.9880

Rule:

*PROVIDER\_ID IN ('48053', '48035', '47922', '47780', '47767', '47651', '47631', '47510', '47468', '47425', '47358', '47278', '47082', '46763', '46657', '46605', '45181', '39834', '39833', '37520', '29436')*  
*AND LEARNERSHIP\_ID IN ('784', '781', '778', '776', '774', '740', '697', '529', '1139')*  
*AND ASSESSOR\_ID IN ('NULL')*  
*AND PROV\_ETQE\_ID IN ('1105', '1103')*  
*AND ETQE\_ID IN ('1127', '1105', '1103')*

AND PROV\_PROVINCE\_DESC IN ('South Africa National')  
AND LSHP\_ETQE\_ID IN ('1127', '1103', '1005')  
AND PROVIDER\_TYPE\_DESC IN ('Education and Training')  
AND PROVIDER\_CLASS\_DESC IN ('Mixed: Public and Private')  
AND 84.3 <= END\_DATE\_IND <= 112.2

#### 7. Cluster 7

% of records: 7.84%

Average probability: 0.9833

Rule:

*LEARNERSHIP\_ID* IN ('99', '98', '97', '96', '95', '94', '93', '88', '744', '717', '465', '39',  
'337', '24', '23', '138', '1327', '112', '111', '110', '107', '104', '103', '102', '101')  
AND *PROVIDER\_ID* IN ('50009', '49726', '49362', '46460', '41565', '41557', '41451',  
'41444', '37970', '37967', '34946', '32807', '26233', '26218', '24850', '21917', '21910',  
'21909', '11100', '11088', '11087', '11069', '11067', '11063')  
AND *ASSESSOR\_ID* IN ('NULL')  
AND *PROV\_ETQE\_ID* IN ('1116', '1115')  
AND *LSHP\_ETQE\_ID* IN ('1116', '1115', '1103')  
AND *ETQE\_ID* IN ('1116', '1115', '1103')  
AND *PROV\_PROVINCE\_DESC* IN ('Western Cape', 'Kwazulu/Natal', 'Gauteng')  
AND *NQF\_LEVEL\_DESC* IN ('Level 7', 'Level 4', 'Level 3', 'Level 2', 'Level 1')  
AND *ENROL\_STATUS\_DESC* IN ('Enrolled')  
AND *PROVIDER\_TYPE\_DESC* IN ('Training', 'Employer', 'Education and Training')

#### 8. Cluster 8

% of records: 5.73%

Average probability: 0.9836

Rule:

*LEARNERSHIP\_ID* IN ('39', '1085', '1080', '1079', '1072', '1070', '1069')  
AND *PROVIDER\_ID* IN ('50240', '49951', '49947', '49943', '49940', '24850', '21830')  
AND *ASSESSOR\_ID* IN ('NULL')  
AND *LSHP\_ETQE\_ID* IN ('1114')  
AND 123.6 <= *START\_DATE\_IND* <= 149.4  
AND *ETQE\_ID* IN ('1114')  
AND *PROV\_ETQE\_ID* IN ('1114')

AND PROVIDER\_TYPE\_DESC IN ('Education and Training')  
AND PROVIDER\_CLASS\_DESC IN ('Private')  
AND 121.5 <= END\_DATE\_IND <= 158.7

### ***K.1.2 Start During, Start After and End After cluster data mining***

This section provides a technical description of the clusters that were generated by cluster data mining the consolidated data category 'Start During, Start After or End After' (see Appendix J.1.8) for learnership enrolment records. A detailed description of the manner in which cluster data mining was conducted can be found in Appendix I.3.

The results of the generated clustering model are significant because the model was measured as being 94.68% accurate. All of the clusters show a tight coupling between data fields that describe the ETQEs, learnerships and providers. This is as a result of the organic relationship between learnerships and ETQEs (learnerships are generally implemented by one ETQE only) and providers and ETQEs (providers generally offer learnerships that are implemented by their primary ETQE).

The model did not generate any clusters that contained less than 1% of the records. The detection of anomalous records based on cluster specific probabilities of less than .6 being allocated to the record was conducted. In this manner it was determined that 3.89% of the records possibly exist in this category as a result of data capturing problems (see Appendix I.3).

The technical description of each of the 8 clusters is provided below. The description is limited to the top 10 attributes of the cluster rule, where each indent of text represents the importance of the attribute. For example:

A line formatted like this represents an importance of 100%

*A line formatted like this represents an importance greater than 75% and less than 100%*

*A line formatted like this represents an importance greater than 50% and less than or equal to 75%*

*A line formatted like this represents an importance less than or equal to 50%*

- Cluster 1

% of records: 27.06%

Average probability: 0.9250

Rule:

LEARNERSHIP\_ID IN ('53', '24')  
 AND PROVIDER\_ID IN ('24763', '24758', '24735', '24728', '24712', '2471', '24671',  
 '24642', '24630', '24629', '24627', '24620', '24598', '24586', '24576', '24545', '24514',  
 '24511', '24510', '2424', '2421', '2391', '2263', '1905')  
 AND LSHP\_ETQE\_ID IN ('1120', '1116')  
 AND ETQE\_ID IN ('1120', '1116')  
 AND PROV\_ETQE\_ID IN ('1120', '1116')  
 AND 89.8 <= START\_DATE\_IND <= 131.8  
 AND 31 <= END\_PROV\_IND <= 91  
 AND NQF\_LEVEL\_DESC IN ('Level 7', 'Level 5')  
 AND PROV\_PROVINCE\_DESC IN ('Western Cape', 'Gauteng')  
 AND PROVIDER\_CLASS\_DESC IN ('Private')

- Cluster 2

% of records: 22.24%

Average probability: 0.9951

Rule:

PROVIDER\_ID IN ('49723', '37631')  
 AND LEARNERSHIP\_ID IN ('460')  
 AND LSHP\_ETQE\_ID IN ('1117')  
 AND ETQE\_ID IN ('1117')  
 AND PROV\_ETQE\_ID IN ('1117')  
 AND NQF\_LEVEL\_DESC IN ('Level 1')  
 AND 115.6 <= END\_DATE\_IND <= 133.4  
 AND 0 <= START\_PROV\_IND <= 12.7  
 AND PROV\_PROVINCE\_DESC IN ('Gauteng', 'Eastern Cape')  
 AND PROVIDER\_CLASS\_DESC IN ('Public')

- Cluster 3

% of records: 17.80%

Average probability: 0.9174

Rule:

PROVIDER\_ID IN ('5994', '5866', '50428', '48693', '41579', '41566', '41459', '38591',  
 '38548', '37832', '33735', '29980', '29278', '29275', '29271', '29269', '29257', '29242',  
 '29240', '29237', '26218', '25429', '24798', '24642', '24641', '2231', '20603', '20589',

'20579', '20357', '1982', '1974', '17119', '14883', '12332', '11101', '11090', '11088',  
 '11087', '11072', '11059', '10782')  
 AND LEARNERSHIP\_ID IN ('99', '96', '95', '929', '892', '825', '801', '800', '714', '641',  
 '640', '528', '518', '515', '351', '350', '341', '322', '321', '320', '313', '306', '305', '24',  
 '1463', '1450', '1325', '1274', '1270', '1269', '1263', '111', '107', '102', '101')  
 AND ETQE\_ID IN ('1127', '1126', '1125', '1116', '1115', '1112', '1111', '1107', '1103')  
 AND PROVIDER\_TYPE\_DESC IN ('Training', 'Employer')  
 AND PROV\_ETQE\_ID IN ('1127', '1126', '1125', '1116', '1115', '1112', '1111', '1107', '1103')  
 AND LSHP\_ETQE\_ID IN ('1127', '1126', '1125', '1116', '1115', '1111', '1107', '1103')  
 AND ASSESSOR\_ID IN ('NULL')  
 AND 0 <= START\_PROV\_IND <= 38.1  
 AND PROV\_PROVINCE\_DESC IN ('Western Cape', 'Undefined', 'Mpumalanga',  
 'Kwazulu/Natal', 'Gauteng')  
 AND NQF\_LEVEL\_DESC IN ('Level 7', 'Level 4', 'Level 3', 'Level 2')

- Cluster 4

% of records: 8.95%

Average probability: 0.9458

Rule:

PRIMARY\_ETQE\_DESC IN ('Not Primary ETQE of provider')  
 AND PROVIDER\_ID IN ('595', '41517', '36340', '29316', '26365', '22745', '2159',  
 '1945', '11772')  
 AND LEARNERSHIP\_ID IN ('97', '900', '899', '873', '462', '461', '288', '240', '189',  
 '1374', '1333', '1242', '1133', '1131', '107')  
 AND PROV\_ETQE\_ID IN ('1126', '1125', '1031')  
 AND 1 <= END\_PROV\_IND <= 16  
 AND PROVIDER\_CLASS\_DESC IN ('Unknown', 'Private')  
 AND 0 <= START\_PROV\_IND <= 12.7  
 AND ETQE\_ID IN ('1126', '1117', '1116', '1115', '1114', '1109', '1106', '1103')  
 AND LSHP\_ETQE\_ID IN ('1126', '1117', '1115', '1109', '1106', '1103')  
 AND ASSESSOR\_ID IN ('NULL')

- Cluster 5

% of records: 7.31%

Average probability: 0.9185

Rule:

*PROVIDER\_ID IN ('5093', '49916', '48792', '43440', '41611', '39735', '36196',  
 '34674', '30204', '29441', '2916', '28710', '2748', '2741', '2627', '22974', '22903',  
 '22060', '22029', '21942', '20357', '19858')*  
*AND LEARNERSHIP\_ID IN ('965', '961', '958', '906', '804', '801', '764', '762', '744',  
 '735', '702', '523', '1529', '1450', '1437', '1407', '1269', '1266', '1197', '1147', '1141',  
 '1139')*  
*AND LSHP\_ETQE\_ID IN ('1127', '1112', '1103', '1102')*  
*AND ETQE\_ID IN ('1127', '1112', '1103', '1102')*  
*AND PROV\_ETQE\_ID IN ('1127', '1103', '1102')*  
*AND PROVIDER\_CLASS\_DESC IN ('Mixed: Public and Private')*  
*AND ASSESSOR\_ID IN ('NULL')*  
*AND 1 <= END\_PROV\_IND <= 31*  
*AND ENROL\_STATUS\_DESC IN ('Enrolled')*  
*AND PROVIDER\_TYPE\_DESC IN ('Education and Training')*

- Cluster 6

% of records: 6.50%

Average probability: 0.9859

Rule:

*PRIMARY\_ETQE\_DESC IN ('Not Primary ETQE of provider')*  
*AND PROVIDER\_ID IN ('41566', '22910')*  
*AND LEARNERSHIP\_ID IN ('895', '894', '1462', '1461', '1460', '1459')*  
*AND ETQE\_ID IN ('1121', '1111')*  
*AND LSHP\_ETQE\_ID IN ('1126', '1111')*  
*AND PROV\_ETQE\_ID IN ('1115', '1103')*  
*AND 0 <= START\_PROV\_IND <= 12.7*  
*AND PROV\_PROVINCE\_DESC IN ('South Africa National', 'Mpumalanga')*  
*AND PROVIDER\_CLASS\_DESC IN ('Mixed: Public and Private')*  
*AND 88.9 <= END\_DATE\_IND <= 169*

- Cluster 7

% of records: 6.17%

Average probability: 0.9282

Rule:

*PROVIDER\_ID IN ('48868', '37505', '35551', '35516', '34674', '28473', '21942', '20725', '19858', '10670')*  
*AND LEARNERSHIP\_ID IN ('958', '744', '1529', '1374', '1139')*  
*AND ASSESSOR\_ID IN ('NULL', '3585132', '3585129', '3557207', '3020580', '3015474')*  
     *AND PROV\_ETQE\_ID IN ('1106', '1105', '1103', '1102')*  
     *AND LSHP\_ETQE\_ID IN ('1106', '1105', '1103', '1102')*  
     *AND 0 <= START\_PROV\_IND <= 12.7*  
     *AND PROVIDER\_CLASS\_DESC IN ('Mixed: Public and Private')*  
     *AND ETQE\_ID IN ('1106', '1105', '1103', '1102')*  
     *AND PROV\_IND\_DESC IN ('Start During, End After')*  
         *AND NQF\_LEVEL\_DESC IN ('Level 4', 'Level 3', 'Level 1')*

- Cluster 8

% of records: 3.96%

Average probability: 0.9752

Rule:

*PRIMARY\_ETQE\_DESC IN ('Not Primary ETQE of provider')*  
     *AND PROVIDER\_ID IN ('34574', '30204', '21942')*  
     *AND LEARNERSHIP\_ID IN ('799', '776', '285', '247', '1266')*  
     *AND PROV\_PROVINCE\_DESC IN ('Free State', 'Eastern Cape')*  
     *AND LSHP\_ETQE\_ID IN ('1106', '1103')*  
     *AND ETQE\_ID IN ('1106', '1103')*  
     *AND 73 <= START\_DATE\_IND <= 131.8*  
         *AND ASSESSOR\_ID IN ('NULL', '4282945', '4282588')*  
         *AND PROV\_ETQE\_ID IN ('1115', '1103', '1102')*  
         *AND NQF\_LEVEL\_DESC IN ('Level 5', 'Level 2')*

## ***K.2 Qualification enrolment***

### ***K.2.1 Start Before, End Before or End During cluster data mining***

This section provides a technical description of the clusters that were generated by cluster data mining the consolidated data category ‘Start Before, End Before or End During’ (see Appendix J.2.7) for qualification enrolment records. A detailed description of the manner in which cluster data mining was conducted can be found in Appendix I.3. The results of the generated clustering model were significant because the model was measured as being 98.55% accurate.

The generated clusters show a tight coupling between data fields that describe the ETQEs, qualifications and providers. This is as a result of the organic relationship between qualifications and ETQEs (qualifications are generally implemented by one ETQE only) and providers and ETQEs (providers generally offer qualifications that are implemented by their primary ETQE).

The model did not generate any clusters that contained less than 1% of the records. The detection of anomalous records based on cluster specific probabilities of less than .6 being allocated to the record was conducted. In this manner it was determined that 0.79% of the records possibly exist in this category as a result of data capturing problems (see Appendix I.3).

The technical description of each of the 8 clusters is provided below. The description is limited to the top 10 attributes of the cluster rule, where each indent of text represents the importance of the attribute. For example:

A line formatted like this represents an importance of 100%

*A line formatted like this represents an importance greater than 75% and less than 100%*

*A line formatted like this represents an importance greater than 50% and less than or equal to 75%*

*A line formatted like this represents an importance less than or equal to 50%*

- Cluster 1

% of records: 20.73%

Average probability: 0.9950

Rule:

LEARNERSHIP\_ID IN ('NULL')

*AND QUALIFICATION\_ID IN ('57625', '50139', '24214', '22507')*  
*AND PROVIDER\_ID IN ('48296', '48295', '48293', '48150', '47993', '47992', '47780', '47651', '47468', '47283', '47215', '47082', '46944', '46763', '37520', '23172')*  
*AND ASSESSOR\_ID IN ('NULL')*  
*AND SUBFIELD\_DESC IN ('Safety in Society')*  
*AND PROV\_ETQE\_ID IN ('1105')*  
*AND ETQE\_ID IN ('1105')*  
*AND FIELD\_DESC IN ('Law, Military Science and Security')*  
*AND PROVIDER\_TYPE\_DESC IN ('Education and Training')*  
*AND PROVIDER\_CLASS\_DESC IN ('Mixed: Public and Private')*

- Cluster 2

% of records: 18.82%

Average probability: 0.9991

Rule:

*LEARNERSHIP\_ID IN ('NULL', '285', '1374')*  
*AND PRIMARY\_ETQE\_DESC IN ('Not Primary ETQE of provider')*  
*AND PROVIDER\_ID IN ('50578', '50114', '21780', '21778', '21337', '21328', '21318')*  
*AND QUALIFICATION\_ID IN ('58778', '23135', '23134', '23133', '23131')*  
*AND SUBFIELD\_DESC IN ('Early Childhood Development')*  
*AND ETQE\_ID IN ('1106')*  
*AND ASSESSOR\_ID IN ('NULL', '9464189', '8145591', '4632260', '4632257', '4631855', '4282957', '4282924', '4282714', '4282640', '4282614', '3557298', '3040050', '3029299', '3027363', '3027229', '3027166', '3018856', '3018255', '3008406')*  
*AND FIELD\_DESC IN ('Education, Training and Development')*  
*AND PROV\_PROVINCE\_DESC IN ('Western Cape', 'North West', 'Free State')*  
*AND PROVIDER\_CLASS\_DESC IN ('Unknown', 'Public')*

- Cluster 3

% of records: 13.20%

Average probability: 0.9866

Rule:

```
ASSESSOR_ID IN ('NULL', '3454471')
AND LEARNERSHIP_ID IN ('NULL', '801', '794', '778', '744', '523', '483', '39',
'240', '1376', '1242', '1085', '1080', '1079', '1077', '1072', '1070', '1069', '1036')
AND PROVIDER_ID IN ('51249', '51225', '51199', '51195', '51191', '51153',
'51145', '51138', '50578', '50489', '50245', '50240', '49951', '49947', '49943',
'49940', '49726', '49724', '48591', '48475', '48375', '48370', '46784', '46460',
'46423', '45504', '45299', '44143', '42946', '41565', '41451', '40194', '40149',
'39834', '39833', '39490', '39104', '39043', '38552', '38476', '37727', '34337',
'33741', '32807', '31909', '29436', '24850', '22236', '21830', '21665', '21650',
'21300', '11087')
AND QUALIFICATION_ID IN ('78982', '78981', '64827', '63426', '63351',
'61467', '58798', '58411', '58393', '58223', '50351', '50302', '49706', '49666',
'49623', '49414', '49148', '49108', '49094', '49038', '49035', '49026', '48989',
'48889', '48778', '48490', '24473', '24290', '24010', '23910', '23290', '23272',
'23271', '23270', '22875', '21032', '21021', '20307', '20211', '20190', '13670')
AND ETQE_ID IN ('1127', '1123', '1117', '1116', '1114', '1113', '1112', '1103',
'1102')
    AND PROV_ETQE_ID IN ('1127', '1123', '1117', '1116', '1115', '1114',
'1113', '1103', '1102', '1033')
    AND PROV_PROVINCE_DESC IN ('Western Cape', 'Limpopo',
'Kwazulu/Natal', 'Gauteng', 'Eastern Cape')
    AND FIELD_DESC IN ('Services', 'Physical, Mathematical, Computer and
Life Sciences', 'Manufacturing, Engineering and Technology', 'Health
Sciences and Social Services', 'Business, Commerce and Management
Studies')
    AND SUBFIELD_DESC IN ('Wholesale and Retail', 'Transport, Operations
and Logistics', 'Promotive Health and Developmental Services', 'Primary
Agriculture', 'Office Administration', 'Manufacturing and Assembly',
'Information Technology and Computer Sciences', 'Finance, Economics and
Accounting', 'Engineering and Related Design', 'Building Construction')
        AND ENROL_TYPE_DESC IN ('Mixed Mode')
```

- Cluster 4

% of records: 10.08%

Average probability: 0.9436

Rule:

*LEARNERSHIP\_ID IN ('NULL', '888')*

*AND PROVIDER\_ID IN ('51339', '50664', '49153', '49045', '44779', '42989', '42963', '42954', '42946', '41493', '39919', '39897', '38476', '37405', '37395', '37392', '36373', '36326', '35735', '35675', '11464')*

*AND ASSESSOR\_ID IN ('NULL')*

*AND QUALIFICATION\_ID IN ('73286', '66266', '61772', '59114', '57954', '50097', '49709', '49708', '35945', '23970', '23673', '23671', '21810', '21808', '20924', '20202')*

*AND PROV\_ETQE\_ID IN ('1126')*

*AND ETQE\_ID IN ('1126')*

*AND ENROL\_TYPE\_DESC IN ('Work Place Learning', 'Residential Learning (i.e. Contact Mode)', 'RPL for Unknown Purpose')*

*AND SUBFIELD\_DESC IN ('Marketing', 'Generic Management', 'Finance, Economics and Accounting', 'Cleaning, Domestic, Hiring, Property and Rescue Services')*

*AND -18.8 <= START\_PROV\_IND <= -1*

*AND FIELD\_DESC IN ('Services', 'Business, Commerce and Management Studies')*

- Cluster 5

% of records: 10.05%

Average probability: 0.9815

Rule:

*LEARNERSHIP\_ID IN ('NULL')*

*AND QUALIFICATION\_ID IN ('58393', '58392')*

*AND PROVIDER\_ID IN ('37975')*

*AND ASSESSOR\_ID IN ('NULL')*

*AND SUBFIELD\_DESC IN ('Finance, Economics and Accounting')*

*AND PROV\_ETQE\_ID IN ('1116')*

*AND ETQE\_ID IN ('1116')*

*AND FIELD\_DESC IN ('Business, Commerce and Management Studies')*

*AND NQF\_LEVEL\_DESC IN ('Level 3')*  
*AND PROVIDER\_TYPE\_DESC IN ('Employer')*

- Cluster 6

% of records: 9.54%

Average probability: 0.9974

Rule:

*QUALIFICATION\_ID IN ('59868', '58777', '58756', '58284', '49031', '49030', '21861', '21832', '21828')*  
*AND PROVIDER\_ID IN ('50456', '46459', '38660', '38601', '38596', '38579', '38574', '38563', '38561', '38560', '38555', '38554', '38551', '38549')*  
*AND LEARNERSHIP\_ID IN ('NULL', '1465', '1463')*  
*AND ASSESSOR\_ID IN ('NULL')*  
*AND SUBFIELD\_DESC IN ('Fabrication and Extraction')*  
*AND PROV\_ETQE\_ID IN ('1111')*  
*AND ETQE\_ID IN ('1111')*  
*AND FIELD\_DESC IN ('Manufacturing, Engineering and Technology')*  
*AND PROV\_PROVINCE\_DESC IN ('North West', 'Limpopo', 'Gauteng')*  
*AND PROVIDER\_TYPE\_DESC IN ('Training')*

- Cluster 7

% of records: 8.95%

Average probability: 0.9805

Rule:

*LEARNERSHIP\_ID IN ('NULL')*  
*AND QUALIFICATION\_ID IN ('58393', '58392', '36230', '20408', '20375')*  
*AND PROVIDER\_ID IN ('50009', '49342', '37975')*  
*AND ASSESSOR\_ID IN ('NULL')*  
*AND SUBFIELD\_DESC IN ('Finance, Economics and Accounting')*  
*AND PROV\_ETQE\_ID IN ('1116')*  
*AND ETQE\_ID IN ('1116')*  
*AND FIELD\_DESC IN ('Business, Commerce and Management Studies')*  
*AND NQF\_LEVEL\_DESC IN ('Level 3')*  
*AND -7.8 <= END\_PROV\_IND <= 0*

- Cluster 8

% of records: 8.63%

Average probability: 0.9748

Rule:

PROVIDER\_ID IN ('49153', '37392')

AND LEARNERSHIP\_ID IN ('NULL', '899', '888')

AND QUALIFICATION\_ID IN ('59114', '35945', '24510', '23850', '23673', '23671', '21810', '20924', '20202')

AND ASSESSOR\_ID IN ('NULL')

AND PROV\_ETQE\_ID IN ('1126')

AND ETQE\_ID IN ('1126')

AND FIELD\_DESC IN ('Business, Commerce and Management Studies')

AND ENROL\_TYPE\_DESC IN ('Unknown', 'Residential Learning (i.e. Contact Mode)', 'RPL for Unknown Purpose')

AND PROVIDER\_TYPE\_DESC IN ('Training')

AND SUBFIELD\_DESC IN ('Office Administration', 'Marketing', 'Generic Management', 'Cleaning, Domestic, Hiring, Property and Rescue Services')

### ***K.2.2 Start During, Start After and End After cluster data mining***

This section provides a technical description of the clusters that were generated by cluster data mining the consolidated data category 'Start During, Start After and End After' (see Appendix J.2.8) for qualification enrolment records. A detailed description of the manner in which cluster data mining was conducted can be found in Appendix I.3. The results of the generated clustering model were significant because the model was measured as being 98.59% accurate.

The generated clusters show a tight coupling between data fields that describe the ETQEs, qualifications and providers. This is as a result of the organic relationship between qualifications and ETQEs (qualifications are generally implemented by one ETQE only) and providers and ETQEs (providers generally offer qualifications that are implemented by their primary ETQE).

The model did not generate any clusters that contained less than 1% of the records. The detection of anomalous records based on cluster specific probabilities of less than .6 being allocated to the record was conducted. In this manner it was determined that 0.15% of the records possibly exist in this category as a result of data capturing problems (see Appendix I.3).

The technical description of each of the 8 clusters is provided below. The description is limited to the top 10 attributes of the cluster rule, where each indent of text represents the importance of the attribute. For example:

A line formatted like this represents an importance of 100%

*A line formatted like this represents an importance greater than 75% and less than 100%*

*A line formatted like this represents an importance greater than 50% and less than or equal to 75%*

*A line formatted like this represents an importance less than or equal to 50%*

- Cluster 1

% of records: 33.59%

Average probability: 0.9913

Rule:

*LEARNERSHIP\_ID IN ('NULL')*

*AND PROVIDER\_ID IN ('39996', '34574', '31724', '30217', '28446', '28261', '28222', '28156', '26373', '11772')*

*AND QUALIFICATION\_ID IN ('74647', '59293', '58778', '50351', '49665', '35945', '23970', '23850', '23672', '23671', '23135', '20924', '20911')*

*AND ETQE\_ID IN ('1126', '1106')*

*AND ASSESSOR\_ID IN ('NULL', '8145429', '8145122', '5518282', '4283216', '3027298')*

*AND PROV\_ETQE\_ID IN ('1126', '1106', '1103', '1031')*

*AND PROVIDER\_TYPE\_DESC IN ('Education and Training')*

*AND SUBFIELD\_DESC IN ('Office Administration', 'Marketing', 'Generic Management', 'Early Childhood Development')*

*AND FIELD\_DESC IN ('Education, Training and Development', 'Business, Commerce and Management Studies')*

*AND NQF\_LEVEL\_DESC IN ('Level 4', 'Level 2')*

- Cluster 2

% of records: 19.99%

Average probability: 0.9699

Rule:

*LEARNERSHIP\_ID IN ('NULL')*

*AND PROVIDER\_ID IN ('51338', '50008', '48792', '41493', '37520', '37171', '28156', '25429', '22927', '1974', '1915', '18575', '1726', '1578')*

*AND QUALIFICATION\_ID IN ('74647', '59768', '57841', '57625', '49946', '49852', '49709', '49708', '48930')*

*AND ASSESSOR\_ID IN ('NULL')*

*AND SUBFIELD\_DESC IN ('Human Resources', 'Finance, Economics and Accounting')*

*AND FIELD\_DESC IN ('Business, Commerce and Management Studies')*

*AND ETQE\_ID IN ('1127', '1126', '1075')*

*AND PROV\_ETQE\_ID IN ('1127', '1119', '1116', '1110', '1106', '1103', '1079', '1075', '1033', '1031')*

*AND NQF\_LEVEL\_DESC IN ('Level 5', 'Level 4')*

*AND PROVIDER\_TYPE\_DESC IN ('Training', 'Education and Training', 'Education')*

- Cluster 3

% of records: 9.73%

Average probability: 0.9867

Rule:

*PROVIDER\_ID IN ('6970', '43397', '41566', '36340', '34674', '30204', '29441', '28710', '2748', '2741', '22974', '22029', '20357', '19858', '1945', '11772')*

*AND QUALIFICATION\_ID IN ('64866', '64827', '64826', '60312', '60311', '60310', '59317', '58802', '58798', '58244', '57848', '50601', '50302', '48993', '48987', '48815', '48812', '24290', '23270', '22787', '22459', '21829', '21029', '20830', '20730')*

*AND ETQE\_ID IN ('1111', '1109', '1103', '1102')*

*AND ASSESSOR\_ID IN ('NULL')*

*AND FIELD\_DESC IN ('Physical Planning and Construction', 'Manufacturing, Engineering and Technology')*

*AND SUBFIELD\_DESC IN ('Transport, Operations and Logistics', 'Manufacturing and Assembly', 'Fabrication and Extraction', 'Engineering and Related Design', 'Building Construction')*

*AND LEARNERSHIP\_ID IN ('NULL', '958', '804', '799', '764', '762', '744', '702', '683', '240', '189', '1462', '1461', '1460', '1459', '1450', '1274', '1270', '1269', '1266', '1243', '1242', '1147', '1141', '1139')*

*AND 0 <= START\_PROV\_IND <= 12.7*

*AND PROV\_ETQE\_ID IN ('1126', '1115', '1103', '1102', '1031')*

*AND ENROL\_TYPE\_DESC IN ('Mixed Mode')*

- Cluster 4

% of records: 9.36%

Average probability: 1.0000

Rule:

*QUALIFICATION\_ID IN ('60006', '58594', '50139')*

*AND PROVIDER\_ID IN ('46926', '46877', '38426', '35551', '35516', '30139', '20772', '20725', '2071')*

*AND LEARNERSHIP\_ID IN ('NULL')*

*AND PROV\_ETQE\_ID IN ('1105')*

*AND SUBFIELD\_DESC IN ('Safety in Society')*

*AND ETQE\_ID IN ('1105')*

*AND ASSESSOR\_ID IN ('NULL')*

*AND PROV\_PROVINCE\_DESC IN ('South Africa National')*

*AND FIELD\_DESC IN ('Law, Military Science and Security')*

*AND 0 <= START\_PROV\_IND <= 12.7*

- Cluster 5

% of records: 9.27%

Average probability: 0.9827

Rule:

*PROVIDER\_ID IN ('48693', '44729', '44486', '42050', '41517', '41483', '41474', '41471', '41460', '41459', '41452', '41443', '41438', '40925', '40875', '40623', '39490', '38616', '38548', '36061', '31726', '29278', '29275', '29273', '29271', '29269', '29260', '29257', '29242', '29240', '29237', '20642', '14883', '12332', '11138')*

*AND LEARNERSHIP\_ID IN ('NULL', '945', '892', '322', '321', '320', '305')*  
*AND QUALIFICATION\_ID IN ('67452', '67447', '61587', '61586', '61566', '59406',*  
*'58756', '58043', '49094', '49030', '48982', '48492', '48491', '48490', '23850', '23695',*  
*'23694', '21828', '20830', '14128')*  
*AND ASSESSOR\_ID IN ('NULL')*  
*AND PROV\_ETQE\_ID IN ('1126', '1125', '1122', '1111', '1110', '1107')*  
*AND SUBFIELD\_DESC IN ('Public Administration', 'Nature Conservation',*  
*'Manufacturing and Assembly', 'Hospitality, Tourism, Travel, Gaming and Leisure',*  
*'Fabrication and Extraction', 'Electrical Infrastructure Construction')*  
*AND ETQE\_ID IN ('1126', '1125', '1122', '1111', '1110', '1107')*  
*AND ENROL\_TYPE\_DESC IN ('Unknown', 'Mixed Mode')*  
*AND 0 <= START\_PROV\_IND <= 12.7*  
*AND PROVIDER\_CLASS\_DESC IN ('Private')*

- Cluster 6

% of records: 8.53%

Average probability: 1.0000

Rule:

*LEARNERSHIP\_ID IN ('53')*  
*AND QUALIFICATION\_CLASS\_DESC IN ('Regular-ELOAC')*  
*AND QUALIFICATION\_ID IN ('48550')*  
*AND PROVIDER\_ID IN ('1905')*  
*AND PROV\_ETQE\_ID IN ('1120')*  
*AND SUBFIELD\_DESC IN ('Finance, Economics and Accounting')*  
*AND ETQE\_ID IN ('1120')*  
*AND ASSESSOR\_ID IN ('NULL')*  
*AND PROV\_PROVINCE\_DESC IN ('Gauteng')*  
*AND 46.3 <= END\_PROV\_IND <= 61.4*

- Cluster 7

% of records: 5.64%

Average probability: 0.9590

Rule:

*NQF\_LEVEL\_DESC IN ('Level 7')*

AND QUALIFICATION\_CLASS\_DESC IN ('Regular-Provider-ELOAC')  
 AND QUALIFICATION\_ID IN ('73729', '20409', '20408')  
 AND QUALIFICATION\_TYPE\_DESC IN ('Post Graduate Diploma')  
 AND LEARNERSHIP\_ID IN ('NULL', '24')  
 AND PROV\_ETQE\_ID IN ('1116')  
 AND SUBFIELD\_DESC IN ('Finance, Economics and Accounting')  
 AND ETQE\_ID IN ('1116')  
 AND ASSESSOR\_ID IN ('NULL')  
 AND PROVIDER\_ID IN ('50008', '49473', '49397', '49388', '38015', '24877', '24861',  
 '24842', '24806', '24780', '24775', '24773', '24770', '24769', '24766', '24763', '24758',  
 '24740', '24737', '24735', '24734', '24728', '24720', '24719', '2471', '24687', '24686',  
 '24683', '2468', '24671', '24662', '24661', '24659', '2465', '24647', '24645', '24642',  
 '24639', '24638', '24630', '2463', '24629', '24628', '24627', '24626', '24620', '24616',  
 '24598', '2459', '24586', '24580', '2458', '24578', '24576', '24574', '24572', '24564',  
 '24561', '24560', '24558', '24554', '24547', '24545', '24543', '24537', '24534', '2452',  
 '24519', '24514', '24511', '24510', '2435', '2432', '2427', '2425', '2424', '2422', '2421',  
 '2419', '2414', '2391', '2361', '2346', '2339', '2337', '2335', '2326', '23252', '23248',  
 '23244', '2309', '2306', '2283', '2276', '2274', '2272', '2268', '2263')

- Cluster 8

% of records: 3.89%

Average probability: 0.9988

Rule:

FIELD\_DESC IN ('Health Sciences and Social Services')  
 AND PROVIDER\_ID IN ('39001', '38989', '37747', '2159')  
 AND QUALIFICATION\_ID IN ('49623', '24010')  
 AND LEARNERSHIP\_ID IN ('NULL')  
 AND PROV\_ETQE\_ID IN ('1117')  
 AND SUBFIELD\_DESC IN ('Promotive Health and Developmental Services')  
 AND ETQE\_ID IN ('1117')  
 AND PROV\_PROVINCE\_DESC IN ('Western Cape', 'Kwazulu/Natal')  
 AND ASSESSOR\_ID IN ('NULL')  
 AND NQF\_LEVEL\_DESC IN ('Level 4', 'Level 1')

### ***K.3 Unit Standard enrolment***

#### ***K.3.1 Start Before, End Before or End During cluster data mining***

This section provides a technical description of the clusters that were generated by cluster data mining the consolidated data category ‘Start Before, End Before or End During’ (see Appendix J.3.7) for unit standard enrolment records. A detailed description of the manner in which cluster data mining was conducted can be found in Appendix I.3. The results of the generated clustering model were significant because the model was measured as being 97.72% accurate.

The generated clusters show a tight coupling between data fields that describe the ETQEs, unit standards and providers. This is as a result of the organic relationship between unit standards and ETQEs (unit standards are generally implemented by one ETQE only) and providers and ETQEs (providers generally offer unit standards that are implemented by their primary ETQE).

The model did not generate any clusters that contained less than 1% of the records. The detection of anomalous records based on cluster specific probabilities of less than .6 being allocated to the record was conducted. In this manner it was determined that 3.87% of the records possibly exist in this category as a result of data capturing problems (see Appendix I.3).

The technical description of each of the 8 clusters is provided below. The description is limited to the top 10 attributes of the cluster rule, where each indent of text represents the importance of the attribute. For example:

A line formatted like this represents an importance of 100%

*A line formatted like this represents an importance greater than 75% and less than 100%*

*A line formatted like this represents an importance greater than 50% and less than or equal to 75%*

*A line formatted like this represents an importance less than or equal to 50%*

- Cluster 1

% of records: 33.12%

Average probability: 0.9959

Rule:

ASSESSOR\_ID IN (NULL)

AND PROVIDER\_ID IN ("11088", "38542", "38551", "38554", "38555",  
"38558", "38560", "38561", "38563", "38574", "38575", "38584", "38590",  
"38601", "38602", "38603", "38604", "38607", "38608", "38610", "38621",  
"38659", "38660", "38662", "38676", "46459", "50456")

AND QUALIFICATION\_ID IN ("21828", "49030", "58284", "58756", "58777",  
NULL)

AND UNIT\_STANDARD\_ID IN ("10024", "10025", "10026", "10568", "10570",  
"10572", "10573", "10653", "10757", "10758", "10762", "10804", "110092",  
"110135", "110138", "110139", "110144", "110162", "110234", "110434",  
"110447", "114291", "11490", "115104", "115109", "115110", "115118",  
"115122", "115767", "116454", "116525", "116528", "116537", "116544",  
"116550", "116551", "116653", "116674", "116676", "116687", "116949",  
"119109", "119112", "119115", "119129", "119136", "119137", "119144",  
"119380", "119381", "119384", "119385", "119390", "119392", "119393",  
"119584", "119585", "119648", "119649", "119652", "119653", "119654",  
"119656", "119657", "119980", "119982", "119987", "119988", "12232",  
"12434", "12479", "12488", "12526", "14128", "14673", "14951", "15282",  
"15297", "15301", "15316", "243781", "243789", "243791", "243793", "243794",  
"243796", "243801", "244376", "244377", "244378", "244379", "244380",  
"244381", "244382", "244383", "244384", "244385", "244386", "244387",  
"244388", "244389", "244390", "244391", "244392", "244393", "244394",  
"244395", "244396", "244397", "244398", "244399", "244400", "244401",  
"244402", "244403", "244404", "244405", "244406", "244407", "244408",  
"244409", "244410", "244411", "244412", "244413", "244414", "244415",  
"244416", "244417", "244418", "244419", "244420", "244421", "244422",  
"244423", "244424", "244425", "244426", "244427", "244428", "244429",  
"244430", "244431", "244432", "244433", "244434", "244435", "244436",  
"244437", "244438", "244439", "244440", "244441", "244442", "244443",  
"244444", "244445", "244446", "244447", "244448", "244449", "244450",  
"244451", "244456", "244457", "244458", "244459", "244460", "244461",  
"244462", "244463", "244464", "244465", "244466", "244467", "244468",  
"244469", "244470", "244477", "244478", "244483", "244484", "244491",  
"244493", "244499", "244505", "252611", "252671", "253042", "253813",  
"253832", "253838", "253854", "254594", "7464", "7465", "7467", "7468",

"7478", "7480", "7486", "7519", "7520", "7524", "7525", "7526", "7530",  
 "7543", "7545", "7547", "7552", "7890", "8979", "8980", "9029", "9616",  
 "9617", "9618", "9619", "9695", "9706", "9707", "9708", "9710", "9711",  
 "9717")  
 AND PROV\_ETQE\_ID IN ("1111")  
 AND ETQE\_ID IN ("1111")  
 AND FIELD\_DESC IN ("Manufacturing, Engineering and Technology")  
 AND SUBFIELD\_DESC IN ("Engineering and Related Design", "Fabrication and  
 Extraction", "Mathematical Sciences", "Preventive Health")  
 AND PROVIDER\_TYPE\_DESC IN ("Training")  
 AND PROVIDER\_CLASS\_DESC IN ("Private")

- Cluster 2

% of records: 19.45%

Average probability: 0.9574

Rule:

ASSESSOR\_ID IN ("3013505", NULL)  
 AND PROVIDER\_ID IN ("29600", "37392", "37975", "44880", "46460",  
 "46461", "48790", "48791", "48793", "48794", "49215", "49342", "50014",  
 "51339")  
 AND QUALIFICATION\_ID IN ("23990", "36230", "49666", "49708", "49852",  
 "49946", "57625", "57934", "58392", "58393", NULL)  
 AND UNIT\_STANDARD\_ID IN ("10186", "10187", "10409", "10995", "10997",  
 "10998", "11000", "110545", "113920", "113924", "113926", "113928",  
 "113938", "113940", "113945", "114223", "114226", "114232", "114750",  
 "114752", "114753", "114759", "114776", "11490", "114949", "114960",  
 "114963", "114973", "114977", "114983", "114987", "114992", "114994",  
 "114995", "114996", "114997", "115000", "115001", "115002", "116957",  
 "116983", "117128", "117134", "117138", "117143", "117144", "117145",  
 "117146", "117147", "117149", "117150", "117152", "117163", "117166",  
 "117172", "117173", "117175", "117188", "117258", "117261", "117434",  
 "117435", "117436", "117437", "117438", "117439", "117440", "117441",  
 "117442", "117443", "117444", "117512", "117944", "119277", "119278",  
 "119279", "119282", "119474", "119476", "119479", "119482", "119484",

"119486", "119488", "119489", "119693", "119698", "119699", "119932",  
"120014", "120015", "120022", "120023", "120025", "120026", "120028",  
"120029", "120030", "120031", "120032", "120033", "120035", "120036",  
"120037", "120039", "120040", "120043", "120092", "120123", "120127",  
"120132", "120133", "120135", "120137", "120138", "120139", "120141",  
"120144", "120145", "120146", "120149", "120152", "12152", "12170", "12181",  
"12183", "12184", "12202", "12352", "12564", "12565", "12567", "12952",  
"12962", "12992", "12993", "12998", "12999", "13000", "13005", "13006",  
"13007", "13008", "13009", "13011", "13012", "13013", "13014", "13015",  
"13016", "13017", "13031", "13032", "13033", "13035", "13036", "13037",  
"13042", "13083", "13084", "13086", "13957", "13958", "14332", "14333",  
"14334", "14523", "14526", "14528", "14531", "14534", "14535", "14536",  
"14537", "14538", "14539", "14540", "14542", "14543", "14544", "14545",  
"14546", "14547", "14548", "14549", "14550", "14552", "14568", "15008",  
"15244", "230087", "230088", "230090", "230091", "230092", "230094",  
"230095", "242571", "242572", "242573", "242574", "242575", "242576",  
"242577", "242578", "242579", "242580", "242581", "242582", "242583",  
"242584", "242585", "242586", "242587", "242588", "242589", "242590",  
"242591", "242592", "242593", "242594", "242595", "242596", "242597",  
"242598", "242599", "242600", "242601", "242602", "242603", "242604",  
"242605", "242606", "242607", "242608", "242609", "242610", "242611",  
"242612", "242613", "242614", "242615", "242616", "242617", "242618",  
"242619", "242620", "242621", "242622", "242623", "242624", "242625",  
"242626", "242627", "242628", "242629", "242630", "242631", "242632",  
"242633", "242634", "242635", "242636", "242672", "243151", "243154",  
"243161", "243170", "243959", "243960", "243961", "243962", "7240", "7248",  
"7253", "7261", "7354", "7473", "7485", "8664", "8985", "9027", "9029",  
"9030", "9032", "9033", "9320")

AND SUBFIELD\_DESC IN ("Finance, Economics and Accounting", "Information  
Technology and Computer Sciences", "Language", "Mathematical Sciences")

AND PROV\_PROVINCE\_DESC IN ("Gauteng")

AND PROV\_ETQE\_ID IN ("1031", "1116", "1127")

AND ETQE\_ID IN ("1116", "1127")

AND ENROL\_TYPE\_DESC IN ("Mixed Mode")

*AND FIELD\_DESC IN ("Business, Commerce and Management Studies",  
"Communication Studies and Language", "Physical, Mathematical, Computer and Life  
Sciences")*

- Cluster 3

% of records: 16.30%

Average probability: 0.9770

Rule:

ASSESSOR\_ID IN (NULL)

*AND QUALIFICATION\_ID IN ("20513", "22507", "50139", "57625", "57730",  
"58594", "60006", "61746", NULL)*

*AND UNIT\_STANDARD\_ID IN ("10765", "10767", "10771", "10773", "113869",  
"113921", "113926", "113928", "113941", "114223", "114958", "114960",  
"114970", "114971", "114972", "114974", "114975", "114976", "114978",  
"114979", "114983", "114987", "114990", "114991", "114996", "115002",  
"11513", "11514", "11515", "11516", "11517", "11518", "11519", "11522",  
"11525", "11526", "11530", "115789", "116146", "116148", "116150", "116151",  
"116158", "116160", "116162", "116551", "117134", "117146", "117150",  
"117188", "117722", "11830", "11834", "119359", "119474", "119482",  
"119484", "119489", "119666", "119667", "119668", "119669", "119693",  
"11990", "11991", "11992", "11993", "119933", "11994", "11995", "11996",  
"11997", "11998", "11999", "12000", "12001", "120014", "120015", "12002",  
"12003", "120036", "120039", "12004", "12005", "12006", "12007", "12008",  
"12009", "120092", "12010", "12011", "12012", "120127", "12013", "120132",  
"120135", "120138", "120141", "120144", "120145", "120327", "120493",  
"120494", "120495", "120496", "120497", "120498", "120499", "120500",  
"120501", "120502", "120503", "120504", "120505", "120506", "120508",  
"120509", "120510", "120511", "120512", "12345", "123528", "123531",  
"123532", "123536", "12501", "12564", "12566", "13929", "13953", "14135",  
"14139", "14147", "14148", "14359", "230087", "230088", "230090", "230091",  
"230092", "230094", "230095", "242571", "242572", "242573", "242574",  
"242575", "242576", "242577", "242578", "242579", "242580", "242581",  
"242582", "242583", "242584", "242585", "242586", "242587", "242588",  
"242589", "242590", "242591", "242592", "242593", "242594", "242595",*

"242596", "242597", "242598", "242599", "242600", "242601", "242602",  
"242603", "242604", "242605", "242606", "242607", "242608", "242609",  
"242610", "242611", "242612", "242613", "242614", "242615", "242616",  
"242617", "242618", "242619", "242620", "242621", "242622", "242623",  
"242624", "242625", "242626", "242627", "242628", "242629", "242630",  
"242631", "242632", "242633", "242634", "242635", "242636", "242671",  
"242672", "242842", "243159", "243165", "243170", "243171", "243242",  
"244193", "244194", "244196", "244198", "244199", "244201", "244206",  
"244352", "244591", "244595", "246711", "253997", "253999", "254003",  
"254007", "254010", "7473", "8617", "8985", "8986", "8987", "8990", "8992",  
"8993", "8996", "9027", "9029", "9030", "9241", "9981", "9982")  
AND PROVIDER\_ID IN ("21016", "23169", "23171", "23172", "35493",  
"35542", "36189", "36191", "37520", "38346", "39834", "46508", "46514",  
"46581", "46598", "46624", "46637", "46648", "46658", "46700", "46733",  
"46735", "46736", "46783", "46784", "46907", "46931", "46942", "46954",  
"47109", "47213", "47215", "47216", "47240", "47278", "47325", "47350",  
"47390", "47409", "47412", "47429", "47446", "47480", "47488", "47497",  
"47509", "47510", "47512", "47534", "47554", "47573", "47620", "47627",  
"47651", "47742", "47762", "47782", "47798", "47999", "48150", "48220",  
"48242", "48244", "48253", "48293", "48295", "48296", "52648")  
AND PROV\_ETQE\_ID IN ("1105")  
AND SUBFIELD\_DESC IN ("Finance, Economics AND Accounting",  
"Language", "Safety in Society")  
AND PROV\_PROVINCE\_DESC IN ("South Africa National")  
AND PROVIDER\_TYPE\_DESC IN ("Education AND Training")  
AND PROVIDER\_CLASS\_DESC IN ("Mixed: Public AND Private")  
AND ETQE\_ID IN ("1105", "1127")

- Cluster 4

% of records: 11.39%

Average probability: 0.9725

Rule:

ASSESSOR\_ID IN (NULL)

AND PROVIDER\_ID IN ("11088", "11464", "35675", "35683", "35735",  
"36326", "36373", "37377", "37392", "37395", "37405", "37872", "38476",  
"39897", "39919", "42946", "42963", "42989", "49045", "51499")

AND QUALIFICATION\_ID IN ("20202", "20924", "21808", "21810", "21813",  
"23671", "23672", "23673", "23850", "23970", "24510", "35945", "48672",  
"49665", "49852", "49946", "50097", "57625", "57954", "58080", "59114",  
"61772", "66266", "72026", "73286", NULL)

AND UNIT\_STANDARD\_ID IN ("10028", "10029", "10030", "10031", "10032",  
"10033", "10034", "10035", "10036", "10037", "10038", "10039", "10040",  
"10041", "10042", "10043", "10044", "10054", "10055", "10060", "10071",  
"10152", "10187", "10330", "10338", "10339", "10340", "10341", "10343",  
"10344", "10345", "10348", "10365", "10366", "10367", "10370", "10371",  
"10375", "10394", "10395", "10398", "10404", "10405", "10406", "10408",  
"10729", "10730", "10731", "10734", "10995", "10998", "110016", "110017",  
"110020", "110026", "110038", "110040", "110043", "110046", "110081",  
"11252", "11258", "113920", "113926", "113928", "113940", "113941",  
"113945", "114226", "114232", "114600", "114601", "114602", "114609",  
"114610", "114613", "114617", "114822", "114949", "114953", "114958",  
"114960", "114963", "114969", "114976", "114977", "114987", "114991",  
"114995", "115002", "115205", "116411", "116962", "116983", "117128",  
"117134", "117138", "117144", "117146", "117150", "117173", "117175",  
"11830", "11833", "119282", "119385", "119471", "119473", "119474",  
"119476", "119477", "119479", "119480", "119482", "119483", "119484",  
"119486", "119488", "119489", "119648", "119652", "119653", "119657",  
"119683", "119685", "119686", "119687", "119689", "119690", "119691",  
"119693", "119911", "119916", "120022", "120023", "120025", "120028",  
"120029", "120030", "120032", "120033", "120036", "120037", "120043",  
"120125", "120127", "120132", "120135", "120137", "120138", "120139",  
"120140", "120141", "120144", "120145", "120149", "120152", "120155",  
"120389", "120390", "120399", "120401", "120402", "120404", "120406",  
"12170", "12171", "12172", "12181", "12198", "12340", "12352", "12434",  
"12461", "12564", "12565", "12567", "13014", "13435", "13437", "13443",  
"13445", "13459", "13889", "13890", "13891", "13900", "13901", "13902",  
"13903", "13928", "13929", "13931", "13932", "13933", "13934", "13935",

"13936", "13945", "13946", "13947", "13948", "13949", "13950", "13951",  
"13952", "13953", "13954", "13957", "13958", "13959", "13960", "13961",  
"13962", "13964", "13965", "13966", "13969", "13970", "13971", "14101",  
"14333", "14336", "14355", "14356", "14357", "14358", "14359", "14360",  
"14361", "14363", "14365", "14366", "14367", "14368", "14369", "14370",  
"14372", "14374", "14376", "14569", "14673", "14676", "14682", "14684",  
"14964", "15025", "15076", "15106", "15231", "15233", "15236", "15237",  
"15240", "15242", "15243", "15244", "15246", "15248", "15250", "15251",  
"15252", "15254", "15255", "242601", "242610", "242672", "242831", "242832",  
"242834", "242836", "242839", "243206", "243208", "243210", "243211",  
"243212", "243214", "243215", "243216", "243218", "243219", "243220",  
"243221", "243222", "243223", "243224", "246750", "246751", "246752",  
"246753", "246754", "246755", "246756", "263373", "263451", "263472",  
"263473", "263491", "263531", "263551", "7236", "7240", "7254", "7261",  
"7464", "7466", "7468", "7473", "7478", "7480", "7481", "7485", "7486",  
"7497", "7502", "7564", "7583", "7584", "7585", "7587", "7588", "7590",  
"7592", "7723", "7802", "7808", "7813", "7853", "7877", "7880", "7893",  
"7928", "8017", "8121", "8435", "8437", "8511", "8572", "8635", "8664",  
"8979", "8980", "8981", "8982", "8983", "8984", "8985", "8986", "8987",  
"8989", "8990", "8991", "8992", "8993", "8994", "8996", "9003", "9012",  
"9021", "9024", "9025", "9026", "9027", "9029", "9030", "9032", "9033",  
"9241", "9261", "9319", "9320", "9373", "9374", "9550", "9977")

AND PROV\_ETQE\_ID IN ("1126")

AND PROV\_PROVINCE\_DESC IN ("Gauteng")

AND ETQE\_ID IN ("1126", "1127")

AND SUBFIELD\_DESC IN ("Cleaning, Domestic, Hiring, Property and Rescue Services",  
"Finance, Economics and Accounting", "Generic Management", "Hospitality, Tourism, Travel,  
Gaming and Leisure", "Information Technology and Computer Sciences", "Language",  
"Marketing", "Mathematical Sciences", "Office Administration", "People/Human-Centred  
Development", "Project Management")

AND PROVIDER\_TYPE\_DESC IN ("Education and Training", "Training")

AND UNIT\_STD\_TYPE\_DESC IN ("Regular", "Regular-Fundamental")

- Cluster 5

% of records: 6.37%

Average probability: 0.9416

Rule:

ASSESSOR\_ID IN (NULL)

AND PROVIDER\_ID IN ("11068", "11088", "11138", "11200", "21830",  
"21917", "21922", "22236", "22279", "22299", "26227", "29888", "32807",  
"37727", "37872", "38959", "39104", "39834", "40147", "40194", "40195",  
"41451", "41525", "41565", "42946", "44143", "44624", "44895", "44899",  
"44964", "44973", "45223", "45299", "45300", "45331", "49183", "49724",  
"49726", "49756", "49935", "49940", "49943", "49945", "49951", "50114",  
"50240", "50245", "50456", "50489", "50873", "50933", "51075", "51339",  
"51498", "51499", "51567", "51771", "51820", "52019", "52629", "52716")

AND QUALIFICATION\_ID IN ("20211", "21828", "23134", "23270", "23910",  
"35970", "48490", "48672", "49108", "49297", "49414", "49614", "49623",  
"49706", "49708", "50302", "58223", "58532", "58756", "58777", "59317",  
"59382", "63351", "63426", "64766", "65426", "71967", NULL)

AND UNIT\_STANDARD\_ID IN ("10246", "10269", "10366", "10370", "10371",  
"10375", "10972", "10997", "110061", "110072", "110092", "110093", "110096",  
"110548", "11258", "113846", "113869", "113894", "113926", "113932",  
"113983", "11423", "11424", "11425", "11426", "11427", "11428", "11429",  
"11430", "11431", "114356", "114358", "114360", "114367", "114370",  
"114374", "114375", "114378", "114379", "114495", "114615", "11490",  
"114904", "114906", "114907", "114908", "114909", "114910", "114911",  
"114912", "114913", "114914", "114915", "114916", "114917", "114918",  
"114919", "114920", "114921", "114922", "114923", "114924", "114925",  
"114926", "114927", "114928", "114929", "114936", "114958", "114959",  
"115110", "115118", "115122", "115205", "115772", "115840", "115872",  
"115895", "116121", "116181", "116185", "116217", "116219", "116220",  
"116221", "116223", "116235", "116252", "116312", "116314", "116525",  
"116544", "116550", "116551", "116836", "116845", "116891", "116916",  
"117008", "117016", "117034", "117046", "117162", "117172", "117173",  
"117177", "117182", "117433", "117510", "117515", "117521", "117522",  
"117667", "117887", "117894", "117904", "117908", "117909", "117914",  
"117915", "117916", "117917", "117918", "118045", "118046", "118047",  
"118050", "118054", "118060", "118062", "11835", "119248", "119471",

"119473", "119474", "119477", "119480", "119481", "119482", "119484",  
"119489", "119520", "119522", "119525", "119526", "119527", "119576",  
"119577", "119580", "119581", "119582", "119584", "119585", "119648",  
"119652", "119653", "119657", "119683", "119691", "119752", "119755",  
"119761", "119763", "119765", "119767", "119768", "119769", "119973",  
"119974", "119975", "119976", "119977", "119978", "119979", "120053",  
"12011", "120317", "120389", "120513", "12157", "12172", "12236", "123387",  
"123391", "12450", "12461", "12479", "12482", "12483", "12488", "12493",  
"12500", "12501", "12554", "12684", "13173", "13186", "13234", "13237",  
"13240", "13501", "13665", "13900", "13902", "13928", "13929", "13931",  
"13934", "13942", "13948", "13949", "13958", "13964", "13965", "13968",  
"13969", "13971", "14012", "14013", "14015", "14128", "14199", "14359",  
"14376", "14522", "14597", "14611", "14624", "14625", "14628", "14630",  
"14635", "14638", "14639", "14640", "14643", "14644", "14650", "14651",  
"14653", "14654", "14655", "14658", "14662", "14663", "14667", "14668",  
"14671", "14672", "14673", "14674", "14676", "14679", "14684", "14685",  
"14687", "14688", "14809", "15110", "15249", "230015", "230419", "230434",  
"242685", "242828", "242829", "242832", "242833", "242836", "242991",  
"243000", "243003", "243004", "243013", "243035", "243682", "243688",  
"243689", "243690", "243693", "243695", "243696", "243697", "243698",  
"243722", "243729", "243820", "243821", "243822", "243823", "243825",  
"243826", "243827", "243965", "244521", "252037", "252038", "252039",  
"252042", "252043", "252044", "252046", "252049", "252051", "252052",  
"252053", "252054", "252057", "252059", "252060", "252061", "252219",  
"252220", "252227", "252228", "252428", "252430", "252431", "252432",  
"252433", "252435", "252436", "252438", "252439", "252440", "252442",  
"252444", "252445", "252446", "252448", "252449", "252450", "252451",  
"252452", "252453", "252454", "252456", "252457", "254611", "254612",  
"254613", "255511", "255512", "255513", "255514", "255515", "255516",  
"255517", "255531", "258172", "258173", "258174", "258175", "258176",  
"258177", "258178", "258179", "258192", "258193", "258194", "258195",  
"258196", "258232", "258233", "258234", "258235", "258236", "258237",  
"258238", "259621", "260480", "260615", "261676", "261680", "261681",  
"262723", "335934", "7192", "7418", "7464", "7465", "7466", "7467", "7468",

"7473", "7478", "7480", "7481", "7485", "7486", "7497", "7506", "7519",  
"7520", "7524", "7525", "7526", "7530", "7543", "7545", "7547", "7552",  
"7734", "7802", "7882", "7890", "8014", "8017", "8033", "8054", "8055",  
"8056", "8204", "8437", "8510", "8572", "8664", "8680", "8959", "8979",  
"8980", "8981", "8984", "8985", "8986", "8987", "8988", "8989", "8990",  
"8991", "8992", "8993", "8996", "9024", "9025", "9026", "9027", "9029",  
"9030", "9032", "9033", "9079", "9319", "9320", "9523", "9840", "9899",  
"9981", "9982", "9990")

AND PROV\_ETQE\_ID IN ("1102", "1103", "1109", "1111", "1112", "1114",  
"1117")

AND ETQE\_ID IN ("1102", "1103", "1109", "1111", "1112", "1114", "1117")

AND PROVIDER\_TYPE\_DESC IN ("Education AND Training")

AND ENROL\_TYPE\_DESC IN ("Mixed Mode")

AND PROV\_PROVINCE\_DESC IN ("Gauteng", "Kwazulu/Natal", "Western Cape")

AND SUBFIELD\_DESC IN ("Adult Learning", "Building Construction", "Curative Health",  
"Generic Management", "Hospitality, Tourism, Travel, Gaming and Leisure", "Human  
Resources", "Language", "Manufacturing and Assembly", "Marketing", "Mathematical  
Sciences", "Office Administration", "People/Human-Centred Development", "Preventive  
Health", "Promotive Health and Developmental Services", "Secondary Agriculture", "Transport,  
Operations and Logistics", "Wholesale and Retail")

- Cluster 6

% of records: 6.05%

Average probability: 0.9858

Rule:

PRIMARY\_ETQE\_DESC IN ("Not Primary ETQE of provider")

AND PROVIDER\_ID IN ("1988", "21318", "21328", "21337", "21375", "21778",  
"21780", "39919", "49583", "50114", "50578")

AND UNIT\_STANDARD\_ID IN ("10305", "10306", "10311", "10312", "114493",  
"114600", "114602", "114609", "114613", "114955", "114959", "115770",  
"115776", "119474", "119476", "119479", "119482", "119484", "119486",  
"119488", "119489", "12851", "12856", "12859", "13660", "13864", "13865",  
"13866", "13867", "13868", "13869", "13870", "13871", "13872", "13873",  
"13942", "14599", "15113", "15234", "15235", "15238", "15244", "15245",  
"15249", "242829", "242833", "242836", "244273", "244274", "244276",

"244277", "244479", "244485", "244486", "244489", "244492", "244495",  
"244497", "244498", "244501", "244502", "7417", "7418", "7420", "7422",  
"7423", "7424", "7425", "7426", "7427", "7465", "7466", "7467", "7468",  
"7470", "7478", "7480", "7481", "7485", "7541", "7543", "7545", "7547",  
"7551", "7552", "7995", "8664", "8991", "8992", "8993", "8996", "9032",  
"9033", "9949", "9952", "9974")

AND QUALIFICATION\_ID IN ("23131", "23133", "23134", "23135", "58778",  
NULL)

AND ASSESSOR\_ID IN ("3008406", "3011448", "3014785", "3018255",  
"3018856", "3024384", "3024481", "3024858", "3027166", "3027229",  
"3027363", "3028901", "3029299", "3031299", "3031754", "3032283",  
"3032933", "3033082", "3033416", "3040050", "3313040", "3313057",  
"3313117", "3313287", "3313288", "3557298", "4282614", "4282640",  
"4282714", "4282924", "4282927", "4282957", "4283013", "4631724",  
"4631855", "4631954", "4632161", "4632257", "4632259", "4632260",  
"4632261", "4632262", "4632265", "5013309", "5013828", "5518246",  
"5518570", "6055536", "6056024", "7309219", "7309240", "8145095",  
"8145588", "8145591", "9463777", "9464189", "9464322", NULL)

AND PROV\_ETQE\_ID IN ("1033", "1106")

AND ETQE\_ID IN ("1106")

AND SUBFIELD\_DESC IN ("Adult Learning", "Communication Studies", "Early  
Childhood Development", "Generic Management", "Higher Education and  
Training", "Language", "Mathematical Sciences", "People/Human-Centred  
Development")

AND ENROL\_TYPE\_DESC IN ("Residential Learning (i.e. Contact Mode)")

AND FIELD\_DESC IN ("Business, Commerce and Management Studies", "Communication  
Studies AND Language", "Education, Training AND Development", "Physical, Mathematical,  
Computer and Life Sciences")

- Cluster 7

% of records: 4.38%

Average probability: 0.9273

Rule:

PRIMARY\_ETQE\_DESC IN ("Not Primary ETQE of provider")

AND ASSESSOR\_ID IN ("17162273", "3013505", "3483001", "3483283",  
"4688814", "4783541", "4783572", "4783603", "6129807", "6130027",  
"7363698", "7363790", "7363870", "8218000", NULL)

AND PROVIDER\_ID IN ("11100", "21318", "21323", "21336", "21342",  
"21343", "21344", "21448", "21468", "21537", "21550", "21561", "21637",  
"21650", "21665", "21768", "21917", "33812", "35404", "35405", "35406",  
"35409", "35411", "39490", "41433", "41444", "41451", "41493", "41565",  
"42946", "44143", "44880", "46423", "46460", "48370", "48475", "48477",  
"48478", "48501", "48508", "48534", "48554", "48558", "48559", "48562",  
"48566", "48591", "49816", "50014", "51338", "51339", "51499")

AND QUALIFICATION\_ID IN ("14128", "14868", "21021", "21022", "22787",  
"22788", "22875", "22886", "23270", "23272", "23290", "23291", "23696",  
"24290", "24473", "36230", "36453", "48490", "48780", "48976", "48982",  
"48987", "48989", "48992", "48993", "48994", "49026", "49065", "49094",  
"49643", "49666", "49706", "49708", "49709", "49946", "50389", "50559",  
"57625", "57934", "58392", "58393", "58551", "58552", "58798", "63490",  
"64826", "64827", "65426", "66791", "71967", "79303", NULL)

AND UNIT\_STANDARD\_ID IN ("10001", "10019", "10020", "10054", "10152",  
"10186", "10187", "10246", "10269", "10270", "10341", "10995", "10997",  
"10998", "11000", "11002", "110040", "110070", "110305", "110306", "110545",  
"11219", "11252", "11258", "113880", "113941", "114093", "114653", "114750",  
"114752", "114753", "114759", "11490", "114908", "114911", "114958",  
"114960", "114976", "114991", "115075", "115128", "115408", "115410",  
"11550", "116070", "116072", "116077", "116079", "116080", "116081",  
"116082", "116086", "116087", "116089", "116093", "116094", "116096",  
"116097", "116098", "116100", "116126", "116127", "116130", "116131",  
"116132", "116137", "116138", "116139", "116140", "116141", "116142",  
"116143", "116144", "116145", "116165", "116167", "116173", "116174",  
"116175", "116176", "116177", "116180", "116181", "116182", "116183",  
"116184", "116185", "116186", "116189", "116214", "116218", "116225",  
"116226", "116240", "116246", "116247", "116249", "116251", "116252",  
"116255", "116256", "116258", "116260", "116261", "116263", "116295",  
"116296", "116303", "116305", "116306", "116307", "116308", "116310",  
"116312", "116314", "116318", "116319", "116320", "116321", "116322",

"116323", "116326", "116328", "116329", "116331", "116332", "116333",  
"116334", "116336", "116337", "116338", "116339", "116351", "116352",  
"116356", "116362", "116363", "116370", "116372", "116654", "116655",  
"116660", "116662", "116664", "116670", "116701", "116713", "116718",  
"116737", "116947", "116948", "116949", "116952", "116953", "116954",  
"116957", "116959", "116960", "116962", "116983", "117173", "117434",  
"117435", "117436", "117437", "117438", "117439", "117440", "117441",  
"117442", "117443", "117444", "117512", "117602", "117609", "117610",  
"117641", "117643", "117644", "117649", "117884", "117887", "117888",  
"117894", "117914", "117918", "117919", "117940", "117941", "117942",  
"117944", "117945", "119074", "119076", "119095", "119156", "119362",  
"119471", "119473", "119474", "119476", "119477", "119479", "119480",  
"119481", "119482", "119483", "119484", "119486", "119488", "119489",  
"119683", "119685", "119687", "119689", "119690", "119691", "119729",  
"119752", "119753", "119754", "119755", "119760", "119761", "119762",  
"119763", "119765", "119766", "119767", "119768", "119769", "119770",  
"120123", "12053", "12152", "12170", "12171", "12172", "12228", "12229",  
"12231", "12232", "12233", "12235", "12236", "12256", "12257", "12260",  
"12263", "123615", "12370", "12434", "12446", "12461", "12473", "12474",  
"12478", "12479", "12480", "12482", "12483", "12486", "12487", "12488",  
"12490", "12493", "12494", "12498", "12500", "12501", "12505", "12554",  
"12684", "12952", "12962", "12992", "12993", "12998", "12999", "13000",  
"13005", "13006", "13007", "13008", "13009", "13011", "13012", "13013",  
"13014", "13015", "13016", "13017", "13031", "13032", "13033", "13035",  
"13036", "13037", "13042", "13083", "13084", "13086", "13151", "13174",  
"13176", "13179", "13182", "13184", "13186", "13188", "13189", "13191",  
"13193", "13221", "13222", "13231", "13233", "13234", "13235", "13236",  
"13237", "13238", "13239", "13240", "13251", "13271", "13275", "13293",  
"13294", "13295", "13296", "13299", "13300", "13314", "13320", "13344",  
"13654", "13664", "13666", "13694", "13929", "13932", "13949", "13958",  
"13994", "14012", "14013", "14015", "14037", "14071", "14101", "14231",  
"14241", "14242", "14243", "14462", "14508", "14510", "14511", "14568",  
"14597", "14603", "14607", "14612", "14617", "14626", "14629", "14649",  
"14667", "14671", "14674", "14679", "14682", "14684", "14689", "14690",

"14691", "14693", "14696", "14700", "14723", "14729", "14730", "14739",  
"14899", "14900", "14901", "14902", "14903", "14904", "14905", "14906",  
"14907", "14908", "14909", "14910", "14911", "14912", "14929", "14930",  
"14934", "14935", "14964", "15117", "15118", "15122", "15127", "15131",  
"15133", "15134", "15135", "15233", "15244", "230087", "230090", "230092",  
"242572", "242574", "242575", "242576", "242579", "242580", "242583",  
"242584", "242585", "242586", "242587", "242588", "242589", "242591",  
"242592", "242594", "242595", "242600", "242601", "242602", "242609",  
"242611", "242613", "242617", "242624", "242631", "242828", "242838",  
"243072", "243073", "243080", "243081", "243083", "243084", "243085",  
"243086", "243089", "243092", "243093", "243105", "243768", "243774",  
"243959", "243960", "243961", "243962", "244065", "244066", "244068",  
"244070", "244073", "244074", "244180", "244521", "244703", "252210",  
"252214", "252217", "252231", "252235", "253457", "254237", "259621",  
"260177", "260454", "260654", "260655", "260696", "260734", "260735",  
"260736", "260737", "260738", "261674", "261675", "261676", "261677",  
"261678", "261679", "261680", "261681", "261682", "261683", "261694",  
"261695", "261696", "261697", "261698", "261714", "261734", "261754",  
"263993", "335872", "7464", "7465", "7466", "7467", "7468", "7469", "7473",  
"7478", "7480", "7481", "7485", "7486", "7497", "7506", "7524", "7525",  
"7526", "7528", "7530", "7564", "7585", "7589", "7626", "7676", "7677",  
"7678", "7765", "7779", "7802", "7810", "7816", "7817", "7819", "7822",  
"7825", "7826", "7827", "7828", "7829", "8364", "8388", "8403", "8405",  
"8425", "8664", "8665", "8679", "8979", "8980", "8981", "8984", "8985",  
"8986", "8987", "8988", "8989", "8990", "8991", "8992", "8993", "8996",  
"9024", "9025", "9026", "9027", "9029", "9030", "9032", "9033", "9285",  
"9319", "9320", "9339", "9374", "9460", "9523", "9543", "9545", "9546",  
"9547", "9550", "9689", "9856", "9893", "9894", "9895", "9896", "9897",  
"9898", "9899", "9905", "9930", "9931", "9977", "9981", "9982", "9984",  
"9985", "9986", "9988", "9990")

AND ETQE\_ID IN ("1075", "1103", "1109", "1112", "1116", "1127")

AND PROV\_ETQE\_ID IN ("1031", "1033", "1075", "1112", "1115")

AND SUBFIELD\_DESC IN ("Building Construction", "Engineering and Related Design",  
 "Finance, Economics and Accounting", "Generic Management", "Hospitality, Tourism, Travel,  
 Gaming and Leisure", "Human  
 Resources", "Information Technology and Computer Sciences", "Language", "Life Sciences",  
 "Manufacturing and Assembly", "Mathematical Sciences", "Primary Agriculture")  
 AND ENROL\_TYPE\_DESC IN ("Mixed Mode")  
 AND PROV\_PROVINCE\_DESC IN ("Eastern Cape", "Gauteng", "Kwazulu/Natal",  
 "Limpopo", "Mpumalanga", "Western Cape")

- Cluster 8

% of records: 2.93%

Average probability: 0.9574

Rule:

UNIT\_STD\_TYPE\_DESC IN ("Regular-Fundamental")

AND ASSESSOR\_ID IN ("6130511", NULL)

AND UNIT\_STANDARD\_ID IN ("10152", "10156", "10157", "10330", "110040",  
 "11258", "114063", "114065", "114066", "114067", "114068", "114071",  
 "114072", "114073", "114075", "114076", "114077", "114083", "114089",  
 "114091", "114093", "114653", "115001", "115375", "115376", "115379",  
 "115382", "115384", "115390", "115401", "115408", "115409", "115448",  
 "116406", "116947", "116948", "116949", "116952", "116953", "116954",  
 "116957", "116959", "116960", "116962", "117173", "117884", "117919",  
 "117940", "117941", "117942", "117943", "117944", "117945", "119095",  
 "119381", "119385", "119390", "119393", "119474", "119476", "119479",  
 "119482", "119484", "119486", "119489", "120396", "12170", "12171", "12434",  
 "12461", "13932", "13948", "14673", "14925", "14926", "14927", "14929",  
 "14930", "14932", "14934", "14935", "14936", "14937", "14938", "14939",  
 "14941", "14943", "14944", "14945", "14946", "14947", "14948", "14949",  
 "14950", "14951", "14952", "14953", "14954", "14955", "14956", "14957",  
 "14958", "14959", "14960", "14961", "14962", "14963", "14964", "14965",  
 "14967", "14969", "14970", "14980", "258892", "258893", "258894", "258895",  
 "258896", "258897", "258898", "258899", "258900", "258914", "258915",  
 "7466", "7473", "7481", "7485", "7486", "7497", "7584", "7802", "8985",  
 "8986", "8987", "8989", "8990", "9024", "9025", "9026", "9027", "9029",  
 "9030", "9032", "9033", "9549")

*AND PROVIDER\_ID IN ("28903", "34337", "39043", "39833", "39834",  
 "39852", "45299", "45504", "46460", "47630", "49273", "49275", "49724",  
 "49726", "49756", "50456", "50489", "51082", "51086", "51095", "51102",  
 "51108", "51119", "51125", "51129", "51133", "51134", "51138", "51144",  
 "51149", "51153", "51158", "51176", "51191", "51195", "51198", "51199",  
 "51209", "51213", "51217", "51222", "51225", "51239", "51244", "51249",  
 "51250", "51255", "51265", "51481", "51499", "51588", "51590", "51771",  
 "51819", "52019", "52648")*

*AND QUALIFICATION\_ID IN ("20202", "21828", "23270", "23910", "24310",  
 "48590", "48688", "48889", "49094", "49108", "49623", "49706", "58393",  
 "58411", "58756", "58777", "58995", "59317", "59382", "61467", "78981",  
 "78982", NULL)*

*AND PROV\_ETQE\_ID IN ("1103", "1123")*

*AND ETQE\_ID IN ("1103", "1123")*

*AND SUBFIELD\_DESC IN ("Information Technology and Computer Sciences",  
 "Language", "Mathematical Sciences", "Project Management")*

*AND PROVIDER\_TYPE\_DESC IN ("Education and Training")*

*AND ENROL\_TYPE\_DESC IN ("Mixed Mode")*

### ***K.3.2 Start During, Start After and End After cluster data mining***

This section provides a technical description of the clusters that were generated by cluster data mining the consolidated data category ‘Start During, Start After and End After’ (see Appendix J.3.8) for unit standard enrolment records. A detailed description of the manner in which cluster data mining was conducted can be found in Appendix I.3. The results of the generated clustering model were significant because the model was measured as being 94.45% accurate.

The generated clusters show a tight coupling between data fields that describe the ETQEs, unit standards and providers. This is as a result of the organic relationship between unit standards and ETQEs (unit standards are generally implemented by one ETQE only) and providers and ETQEs (providers generally offer unit standards that are implemented by their primary ETQE).

The model did not generate any clusters that contained less than 1% of the records. The detection of anomalous records based on cluster specific probabilities of less than .6 being allocated to the record was conducted. In this manner it was determined that 8.04% of the records possibly exist in this category as a result of data capturing problems (see Appendix I.3).

The technical description of each of the 8 clusters is provided below. The description is limited to the top 10 attributes of the cluster rule, where each indent of text represents the importance of the attribute. For example:

A line formatted like this represents an importance of 100%

*A line formatted like this represents an importance greater than 75% and less than 100%*

*A line formatted like this represents an importance greater than 50% and less than or equal to 75%*

*A line formatted like this represents an importance less than or equal to 50%*

- Cluster 1

% of records: 21.64%

Average probability: 0.9291

Rule:

*ASSESSOR\_ID IN ("17369081", "3013505", "5115688", "7339985", "7355575", "NULL")*

*AND PROVIDER\_ID IN ("11193", "11194", "11343", "11772", "12916", "17119", "17127", "1726", "18701", "18705", "1974", "20589", "22680", "23033", "29242", "29323", "31724", "32915", "36061", "36683", "37171", "37715", "37975", "38548", "38607", "38637", "38642", "38989", "39084", "39490", "39904", "41433", "41482", "41493", "41540", "41933", "42152", "43923", "44003", "44010", "49185", "49200", "49342", "49381", "49420", "49421", "49717", "49783", "49792", "52070", "594")*

*AND QUALIFICATION\_ID IN ("14127", "14130", "20194", "20911", "20924", "21813", "23391", "23672", "35945", "36230", "48492", "48590", "48982", "49030", "49106", "49622", "49623", "49708", "49709", "49946", "50077", "57625", "57841", "58392", "58393", "58756", "58968", "59293", "59382", "60172", "60207", "61686", "61726", "62266", "62606", "63806", "65506", "74166", "NULL")*

*AND UNIT\_STANDARD\_ID IN ("14336", "14434", "14443", "14500", "14523", "14534", "14535", "14536", "14537", "14538", "14540", "14542", "14548", "14550", "14551", "14552", "14568", "14569", "14626", "14684", "14996", "14997", "15008",*

"15011", "15012", "15016", "15025", "15231", "15232", "15234", "15237", "15244",  
"15246", "15247", "15249", "15251", "15282", "15316", "230087", "230088",  
"230092", "230094", "230095", "230465", "242571", "242590", "242591", "242597",  
"242601", "242610", "242672", "242682", "242827", "242831", "242875", "242877",  
"242883", "242887", "242897", "242917", "242918", "242919", "242920", "242931",  
"243000", "243003", "243004", "243008", "243013", "243151", "243170", "243280",  
"243281", "243313", "243315", "243697", "243821", "243841", "243959", "243960",  
"243961", "243962", "244192", "244200", "244271", "244376", "244377", "244380",  
"244381", "244385", "244395", "244396", "244397", "244399", "244400", "244401",  
"244402", "244403", "244405", "244409", "244410", "244415", "244422", "244425",  
"244427", "244428", "244430", "244432", "244433", "244434", "244437", "244439",  
"244442", "244446", "244450", "244451", "244459", "244460", "244462", "244463",  
"244464", "244465", "244467", "244470", "244521", "244617", "246458", "246462",  
"246463", "246757", "252050", "252208", "252210", "252211", "252212", "252213",  
"252214", "252218", "252219", "252220", "252221", "252223", "252226", "252227",  
"252228", "252231", "252233", "252234", "252235", "252410", "252421", "252428",  
"252430", "252431", "252432", "252433", "252434", "252435", "252436", "252438",  
"252439", "252440", "252441", "252442", "252443", "252444", "252445", "252446",  
"252447", "252448", "252449", "252450", "252451", "252452", "252453", "252454",  
"252455", "252456", "252457", "252529", "252530", "254083", "254084", "254086",  
"254087", "254089", "254090", "254091", "254093", "254094", "254114", "254116",  
"254118", "254120", "254131", "254132", "254133", "254134", "254135", "254138",  
"254140", "254142", "254151", "254211", "254239", "255491", "255992", "255993",  
"255994", "255995", "255996", "255997", "255998", "255999", "256000", "256001",  
"256002", "256003", "256004", "256492", "256493", "256494", "256495", "256496",  
"256497", "256498", "256499", "256500", "256501", "256502", "256512", "256513",  
"256514", "256552", "256553", "256572", "256573", "256574", "258237", "258934",  
"258935", "258936", "258937", "258938", "258939", "258940", "258942", "258943",  
"258944", "258945", "258946", "258948", "258949", "258951", "258952", "258953",  
"258954", "258955", "258956", "258958", "258959", "258974", "258977", "258979",  
"258984", "259034", "259160", "259621", "260614", "261674", "261678", "261754",  
"261836", "264277", "264415", "264420", "265016", "265019", "265022", "265024",  
"265025", "265051", "337076", "337078", "337080", "377913", "377918", "377970",  
"6816", "6877", "6900", "6914", "6917", "6941", "6944", "6957", "6974", "7192",

"7485", "7590", "7619", "7622", "7675", "7677", "7680", "7717", "7723", "7765",  
"7779", "7799", "7801", "7802", "7803", "7804", "7805", "7807", "7808", "7810",  
"7811", "7813", "7816", "7817", "7826", "7828", "7829", "7835", "7853", "7865",  
"7877", "7885", "7886", "7899", "8014", "8017", "8055", "8437", "8497", "8578",  
"8635", "8664", "8665", "8679", "8681", "9003", "9014", "9025", "9026", "9241",  
"9259", "9547", "9550", "9619", "9706", "9708", "9710", "9717", "9898", "9977",  
"9981", "9990", "10025", "10028", "10029", "10030", "10031", "10032", "10033",  
"10034", "10035", "10036", "10037", "10038", "10039", "10040", "10041", "10042",  
"10043", "10044", "10045", "10046", "10048", "10054", "10055", "10061", "10152",  
"10165", "10186", "10187", "10188", "10211", "10225", "10227", "10231", "10232",  
"10233", "10234", "10235", "10236", "10237", "10250", "10251", "10269", "10272",  
"10277", "10278", "10282", "10286", "10287", "10381", "10385", "10390", "10394",  
"10395", "10402", "10404", "10405", "10406", "10408", "10409", "10410", "10568",  
"10572", "10573", "10729", "10730", "10731", "10733", "10735", "10757", "10758",  
"10995", "10997", "10998", "11000", "11002", "110020", "110026", "110040",  
"110061", "110092", "110135", "110139", "110144", "110404", "110545", "11252",  
"11258", "11290", "11303", "113846", "113876", "113893", "113894", "113910",  
"113915", "113921", "113926", "113928", "113931", "113933", "113935", "113938",  
"113939", "113941", "113972", "113973", "113977", "114062", "114067", "114072",  
"114232", "114243", "114290", "114291", "114295", "114405", "114423", "114508",  
"114601", "114603", "114750", "114752", "114753", "114754", "114755", "114759",  
"114799", "114890", "114899", "11490", "114902", "114903", "114928", "114949",  
"114953", "114958", "114960", "114962", "114969", "114971", "114972", "114974",  
"114976", "114977", "114978", "114979", "114981", "114983", "114987", "114988",  
"114990", "114991", "114994", "114995", "114996", "115000", "115001", "115002",  
"115191", "115847", "115924", "115982", "116097", "116235", "116243", "116260",  
"116261", "116291", "116311", "116356", "116357", "116358", "116359", "116360",  
"116361", "116362", "116363", "116364", "116365", "116368", "116370", "116374",  
"116375", "116377", "116378", "116379", "116380", "116381", "116454", "116485",  
"116551", "116604", "116653", "116672", "116674", "116687", "116713", "116714",  
"116719", "116724", "116729", "116734", "116736", "116737", "116944", "116945",  
"116949", "116955", "116957", "117008", "117016", "117020", "117034", "117066",  
"117099", "117121", "117125", "117128", "117134", "117135", "117137", "117138",  
"117139", "117140", "117141", "117142", "117143", "117144", "117145", "117146",

"117147", "117148", "117149", "117150", "117152", "117153", "117154", "117158",  
"117163", "117166", "117172", "117173", "117175", "117188", "117258", "117261",  
"117407", "117434", "117435", "117436", "117437", "117438", "117439", "117440",  
"117441", "117442", "117443", "117444", "117512", "117524", "117887", "117888",  
"117894", "117900", "117914", "117916", "117942", "117943", "117944", "118045",  
"11833", "11834", "119136", "11924", "11926", "119282", "119348", "119349",  
"119351", "119353", "119357", "119359", "119360", "119362", "119365", "119367",  
"119368", "119369", "119370", "119495", "119570", "119571", "119572", "119573",  
"119574", "119575", "119576", "119577", "119579", "119580", "119581", "119582",  
"119584", "119693", "120014", "120022", "120026", "120032", "120033", "120036",  
"120037", "120092", "120123", "120127", "120135", "120137", "120138", "120140",  
"120141", "120145", "120149", "120153", "120320", "120321", "120324", "120327",  
"120389", "120406", "120408", "120409", "120410", "120411", "120504", "120513",  
"12065", "12075", "12152", "12155", "12156", "12157", "12181", "123151",  
"123154", "12357", "12450", "12478", "12482", "12483", "12500", "12501",  
"12564", "12567", "12952", "12962", "12992", "12993", "12998", "12999", "13000",  
"13005", "13006", "13007", "13008", "13009", "13011", "13012", "13013", "13014",  
"13015", "13016", "13017", "13031", "13032", "13033", "13035", "13036", "13037",  
"13042", "13083", "13084", "13086", "13119", "13234", "13238", "13251", "13808",  
"13929", "13953", "13957", "13959", "13962", "13971", "14015", "14151", "14152",  
"14199", "14332", "14333")

AND SUBFIELD\_DESC IN ("Civil Engineering Construction", "Curative Health", "Electrical  
Infrastructure Construction", "Environmental Sciences", "Fabrication and Extraction",  
"Finance, Economics and Accounting", "Generic Management", "Hospitality, Tourism, Travel,  
Gaming and Leisure", "Human Resources", "Information Technology and Computer  
Sciences", "Manufacturing and Assembly", "Marketing", "People/Human-Centred  
Development", "Procurement", "Promotive Health and Developmental Services", "Public  
Administration", "Transport, Operations and Logistics")

AND PROVIDER\_CLASS\_DESC IN ("Private")

AND PROVIDER\_TYPE\_DESC IN ("Education", "Education and Training", "Training")

AND ETQE\_ID IN ("1075", "1100", "1102", "1110", "1111", "1116", "1117",  
"1118", "1122", "1125", "1126", "1127")

AND PRIMARY\_ETQE\_DESC IN ("Primary ETQE of provider")

AND FIELD\_DESC IN ("Business, Commerce and Management Studies", "Health  
Sciences and Social Services", "Manufacturing, Engineering and Technology",  
"Physical Planning and Construction", "Services")

- Cluster 2

% of records: 20.83%

Average probability: 0.9979

Rule:

ASSESSOR\_ID IN ("NULL")

AND QUALIFICATION\_ID IN ("49614", "50139", "58594", "60006", "61746", "NULL")

AND UNIT\_STANDARD\_ID IN ("10765", "113869", "113926", "113941", "114958", "114996", "11522", "11525", "11530", "116551", "116611", "117722", "117906", "119359", "119474", "119482", "119484", "119489", "119666", "119667", "119668", "119669", "11998", "12002", "12007", "120493", "120494", "120495", "120496", "120497", "120498", "120499", "120500", "120501", "120502", "120503", "120504", "120505", "120506", "120508", "120509", "120511", "123527", "123528", "123529", "123530", "123531", "123532", "123535", "123536", "12501", "13929", "13953", "15113", "230069", "242842", "243205", "243207", "244193", "244194", "244195", "244196", "244198", "244199", "244201", "244206", "244352", "244595", "244622", "246710", "246711", "252191", "253991", "253995", "253996", "253997", "253999", "254000", "254002", "254003", "254004", "254005", "254007", "254010", "336676", "7473", "9027", "9029", "9030")

AND PROVIDER\_ID IN ("17134", "17141", "17146", "17149", "17156", "2066", "20689", "20707", "20725", "20745", "20770", "20772", "20785", "20796", "20803", "20917", "20964", "21039", "21089", "21121", "21164", "21185", "22273", "23092", "23096", "25139", "25152", "25158", "25169", "28927", "28937", "28953", "30139", "32878", "35466", "35471", "35501", "35509", "35516", "35537", "35545", "35551", "37505", "37523", "37532", "37540", "38348", "38377", "38397", "38420", "46514", "46808", "46926", "47700", "47762")

AND ETQE\_ID IN ("1105")

AND PROV\_ETQE\_ID IN ("1105")

AND 0 <= START\_PROV\_IND <= 13.2

AND PROV\_PROVINCE\_DESC IN ("South Africa National")

AND SUBFIELD\_DESC IN ("Finance, Economics and Accounting", "Language", "Manufacturing and Assembly", "Mathematical Sciences", "People/Human-Centred Development", "Preventive Health", "Safety in Society")

AND FIELD\_DESC IN ("Business, Commerce and Management Studies", "Law, Military Science and Security")

- Cluster 3

% of records: 15.95%

Average probability: 0.8898

Rule:

ASSESSOR\_ID IN ("NULL")

AND UNIT\_STANDARD\_ID IN ("10152", "110016", "110020", "110026", "110038", "110040", "110043", "114508", "114606", "114928", "115872", "116537", "116947", "116948", "116954", "116959", "116962", "119095", "119474", "119476", "119479", "119482", "119483", "119484", "119486", "119488", "119489", "120365", "120513", "12170", "12434", "13929", "13932", "13945", "13946", "13947", "13950", "13951", "13953", "13958", "13960", "13962", "14964", "15251", "242828", "242836", "243697", "244606", "244628", "7473", "7485", "7584", "7587", "8985", "8986", "8987", "8989", "8990", "8991", "8992", "8993", "9025", "9026", "9027", "9029", "9030", "9032", "9033")

AND PROVIDER\_ID IN ("11098", "11343", "11772", "17119", "1974", "20689", "26373", "36683", "37171", "37975", "41493", "42152", "42351", "44010", "49200", "49342", "49420", "49792")

AND QUALIFICATION\_ID IN ("20924", "23671", "23672", "23673", "23970", "35945", "48492", "49106", "49665", "49708", "57841", "58392", "58393", "58531", "59293", "60207", "61566", "62266", "62606", "63806", "74647", "78981", "NULL")

AND SUBFIELD\_DESC IN ("Generic Management", "Information Technology and Computer Sciences", "Language", "Mathematical Sciences", "Office Administration", "People/Human-Centred Development", "Preventive Health")

AND ETQE\_ID IN ("1075", "1100", "1116", "1123", "1125", "1126")

AND PROVIDER\_TYPE\_DESC IN ("Education and Training")

AND PROVIDER\_CLASS\_DESC IN ("Private")

AND PROV\_ETQE\_ID IN ("1075", "1100", "1116", "1123", "1125", "1126")

AND FIELD\_DESC IN ("Business, Commerce and Management Studies", "Communication Studies and Language", "Physical, Mathematical, Computer and Life Sciences")

- Cluster 4

% of records: 12.64%

Average probability: 0.9649

Rule:

*ASSESSOR\_ID* IN ("3015274", "3029470", "5657633", "NULL")  
*AND PROVIDER\_ID* IN ("11059", "11065", "11072", "11073", "11087", "11088",  
"11098", "20357", "20589", "2169", "2171", "2172", "2183", "2224", "2251",  
"24993", "28710", "29272", "29275", "29278", "31726", "34485", "34583", "41445",  
"41474", "41566", "49128", "49421", "5994")  
*AND UNIT\_STANDARD\_ID* IN ("10187", "10212", "10272", "10287", "10505",  
"10539", "10591", "10599", "10602", "10604", "10615", "10630", "10639", "10647",  
"10648", "10696", "10701", "10702", "10757", "110024", "110092", "110142",  
"110234", "110316", "110317", "110571", "113890", "113894", "113916", "113918",  
"113941", "114211", "114234", "114383", "114423", "114473", "114475", "114476",  
"114479", "114480", "114481", "114482", "114615", "114620", "114622", "114626",  
"114631", "114632", "114637", "114638", "114640", "114641", "114654", "114656",  
"114657", "114658", "114659", "114660", "114661", "114662", "114664", "114666",  
"114667", "114668", "114669", "114670", "114920", "114923", "114927", "114929",  
"114953", "114967", "115205", "115838", "116073", "116076", "116101", "116103",  
"116252", "116537", "116544", "116551", "116731", "116948", "116949", "116954",  
"116957", "116962", "117884", "117891", "117894", "117904", "117917", "117919",  
"117941", "117944", "119095", "119176", "119183", "119344", "119345", "119381",  
"119385", "119390", "119471", "119473", "119474", "119475", "119476", "119477",  
"119479", "119480", "119482", "119483", "119484", "119486", "119488", "119489",  
"11961", "119648", "119652", "119653", "119657", "119683", "11974", "119761",  
"119813", "119814", "119815", "119816", "119817", "119818", "119819", "120252",  
"120254", "120255", "120256", "120258", "120262", "120317", "120392", "120394",  
"12040", "120412", "120414", "120415", "120417", "120418", "120419", "120420",  
"120421", "120422", "120424", "120425", "120427", "120428", "120429", "120430",  
"120433", "120434", "120435", "12113", "12216", "12221", "12224", "12232",  
"12236", "12242", "12269", "12275", "12332", "12333", "12334", "12336",  
"123374", "123376", "123377", "123378", "123384", "123388", "123390", "123391",  
"123392", "12446", "12461", "12472", "12473", "12474", "12478", "12480",  
"12482", "12483", "12493", "12494", "12498", "12500", "12501", "12505", "13156",  
"13182", "13231", "13233", "13234", "13236", "13237", "13238", "13240", "13241",  
"13251", "13277", "13293", "13294", "13295", "13296", "13297", "13299", "13300",

"13314", "13316", "13320", "13342", "13344", "13346", "13351", "13617", "13720",  
"13730", "13737", "13835", "13929", "13964", "13969", "13970", "13978", "13979",  
"13980", "14053", "14054", "14055", "14077", "14079", "14080", "14082", "14113",  
"14123", "14127", "14234", "14359", "14374", "14376", "14490", "14586", "14622",  
"14673", "14681", "14723", "14791", "14800", "14818", "14821", "14897", "14898",  
"14913", "14915", "14920", "15237", "15247", "15253", "15254", "15276", "15279",  
"15293", "15294", "15299", "15308", "15320", "15349", "15351", "15352", "15356",  
"242998", "243073", "243075", "243076", "243080", "243084", "243085", "243089",  
"243092", "243094", "243095", "243098", "244066", "244068", "244078", "244079",  
"244080", "244081", "244082", "244083", "244084", "244085", "244086", "244087",  
"244088", "244090", "244092", "244093", "244094", "244095", "244096", "244097",  
"244098", "244099", "244100", "244101", "244102", "244103", "244104", "244105",  
"244106", "244107", "244108", "244109", "244110", "244111", "244112", "244113",  
"244115", "244125", "244258", "244384", "244703", "246491", "246493", "246494",  
"246495", "246496", "246499", "253378", "253391", "253393", "253403", "253408",  
"253432", "253440", "253451", "253456", "253514", "253575", "253596", "253604",  
"253605", "253609", "253752", "255991", "258936", "258937", "258942", "258949",  
"258952", "258976", "258977", "258978", "258979", "258982", "258983", "258984",  
"258985", "258994", "259014", "259055", "259075", "259094", "259095", "259114",  
"259194", "259197", "259209", "259214", "259217", "259218", "259234", "260655",  
"260736", "262498", "262499", "262500", "262502", "262503", "262505", "336836",  
"7473", "7474", "7485", "7486", "7497", "7502", "7564", "7565", "7584", "7587",  
"7626", "8782", "8783", "8820", "8822", "8823", "8824", "8839", "8887", "8922",  
"8940", "8941", "8959", "8960", "8961", "8962", "8963", "8979", "8980", "8981",  
"8984", "8985", "8986", "8987", "8989", "8990", "9024", "9025", "9026", "9027",  
"9029", "9030", "9032", "9033", "9059", "9061", "9063", "9068", "9069", "9070",  
"9071", "9079", "9080", "9122", "9128", "9130", "9132", "9139", "9142", "9143",  
"9148", "9153", "9285", "9339", "9543", "9545", "9546", "9547", "9550", "9616",  
"9626", "9628", "9629", "9630", "9631", "9633", "9634", "9635", "9639", "9640",  
"9711", "9898", "9899", "9914")

AND QUALIFICATION\_ID IN ("20211", "20671", "21829", "23694", "23695",  
"48812", "49760", "50100", "50322", "50323", "50324", "57711", "58392", "58531",  
"58532", "58555", "58968", "58972", "59195", "59196", "59197", "60310", "60311",  
"60312", "61566", "61586", "63347", "63349", "63487", "63491", "63497", "63501",

"64826", "64827", "65206", "66226", "67448", "72089", "72091", "74246", "78400",  
 "78401", "78402", "79963", "NULL")  
 AND ETQE\_ID IN ("1107", "1115")  
 AND PROV\_ETQE\_ID IN ("1107", "1115")  
 AND SUBFIELD\_DESC IN ("Electrical Infrastructure Construction", "Engineering  
 and Related Design", "Fabrication and Extraction", "Information Technology and  
 Computer Sciences", "Language", "Manufacturing and Assembly", "Mathematical  
 Sciences")  
 AND PROVIDER\_TYPE\_DESC IN ("Employer")  
 AND FIELD\_DESC IN ("Communication Studies and Language", "Manufacturing,  
 Engineering and Technology", "Physical, Mathematical, Computer and Life Sciences")  
 AND PRIMARY\_ETQE\_DESC IN ("Primary ETQE of provider")

- Cluster 5

% of records: 12.01%

Average probability: 0.9179

Rule:

ASSESSOR\_ID IN ("NULL")  
 AND PROVIDER\_ID IN ("12666", "12916", "13248", "15316", "15841", "16886",  
 "1726", "1915", "20933", "22680", "28156", "28222", "28406", "30104", "30794",  
 "31724", "32915", "35608", "36340", "37171", "37904", "38989", "39546", "39904",  
 "41482", "41557", "43923", "44003", "46423", "49421")  
 AND QUALIFICATION\_ID IN ("20830", "23671", "24510", "48932", "49297",  
 "49414", "49622", "49665", "49708", "49709", "49769", "50098", "50326", "57841",  
 "58223", "58325", "58798", "59382", "60286", "63426", "64407", "71767", "72026",  
 "74647", "NULL")  
 AND UNIT\_STANDARD\_ID IN ("10024", "10025", "10054", "10055", "10061",  
 "10165", "10186", "10187", "10188", "10366", "10370", "10371", "10375", "10995",  
 "10997", "11000", "11002", "110061", "110070", "110545", "11258", "11303",  
 "113869", "113926", "113932", "113983", "114243", "114290", "114291", "114602",  
 "114606", "114615", "114896", "114899", "11490", "114902", "114903", "114904",  
 "114906", "114907", "114908", "114909", "114910", "114911", "114912", "114913",  
 "114914", "114915", "114916", "114917", "114918", "114919", "114920", "114921",  
 "114922", "114923", "114924", "114925", "114926", "114927", "114928", "114929",

"114936", "114958", "114974", "114991", "114993", "115110", "115808", "115847",  
"115895", "116097", "116274", "116292", "116397", "116541", "116944", "116947",  
"116948", "116949", "116954", "116959", "116960", "116962", "117016", "117172",  
"117173", "117433", "117512", "117887", "117888", "117894", "117904", "117908",  
"117914", "117915", "117916", "117917", "117918", "117919", "117940", "117941",  
"117942", "117960", "118045", "118050", "11830", "119095", "11923", "11924",  
"11926", "11928", "119358", "119360", "119368", "119379", "119381", "119390",  
"119471", "119473", "119474", "119475", "119476", "119477", "119479", "119480",  
"119482", "119483", "119484", "119486", "119488", "119489", "119570", "119571",  
"119573", "119575", "119648", "119652", "119653", "119657", "119683", "119691",  
"119693", "119838", "119839", "119846", "119847", "119930", "119973", "119974",  
"119975", "119976", "119977", "119978", "119979", "120317", "120320", "120322",  
"120324", "120513", "12053", "12152", "12155", "12156", "12157", "12170",  
"12172", "123248", "123385", "123386", "123389", "12450", "12461", "12472",  
"12478", "12480", "12500", "12501", "13184", "13234", "13237", "13238", "13240",  
"13241", "13252", "13275", "13900", "13902", "13928", "13929", "13931", "13932",  
"13933", "13934", "13935", "13936", "13946", "13947", "13948", "13949", "13952",  
"13958", "13964", "13968", "13969", "13971", "14036", "14067", "14358", "14359",  
"14365", "14376", "14444", "14461", "14462", "14551", "14568", "14586", "14673",  
"14676", "14678", "14681", "14682", "14684", "14900", "14928", "15108", "15109",  
"15111", "15231", "15232", "15233", "15234", "15236", "15237", "15238", "15239",  
"15241", "15242", "15245", "15246", "15247", "15249", "15250", "15251", "15252",  
"15254", "242685", "242827", "242828", "242830", "242831", "242832", "242833",  
"242834", "242835", "242836", "242837", "242838", "242839", "242841", "242846",  
"242871", "242873", "242874", "242875", "242877", "242879", "242880", "242881",  
"242882", "242883", "242887", "242891", "242917", "242993", "243013", "243035",  
"243688", "243689", "243690", "243693", "243695", "243696", "243697", "243698",  
"243729", "243820", "243821", "243822", "243823", "243824", "243825", "243826",  
"243827", "243834", "243840", "243841", "244512", "244524", "244591", "244606",  
"252037", "252038", "252039", "252041", "252042", "252043", "252044", "252046",  
"252048", "252049", "252051", "252052", "252053", "252054", "252057", "252058",  
"252059", "252060", "252219", "252220", "252227", "252228", "252259", "254611",  
"254612", "254613", "255517", "258172", "258173", "258174", "258175", "258176",  
"258177", "258178", "258179", "258192", "258193", "258194", "258195", "258196",

"258232", "258233", "258234", "258235", "258236", "258237", "258238", "259954",  
"259955", "259956", "335931", "7465", "7466", "7467", "7469", "7473", "7481",  
"7485", "7486", "7497", "7584", "7587", "7865", "7880", "7893", "7899", "8511",  
"8664", "8979", "8980", "8981", "8984", "8985", "8986", "8987", "8989", "8990",  
"8991", "8992", "8993", "8996", "9024", "9025", "9026", "9027", "9029", "9030",  
"9032", "9033", "9319", "9320", "9374", "9523", "9549", "9899", "9981", "9990")  
AND PROV\_ETQE\_ID IN ("1106", "1110", "1111", "1116", "1120", "1125", "1126")  
AND PROV\_PROVINCE\_DESC IN ("Eastern Cape", "Free State", "Gauteng",  
"Mpumalanga", "Undefined", "Western Cape")  
AND SUBFIELD\_DESC IN ("Building Construction", "Finance, Economics and Accounting",  
"Generic Management", "Human Resources", "Information Technology and Computer  
Sciences", "Language", "Manufacturing and Assembly", "Marketing", "Mathematical  
Sciences", "Office Administration", "People/Human-Centred Development", "Public  
Administration", "Wholesale and Retail")  
AND ETQE\_ID IN ("1075", "1103", "1110", "1114", "1118", "1126")  
AND 1 <= END\_PROV\_IND <= 43.3  
AND 0 <= START\_PROV\_IND <= 52.8

- Cluster 6

% of records: 7.03%

Average probability: 0.9316

Rule:

ASSESSOR\_ID IN ("NULL")

AND PROVIDER\_ID IN ("10053", "10377", "10670", "11691", "12743", "12888",  
"13654", "14741", "15151", "17425", "18274", "18564", "19858", "20357", "21985",  
"22273", "22883", "22927", "23501", "2627", "2741", "28710", "28833", "28868",  
"29433", "29441", "30139", "30204", "30216", "30217", "31909", "32518", "32915",  
"34432", "34485", "34674", "37334", "37520", "37747", "38282", "38989", "39271",  
"39703", "39728", "39735", "41565", "42072", "42373", "42987", "43139", "43177",  
"46423", "49381", "49533", "5994")  
AND QUALIFICATION\_ID IN ("22459", "22787", "23270", "23671", "23850",  
"24150", "35945", "48987", "48993", "49614", "49623", "49665", "49946", "50098",  
"50302", "57934", "58026", "58798", "58799", "58802", "59317", "59382", "64386",  
"64826", "64827", "65206", "65208", "65426", "66286", "67626", "NULL")

AND UNIT\_STANDARD\_ID IN ("10019", "10020", "10039", "10054", "10152",  
"10156", "10163", "10246", "10269", "10341", "110016", "110020", "110026",  
"110038", "110040", "110043", "110081", "123274", "123275", "123277", "123278",  
"12434", "12461", "12472", "12473", "12478", "12479", "12480", "12482", "12483",  
"12486", "12487", "12488", "12490", "12493", "12500", "12501", "12554", "13153",  
"13174", "13176", "13179", "13182", "13184", "13186", "13188", "13189", "13191",  
"13193", "13231", "13234", "13237", "13239", "13251", "13431", "13928", "13929",  
"13931", "13932", "13933", "13934", "13935", "13945", "13947", "13949", "13951",  
"13953", "13958", "13960", "13962", "13964", "13965", "13994", "14012", "14013",  
"14015", "14101", "14241", "14243", "14355", "14356", "14357", "14358", "14359",  
"14360", "14361", "14363", "14365", "14366", "14369", "14370", "14372", "14376",  
"14508", "14510", "14511", "14569", "14597", "14649", "14667", "14671", "14673",  
"14674", "14679", "14682", "14684", "14689", "14690", "14691", "14693", "14696",  
"14699", "14700", "14730", "14899", "15110", "15140", "15251", "242827",  
"242828", "242836", "243073", "243080", "243223", "243310", "243311", "243312",  
"243313", "243314", "243315", "243316", "243317", "243318", "243319", "243320",  
"243768", "243774", "244595", "252210", "252214", "252217", "252231", "252235",  
"252267", "254237", "256616", "259621", "261674", "261675", "261676", "261677",  
"261678", "261679", "261680", "261681", "261682", "261683", "261694", "261695",  
"261696", "261697", "261698", "261714", "261734", "261754", "263701", "263703",  
"7464", "7465", "7466", "7467", "7468", "7469", "7473", "7477", "7478", "7480",  
"7481", "7485", "7486", "7497", "7506", "7524", "7525", "7526", "7528", "7530",  
"7564", "7583", "7585", "7588", "7590", "7807", "7808", "8014", "8055", "8056",  
"8121", "8437", "8635", "8979", "8980", "8981", "8982", "8984", "8985", "8986",  
"8987", "8989", "8990", "8991", "8992", "8993", "8996", "9024", "9025", "9026",  
"9027", "9029", "9030", "9032", "9033", "9259", "9285", "9319", "9320", "9339",  
"9374", "9460", "9545", "9861", "9862", "9864", "9865", "9867", "9870", "9872",  
"9875", "9876", "9891", "9892", "9895", "9897", "9977", "9981", "9986", "9988",  
"9990", "110092", "110097", "11248", "11252", "11263", "113869", "113926",  
"113941", "113983", "114021", "11411", "11415", "11418", "11423", "114232",  
"114243", "11432", "114495", "114606", "114747", "114755", "114906", "114908",  
"114913", "114917", "114923", "114929", "114931", "114935", "114937", "114940",  
"114944", "114945", "114946", "114948", "114949", "114958", "114960", "114963",  
"114969", "114976", "114991", "115089", "115118", "115410", "115840", "115872",

"116181", "116183", "116189", "116217", "116218", "116221", "116223", "116270",  
"116537", "116550", "116551", "116655", "116660", "116737", "116947", "116948",  
"116949", "116952", "116953", "116954", "116955", "116957", "116962", "116983",  
"117016", "117128", "117188", "117258", "117261", "117437", "117516", "117517",  
"117850", "117884", "117919", "117940", "117941", "117942", "119095", "119379",  
"119381", "119390", "119471", "119473", "119474", "119476", "119477", "119479",  
"119480", "119482", "119483", "119484", "119486", "119488", "119489", "119584",  
"119648", "119652", "119653", "119657", "119683", "119684", "119685", "119686",  
"119687", "119688", "119689", "119690", "119691", "119729", "119730", "119761",  
"119767", "120324", "120325", "120327", "120328", "120329", "120378", "120383",  
"120411", "12170", "12171", "12172", "12173", "12232", "12233", "12235",  
"12236", "12257", "12258", "12260")

AND PROV\_ETQE\_ID IN ("1102", "1103", "1105", "1126")

AND ETQE\_ID IN ("1102", "1103", "1109", "1122", "1126", "1127")

AND SUBFIELD\_DESC IN ("Building Construction", "Engineering and Related Design",  
"Finance, Economics and Accounting", "Generic Management", "Hospitality, Tourism, Travel,  
Gaming and Leisure", "Information Technology and Computer Sciences", "Language",  
"Manufacturing and Assembly", "Mathematical Sciences", "Office Administration",  
"People/Human-Centred Development", "Primary Agriculture", "Sport", "Transport,  
Operations and Logistics", "Wholesale and Retail")

AND PROVIDER\_TYPE\_DESC IN ("Education and Training")

AND PROVIDER\_CLASS\_DESC IN ("Mixed: Public and Private")

AND 1 <= END\_PROV\_IND <= 85.6

- Cluster 7

% of records: 6.55%

Average probability: 0.9826

Rule:

ASSESSOR\_ID IN ("NULL")

AND PROVIDER\_ID IN ("14764", "1915", "20933", "22927", "37520")

AND QUALIFICATION\_ID IN ("49106", "49852", "49946", "57625", "57934",  
"NULL")

AND SUBFIELD\_DESC IN ("Finance, Economics and Accounting")

AND UNIT\_STANDARD\_ID IN ("10396", "113920", "113924", "113926", "113928",  
"113929", "113931", "113940", "113941", "113945", "114223", "114226", "114232",

"114755", "114977", "114981", "114983", "114987", "114992", "114994", "114995",  
 "114996", "114997", "115000", "115001", "115002", "117128", "117134", "117138",  
 "117143", "117144", "117145", "117146", "117147", "117149", "117150", "117151",  
 "117152", "117163", "117166", "117172", "117173", "117175", "117188", "117258",  
 "117261", "118022", "119277", "119282", "119693", "119698", "119699", "119911",  
 "119916", "119921", "119926", "119932", "120014", "120015", "120022", "120023",  
 "120025", "120026", "120028", "120029", "120030", "120031", "120032", "120033",  
 "120034", "120035", "120036", "120037", "120039", "120040", "120043", "120092",  
 "120125", "120126", "120127", "120128", "120129", "120130", "120131", "120132",  
 "120133", "120134", "120135", "120136", "120137", "120138", "120139", "120140",  
 "120141", "120142", "120143", "120144", "120145", "120146", "120147", "120148",  
 "120149", "120150", "120151", "120152", "120153", "120154", "120155", "12181",  
 "12564", "12565", "12567", "13929", "13957", "13958", "13959", "13964", "13966",  
 "13970", "14334", "14336", "14523", "14996", "15001", "15008", "230087",  
 "230088", "230090", "230091", "230092", "230094", "230095", "242571", "242572",  
 "242573", "242574", "242575", "242576", "242577", "242578", "242579", "242580",  
 "242581", "242582", "242583", "242584", "242585", "242586", "242587", "242588",  
 "242589", "242590", "242591", "242592", "242593", "242594", "242595", "242596",  
 "242597", "242598", "242599", "242600", "242601", "242602", "242603", "242604",  
 "242605", "242606", "242607", "242608", "242609", "242610", "242611", "242612",  
 "242613", "242614", "242615", "242616", "242617", "242618", "242619", "242620",  
 "242621", "242622", "242623", "242624", "242625", "242626", "242627", "242628",  
 "242629", "242630", "242631", "242632", "242633", "242634", "242635", "242636",  
 "242672", "243159", "243165", "243170", "243171", "252054", "337080", "377930")  
 AND ETQE\_ID IN ("1127")  
 AND FIELD\_DESC IN ("Business, Commerce and Management Studies")  
 AND PROV\_ETQE\_ID IN ("1103", "1105", "1120", "1126")  
 AND ENROL\_TYPE\_DESC IN ("Mixed Mode", "Unknown")  
 AND ENROL\_STATUS\_DESC IN ("Achieved")

- Cluster 8

% of records: 3.08%

Average probability: 0.9893

Rule:

ENROL\_TYPE\_DESC IN ("Residential Learning (i.e. Contact Mode)")

AND UNIT\_STANDARD\_ID IN ("115449", "115770", "115776", "117882",  
"117888", "117891", "117894", "119474", "119476", "119479", "119482", "119484",  
"119486", "119488", "119489", "123414", "123415", "13660", "15234", "15235",  
"15238", "15245", "15249", "244276", "244479", "244485", "244486", "244489",  
"244492", "244495", "244497", "244498", "244501", "244502", "7485", "9032",  
"9033")

AND QUALIFICATION\_ID IN ("58778", "NULL")

AND PROVIDER\_ID IN ("11691", "12743", "13655", "14096", "14607", "14640",  
"15308", "16107", "22708", "26365", "26373", "26416", "28446", "29318", "37237",  
"37395", "39555")

AND ETQE\_ID IN ("1106")

AND ASSESSOR\_ID IN ("3002915", "3005226", "3008430", "3018056", "3018373",  
"3022011", "3027298", "3029329", "3030913", "3031172", "3031732", "3039611",  
"4631819", "4631884", "4631906", "4632003", "5013181", "5518282", "5518427",  
"6055458", "6055748", "6055792", "8145122", "8145323", "8145429", "9050095",  
"9050622", "9463919", "9574862", "9574962", "NULL", "16123834", "17374438",  
"3002886")

AND PROV\_ETQE\_ID IN ("1031", "1110", "1126")

AND SUBFIELD\_DESC IN ("Adult Learning", "Early Childhood Development",  
"Higher Education and Training", "Human Resources", "Language", "Mathematical  
Sciences")

AND PROV\_PROVINCE\_DESC IN ("Gauteng")

AND FIELD\_DESC IN ("Communication Studies and Language", "Education, Training and  
Development", "Physical, Mathematical, Computer and Life Sciences")

## Appendix L

This appendix provides a detailed review of learner enrolment records in relation to whether the provider was accredited to offer the qualification or unit standard for the duration of the learner's active enrolment on the qualification or unit standard. The review focuses on gaining a better understanding of data records that fall into specific categories of the data field PROV\_ACCRED\_IND (see Appendix E.3.7 and Appendix G.3.7).

The appendix was necessitated as a result of the scope and volume of records that infringe on this semantic business rule for qualification and unit standard enrolment records. As a result the structure of this appendix has sub sections that focus on the each of these types of enrolment records.

### ***L.1 Qualification enrolments***

#### ***L.1.1 No Accreditation***

This category indicates that the provider that is linked to the qualification has never had an active accreditation to offer the qualification. This category contains 42.29% of all of the records that infringe on this semantic business rule and this category is of greatest concern to SAQA.

All of the 29 discrete ETQEs in the dataset are linked to this category. Of these records, 79.03% were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 366 qualifications are linked to this category. Of these 366 qualifications, 10 qualifications contribute to 52.55% of records in this category. Most notably, although 20 of the 366 qualifications only constitute 1.20% of the records; the records for these qualifications represent 100% of the qualification enrolment records submitted to the NLRD for the qualifications.

Of the 5669 discrete providers in the dataset, 1279 providers are linked to this category. Of these 1279 providers, 10 providers contribute to 40.91% of the records. Most notably, 491 of the 1279 providers constitute 22.70% of the records; the records for these providers represent 100% of the qualification enrolment records submitted to the NLRD for the providers.

As already noted, this category indicates that the provider has never had an active accreditation to offer the qualification. As a result, providers in this category cannot be reported on in any of the other categories that form part of this research. However, other categories that the providers from this category can exist in are the ‘No Accreditation (Qual Linked to Lshp)’, ‘No Accreditation Predicted’ and ‘ ‘ No Accreditation Predicted (Qual Linked to Lshp)’ categories that have been excluded from this research.

The reader should therefore note that the above statement “...491 of the 1279 providers constitute 22.70% of the records; the records for these providers represent 100% of the qualification enrolment records submitted to the NLRD for the providers.” must further be interpreted to mean that for the remaining 788 providers, 100% of the records for the specific provider fall into the categories ‘No Accreditation’, ‘No Accreditation (Qual Linked to Lshp)’, ‘No Accreditation Predicted’ or ‘ ‘ No Accreditation Predicted (Qual Linked to Lshp)’.

Unlike any of the other categories for this semantic business rule, it is unlikely that qualification enrolment records that appear in the ‘No Accreditation’ category do so as a result of data capturing issues related to the enrolment record. Rather the data capturing or data quality issues reside in the lack of an accreditation of a provider to offer a qualification.

Table L.1.1.1 provides an overview of the records found in this category grouped by submitting ETQE. The table differentiates the number of providers that have the submitting ETQE as their primary ETQE (“Primary ETQE of provider”) and the number of providers where the submitting ETQE is not the primary ETQE of the provider (“Not Primary ETQE of provider”). A submitting ETQE may utilize another ETQE’s providers (Section 3.8.3.5). It is found the same providers that are not accredited have been utilized by more than one ETQE and as a result the count of provider by submitting ETQE is 1309, whereas there are only 1288 discrete providers in this category.

Table L.1.1.1 ‘No Accreditation’ records by submitting ETQE identifier, count of Not Primary ETQE providers, count of Primary ETQE of provider and % qualification enrolment records in the category

Submitting ETQE Identifier	Not Primary ETQE of Provider		Primary ETQE of Provider		Total	
	Count of Provider	% Records	Count of Provider	% Records	Count of Provider	% Records
1034	1	0.21%		0.00%	1	0.21%
1075	1	0.01%	4	1.82%	5	1.82%
1079	4	0.03%	50	0.30%	54	0.33%
1102		0.00%	12	0.14%	12	0.14%
1103	36	0.27%	86	0.35%	122	0.62%
1104		0.00%	3	0.02%	3	0.02%
1105		0.00%	52	29.93%	52	29.93%
1106	4	0.29%	5	0.01%	9	0.30%
1107	6	0.02%	10	0.13%	16	0.15%
1108	1	0.02%	2	1.48%	3	1.51%
1109	7	0.42%	17	0.97%	24	1.39%
1110		0.00%	1	0.02%	1	0.02%
1111	1	0.01%	61	2.45%	62	2.45%
1112	7	0.44%	29	0.89%	36	1.33%
1113		0.00%	41	0.82%	41	0.82%
1114	14	1.36%	23	1.91%	37	3.27%
1115	2	0.02%	23	6.99%	25	7.00%
1116	19	0.14%	419	2.85%	438	2.99%
1117		0.00%	11	0.13%	11	0.13%
1118		0.00%	4	0.06%	4	0.06%
1119	1	0.00%	65	1.83%	66	1.83%
1120	1	0.06%	3	0.02%	4	0.08%
1121		0.00%	2	0.01%	2	0.01%
1122		0.00%	12	0.09%	12	0.09%
1123	1	0.00%	8	0.13%	9	0.13%
1124		0.00%	6	0.08%	6	0.08%
1125		0.00%	42	0.97%	42	0.97%
1126	23	4.16%	163	37.94%	186	42.09%
1127	10	0.07%	16	0.15%	26	0.22%
<b>Grand Total</b>	<b>139</b>	<b>7.51%</b>	<b>1170</b>	<b>92.49%</b>	<b>1309</b>	<b>100.00%</b>

The reader should note that in the case of the overall implementation of a provider accreditation (see Section 4.4.2) it is understandable that an ETQE that is not the primary ETQE of the provider may not be aware that a specific provider has not been awarded an active accreditation by the provider’s primary ETQE. The implementation of accreditations to offer a qualification is however different in that it is at the discretion of the accrediting ETQE to accredit a provider to offer its qualifications.

Analysis of Table L.1.1.1 shows the following notable trends:

- ETQE identifier 1116 has the highest incidence of the number of providers (438) that are not accredited to offer the qualification. Further analysis shows that 58.11% of the providers linked to qualification enrolment records for this ETQE have never been accredited to offer the qualification.

- ETQE identifier 1126 has the second highest incidence of the number of providers (186) that are not accredited to offer the qualification. Further analysis shows that 56.62% of the providers linked to qualification enrolment records for this ETQE have never been accredited to offer the qualification. This ETQE also had the highest incidence of enrolment records in this category (42.09%).
- ETQE identifier 1103 has the third highest incidence of the number of providers (122) that are not accredited to offer the qualification. Further analysis shows that 11.14% of the providers linked to qualification enrolment records for this ETQE have never been accredited to offer the qualification.

### ***L.1.2 Start Before, End During***

This category indicates that the qualification enrolment started before the provider was accredited to offer the qualification and either was achieved or expired whilst the provider was accredited to offer the qualification. This category contains 24.09% of all of the records that infringe on this semantic business rule.

All of the 29 discrete ETQEs in the dataset are linked to this category. Of these records, 50.82% were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 359 qualifications are linked to this category. Of these 359 qualifications, 10 qualifications contribute to 42.16% of records in this category. Most notably, although 4 of the 359 qualifications only constitute 0.11% of the records; the records for these qualifications represent 100% of the qualification enrolment records submitted to the NLRD for the qualifications.

Of the 5669 discrete providers in the dataset, 646 providers are linked to this category. Of these 646 providers, 10 providers contribute to 41.08% of the records. Most notably, although 35 of the 646 providers only constitute 1.38% of the records; the records for these providers represent 100% of the qualification enrolment records submitted to the NLRD for the providers.

The high densities of records that infringe on this rule, in combination with the high percentage of records in this category for some providers, hints at possible systemic issues in regard to provider accreditations to offer qualifications. The patterns in the data are

however not clearly delineated and a further review of qualification enrolment that fall into this category needs to be conducted (refer to Appendix L.1.7).

### ***L.1.3 Start Before, End Before***

This category indicates that the qualification enrolment both started before and either was achieved or expired before the provider was accredited to offer the qualification. This category contains 21.26% of all of the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 27 ETQEs are linked to this category. Of these records, 60.44% were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 258 qualifications are linked to this category. Of these 258 qualifications, 10 qualifications contribute to 55.37% of records in this category. Most notably, although 5 of the 258 qualifications only constitute 0.48% of the records; the records for these qualifications represent 100% of the qualification enrolment records submitted to the NLRD for the qualifications.

Of the 5669 discrete providers in the dataset, 485 providers are linked to this category. Of these 485 providers, 10 providers contribute to 49.66% of the records. Most notably, although 45 of the 485 providers only constitute 2.55% of the records; the records for these providers represent 100% of the qualification enrolment records submitted to the NLRD for the providers.

The high densities of records that infringe on this rule, in combination with the high percentage of records in this category for some providers, hints at possible systemic issues in regard to provider accreditations to offer qualifications. The patterns in the data are however not clearly delineated and a further review of qualification enrolment records that fall into this category was conducted (refer to Appendix L.1.7).

Cross checking the results in this category with the results in the ‘Start Before, End During’ (Appendix L.1.2) category shows some remarkable similarities. The first of which was that three ETQEs identified as top contributors to the ‘Start Before, End During’ category are the top contributors to this category. A precursory review also revealed that 40% of the top 10 providers that contributed to the ‘Start Before, End During’ category are top 10

contributors in this category. These similarities suggested that a further review conducted on these records should be conducted in conjunction with the records in the ‘Start Before, End During’ category.

#### ***L.1.4 Start During, End After***

This category indicates that the qualification enrolment started whilst the provider was accredited to offer the qualification, and either was achieved or expired after the provider was no longer accredited to offer the qualification. This category contains 8.49% of all of the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 26 ETQEs are linked to this category. Of these records, 63.69% were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 206 qualifications are linked to this category. Of these 206 qualifications, 10 qualifications contribute to 67.31% of records in this category. Most notably, although 2 of the 206 qualifications only constitutes 0.12% of the records; the records for these qualifications represent 100% of the qualification enrolment records submitted to the NLRD for the qualification.

Of the 5669 discrete providers in the dataset, 944 providers are linked to this category. Of these 944 providers, 10 providers contribute to 20.68% of the records. Most notably, although 74 of the 944 providers only contribute 2.81% of the records; the records for these providers represent 100% of the qualification enrolment records submitted to the NLRD for the providers.

The high densities of records that infringe on this rule in relation to the number of qualification enrolment records submitted to the NLRD for the provider hints at possible systemic issues in regard to provider accreditations to offer the qualification. The patterns in the data are however not clearly delineated and a further review of qualification enrolment records submitted to the NLRD for these providers needed to be conducted (refer to Appendix L.1.8).

#### ***L.1.5 Start After, End After***

This category indicates that the qualification enrolment both started after and either was achieved or expired after the provider was accredited to offer the qualification. This category contains 3.77% of all of the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 25 ETQEs are linked to this category. Of these records, 48.73% were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 129 qualifications are linked to this category. Of these 129 qualifications, 10 qualifications contribute to 63.35% of records in this category. Most notably, although 3 of the 129 qualifications only constitute 0.58% of the records; the records for these qualifications represent 100% of the qualification enrolment records submitted to the NLRD for the qualifications.

Of the 5669 discrete providers in the dataset, 171 providers are linked to this category. Of these 171 providers, 10 providers contribute to 58.58% of the records. Most notably, although 26 of the 171 providers only contribute 4.72% of the records; the records for these providers represent 100% of the qualification enrolment records submitted to the NLRD for the providers.

The high densities of records that infringe on this rule in relation to the number of qualification enrolment records submitted to the NLRD for the provider hints at possible systemic issues in regard to provider accreditations to offer the qualification. The patterns in the data are however not clearly delineated and a further review of qualification enrolment records submitted to the NLRD for these providers needed to be conducted (refer to Appendix L.1.8).

Cross checking the results in this category with the results in the 'Start During, End After' (Appendix L.1.4) category shows some similarities. The first of which is that one of the ETQEs identified as top contributors to the 'Start During, End After' category is a top contributor to this category. A precursory review also revealed that 10% of the top 10 providers that contributed to the 'Start During, End After' category are top 10 contributors in this category. These similarities suggested that a further review conducted on these records should be conducted in conjunction with the records in the 'Start During, End After' category.

### ***L.1.6 Start Before, End After***

This category indicates that the qualification enrolment both started before the provider was accredited to offer the qualification and either was achieved or expired after the provider was no longer accredited to offer the qualification. This category contains 0.11% of all of the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 9 ETQEs are linked to this category. Of these records, 84.93% were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 18 qualifications are linked to this category. Of these 18 qualifications, 10 qualifications contribute to 94.98% of records in this category.

Of the 5669 discrete providers in the dataset, 58 providers are linked to this category. Of these 58 providers, 10 providers contribute to 62.10% of the records. Most notably, although 17 of the 58 providers only contribute 19.18% of the records; the records for these providers represent 100% of the qualification enrolment records submitted to the NLRD for the providers.

The low incidence of records that fall into this category suggests that these records are found in this category as a result of incorrect data capturing at the source of the data.

### ***L.1.7 Start Before, End Before or End During***

As stated in Appendix L.1.2 and L.1.3, the high density of records submitted to the NLRD for the providers in the categories ‘Start Before, End Before’ and ‘Start Before, End During’, in conjunction with intersections in the top 3 ranked ETQEs and top 10 ranked providers in these categories, suggested that further analysis of these records as a single data set should be conducted.

Further analysis found that these two initial categories shared 215 qualifications and 353 providers. As a result, the categories ‘Start Before, End Before’ and ‘Start Before, End During’ were grouped into a category called ‘Start Before, End Before or End During’ for this analysis. This category indicates that the qualification enrolment started before the

provider was accredited to offer the qualification and either was achieved or expired before or whilst the provider was accredited to offer the qualification. As a result of this consolidation the 'Start Before, End Before or End During' category contains 45.35% of all the records that infringe on this semantic business rule.

All of the 29 ETQEs in the dataset are linked to this category. Slightly more than 55% of these records were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 402 qualifications are linked to this category. Of these 402 qualifications, 10 qualifications contribute to 45.68% of records in this category. Most notably, although 11 of the 402 qualifications only constitute 1.01% of the records; the records for this qualification represent 100% of the qualification enrolment records submitted to the NLRD for the qualifications.

Of the 5669 discrete providers in the dataset, 778 providers are linked to this category. Of these 778 providers, 10 providers contribute to 41.06% of the records. Most notably, although 116 of the 778 providers only constitute 4.02% of the records; the records for these providers represent 100% of the qualification enrolment records submitted to the NLRD for the providers.

The volume of records found in this consolidated category constituted nearly 9% of the total qualification enrolment records that form part of the research. The implementation of data mining techniques, beyond exploratory data mining techniques, is needed to provide SAQA with both a succinct description of these records and to attempt to identify any data records that exist in this category as a result of possible incorrect data capturing at the source of the data.

The clustering data mining technique, as described in Appendix I.3, is applied to this data set in an effort to achieve both these aims.

An initial review of the data in this category, in combination with the understanding that the provider linked to a qualification enrolment record may have a primary ETQE that differs from the ETQE that submitted the qualification enrolment record to the NLRD, prompted the implementation of a new data field on the data set prior to data mining. The indicator

PRIMARY\_ETQE\_DESC was developed as a nominal data value that contains the value 'Primary ETQE of provider' if the ETQE identifier of the ETQE that submitted the qualification enrolment record to the NLRD (ETQE\_ID) was the same as the ETQE identifier of the primary ETQE of the provider (PROV\_ETQE\_ID). The same indicator would have the value 'Not Primary ETQE of provider' if the primary ETQE identifier of the provider differed from the ETQE identifier of the ETQE that submitted the qualification record to the NLRD. It was hoped that the implementation of the PRIMARY\_ETQE\_DESC indicator would allow the data mining algorithm to find discrete patterns in the data related to the utilization of providers that are accredited by an ETQE other than the ETQE that submitted the qualification enrolment record to the NLRD.

The data mining effort results in the development of 8 data clusters (a technical description of each cluster is provided in Appendix M.1.1) that have the following most dominant characteristics:

1. Cluster 1

This cluster describes more than 25% of the records. The cluster is predominantly described as containing qualification enrolment records for 17 qualifications as offered by 22 providers. The records were submitted to the NLRD by ETQE identifier 1126.

2. Cluster 2

This cluster describes nearly 16% of the records. The cluster is relatively diverse in that it describes records submitted by 8 ETQEs, covering 32 different qualifications offered by 58 different providers.

3. Cluster 3

This cluster describes slightly more than 14% of the records. The cluster is diverse in that it describes records submitted by 10 ETQEs, covering 65 different qualifications offered by 88 different providers.

4. Cluster 4

This cluster has a probability of 1 and describes nearly 13.5% of the records. These records were submitted to the NLRD by ETQE identifier 1105 and encompass 4 different qualifications, all with a subfield of 'Safety in Society'. The qualifications were offered by 26 providers all of which have ETQE identifier 1105 as their primary ETQE.

5. Cluster 5

This cluster describes nearly 13% of the records as belonging to 5 qualifications, all with a subfield of 'Finance, Economics and Accounting', offered by 3 providers. The

records were submitted to the NLRD by ETQE identifier 1116 and the same ETQE is the primary ETQE of the providers that offered these qualifications.

6. Cluster 6

The cluster describes nearly 8.5% of the records. The cluster is relatively diverse in that it describes records submitted by 9 ETQEs, covering 21 qualifications offered by 31 providers.

7. Cluster 7

This cluster describes more than 5.5% of the records as being submitted to the NLRD by ETQE identifier 1106. All of these records belong to qualification identifier 58778 which has a subfield of 'Early Childhood Development'. The qualification was offered by 8 providers.

8. Cluster 8

The cluster describes nearly 4.5% of the records as having been submitted to the NLRD by ETQE identifier 1106. These records encompass 4 qualifications offered by 12 providers. The qualifications have a subfield of 'Early Childhood Development' or 'Adult Learning'.

As stated before, this category contains records from 29 different ETQEs. The above description of the 8 clusters generated by the clustering algorithm shows that 5 of these clusters (clusters 1, 4, 5, 7 and 8) each describe records that were submitted to the NLRD by a specific ETQE.

The clustering algorithm does not generate any clusters with a population of less than 1%. A review of the probabilities allocated to each record in a cluster does however allow for the identification of records that are low enough to be considered anomalous as per the criteria defined in Appendix I.3. These records constitute 1.23% of the records found in this category, and possibly exist in this category as a result of data capturing problems at the source of the data.

### ***L.1.8 Start During, Start After and End After***

As stated in Appendix L.1.4 and L.1.5, the high density of records submitted to the NLRD for the providers in the categories 'Start During, End After' and 'Start After, End After', in conjunction with intersections in the top 3 ranked ETQEs and top 10 ranked providers in

these categories, suggested that further analysis of these records as a single data set should be conducted.

Further analysis found that these two initial categories shared 93 qualifications and 80 providers. As a result, the categories 'Start During, End After' and 'Start After, End After' were grouped into a category called 'Start During, Start After and End After' for this analysis. This category indicates that the qualification enrolment started during or after the provider was accredited and either was achieved or expired after the provider was no longer accredited. As a result of this consolidation the 'Start During, Start After and End After' category contains 12.25% of all the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 28 ETQEs are linked to this category. More than 52% of these records were submitted to the NLRD by 3 ETQEs.

Of the 861 discrete qualifications in the dataset, 242 qualifications are linked to this category. Of these 242 qualifications, 10 qualifications contribute to 61.35% of records in this category. Most notably, although 5 of the 242 qualifications only constitute 0.26% of the records; the records for these qualifications represent 100% of the qualification enrolment records submitted to the NLRD for the qualifications.

Of the 5569 discrete providers in the dataset, 1035 providers are linked to this category. Of these 1035 providers, 10 providers contribute to 24.52% of the records. Most notably, although 104 of the 1035 providers only constitute 3.81% of the records; the records for these providers represent 100% of the qualification enrolment records submitted to the NLRD for the providers.

The volume of records found in this consolidated category constitutes 2.41% of the total qualification enrolment records that form part of the research. The implementation of data mining techniques, beyond exploratory data mining techniques, is needed to provide SAQA with both a succinct description of these records and to attempt to identify any data records that exist in this category as a result of possible incorrect data capturing at the source of the data.

The clustering data mining technique, as described in Appendix I.3, is applied to this data set in an effort to achieve both these aims.

An initial review of the data in this category, in combination with the understanding that the provider linked to a qualification enrolment record may have a primary ETQE that differs from the ETQE that submitted the qualification enrolment record to the NLRD, prompted the implementation of a new data field on the data set prior to data mining. The indicator PRIMARY\_ETQE\_DESC was developed as a nominal data value that contains the value 'Primary ETQE of provider' if the ETQE identifier of the ETQE that submitted the qualification enrolment record to the NLRD (ETQE\_ID) was the same as the ETQE identifier of the primary ETQE of the provider (PROV\_ETQE\_ID). The same indicator would have the value 'Not Primary ETQE of provider' if the primary ETQE identifier of the provider differed from the ETQE identifier of the ETQE that submitted the qualification record to the NLRD. It was hoped that the implementation of the PRIMARY\_ETQE\_DESC indicator would allow the data mining algorithm to find discrete patterns in the data related to whether or not the ETQE submitted the data was primary ETQE of the provider that offered the qualification.

The data mining effort results in the development of 8 data clusters (a technical description of each cluster is provided in Appendix M.1.2) that have the following most dominant characteristics:

1. Cluster 1

This cluster describes slightly more than 31% of the records. The cluster predominantly describes enrolments against qualification identifier 48930 as offered by 238 providers. These records were submitted to the NLRD by ETQE identifier 1079. Further analysis found that the 238 providers referred to in these enrolment records constitute nearly 28% of the providers that offered this qualification.

2. Cluster 2

The cluster describes nearly 19% of the records as belonging to 7 qualifications offered by 14 providers. The records in this cluster were submitted to the NLRD by 3 different ETQEs (ETQE identifiers 1075, 1120 and 1127). The qualifications in this cluster generally have subfield descriptions of 'Human Resources' or 'Finance, Economics and Accounting'.

3. Cluster 3

This cluster describes nearly 11% of the records. The cluster is relatively diverse in that it describes records covering 29 different qualifications, offered by 39 different providers. The majority of these enrolments commenced within 10 months of the provider's accreditation to offer the qualification expired.

4. Cluster 4

This cluster describes nearly 10.5% of the records as being submitted to the NLRD by 10 different ETQEs (ETQE identifiers 1126, 1125, 1122, 1116, 1112, 1111, 1109, 1108, 1107 and 1103). The cluster is diverse in that it comprises of 43 qualifications offered by 50 providers. The majority of the 43 qualifications have a qualification class of 'Regular-Unit Stds Based'. Further, the majority of the providers have a provider class of 'Private'.

5. Cluster 5

The cluster is relatively diverse and describes slightly more than 9% of the records as having been submitted to the NLRD by 9 different ETQEs (ETQE identifiers 1126, 1117, 1113, 1111, 1109, 1107, 1105, 1103 and 1102). The cluster encompasses 36 qualifications offered by 47 providers.

6. Cluster 6

The cluster describes slightly more than 7% of the records as belonging to 5 qualifications that have a type of 'National Higher Certificate' or 'National Diploma'. The qualifications are offered by 12 providers and the enrolment records were submitted to the NLRD by ETQE identifier 1106.

7. Cluster 7

This cluster describes nearly 6.5% of the records as belonging to 1 qualification with identifier 58788. This qualification was offered by 7 providers and all of these records were submitted to the NLRD by ETQE identifier 1106.

8. Cluster 8

The cluster describes slightly more than 6% of the records as belonging to 3 providers that offered 12 qualifications that all have a subfield of 'Engineering and Related Design'. All of these enrolment records were submitted to the NLRD by ETQE identifier 1115.

Of the 8 clusters generated 4 provide a very discrete description of the characteristics of the records found in the cluster. The most notable clusters that are generated for this category are clusters 1, 6, 7 and 8. Each of these clusters points to problems related either to specific

qualifications, providers and/or ETQEs. None of the clusters seemed to indicate a trend in regard to the utilization of providers whose primary ETQE is other than that of the submitting ETQE.

The clustering algorithm does not generate any clusters with a population of less than 1%. A review of the probabilities allocated to each record in a cluster does however allow for the identification of records that are low enough to be considered anomalous as per the criteria defined in Appendix I.3. These records constitute 0.81% of the records found in this category, and possibly exist in this category as a result of data capturing problems at the source of the data.

#### ***L.1.9 Summary of semantic infringements by ETQE***

The preceding sections provide the results of records that infringe on this semantic business rule from the granular perspective of the qualification enrolment record in relation to the complete dataset. This approach supports the determination of patterns within the data that point to systemic and anomalous problems within the overall dataset, which in turn lends itself to assessing the quality of the data in the data set.

The approach however ignores the diverse nature of ETQEs, and in particular the volume of the records that each ETQE submits to the NLRD. The final step in the analysis of this semantic business rule provides an overview of the percentage of records, calculated as a percentage of the number of records submitted by the ETQE, which infringe on this semantic business rule.

The results are presented as the percentage of records submitted by the ETQE that fall into a category that describes a semantic business rule issue (see Table L.1.9.1):

Table L.1.9.1 % of records submitted by an ETQE that have a category that describes a semantic business rule issue

ETQE Identifier	% Semantic Rule Issue
1116	56.77%
1075	46.47%
1079	43.89%
1105	40.84%
1121	32.08%
1126	30.54%
1110	24.55%
1108	23.38%
1114	17.68%
1125	16.63%
1115	15.85%
1109	14.94%
1118	14.66%
1124	14.37%
1111	14.12%
1120	11.09%
1106	10.70%
1104	10.65%
1102	9.19%
1119	8.70%
1112	8.41%
1127	8.23%
1107	8.11%
1113	7.55%
1123	6.71%
1103	6.51%
1117	5.66%
1034	4.15%
1122	3.74%

The results clearly illustrate that the infringement of this semantic business rule could be considered systemic at a number of the ETQEs.

#### ***L.1.10 Conclusion***

The analysis of qualification enrolment records in regard to whether the provider was accredited to offer the qualification for the duration of the learner’s active enrolment highlights the possibility of systemic issues in regard to provider accreditations.

The cluster analysis for the ‘Start Before, End Before or End During’ and ‘Start During, Start After and End After’ categories is able to provide a clear description of the data in the categories. Further, a comparison across the two cluster analyses shows that ETQE

identifier 1106 is featured in both categories. The analysis of the ‘No Accreditation’ category highlights possible systemic issues in regard to provider accreditations as implemented by ETQE identifiers 1116, 1126 and 1103.

The cluster analysis of both the ‘Start Before, End Before or End During’ and ‘Start During, Start After and End After’ categories is able to identify records that may exist in these categories as a result of incorrect data capturing on the qualification enrolment record. The analysis of the ‘Start Before, End After’ category in turn allows for the identification of enrolment records that have possibly been captured incorrectly.

Finally, the summary of semantic infringements by ETQE, which shows the percentage of infringements of this semantic business rule calculated as a percentage of the number of qualification enrolment records submitted to the NLRD by ETQE, shows clear trends of a systemic nature at some ETQEs.

## ***L.2 Unit Standard enrolments***

### ***L.2.1 No Accreditation***

This category indicates that the provider that is linked to the unit standard has never had an active accreditation to offer the unit standard. This category contains 62.77% of all of the records that infringe on this semantic business rule and this category is of greatest concern to SAQA.

Of the 29 discrete ETQEs in the dataset, 27 ETQEs are linked to this category. Of these records, 45.25% were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 6910 are linked to this category. Of these 6910 unit standards, 10 unit standards contribute to 5.70% of records in this category. Most notably, although 537 of the 9124 unit standards only constitute 0.01% of the records; the records for these unit standards represent 100% of the unit standard enrolment records submitted to the NLRD.

Of the 6254 discrete providers in the dataset, 3727 providers are linked to this category. Of these 3727 providers, 10 providers contribute to 22.44% of the records. Most notably, 642 of the 3727 providers constitute 0.04% of the records; the records for these providers

represent 100% of the unit standard enrolment records submitted to the NLRD for the providers.

As already noted, this category indicates that the provider has never had an active accreditation to offer the unit standard. As a result, providers in this category cannot be reported on in any of the other categories that form part of this research. However, other categories that the providers from this category can exist in are the 'No Accreditation (UStd Linked to Lshp)', 'No Accreditation Predicted' and ' ' No Accreditation Predicted (UStd Linked to Lshp)' categories that have been excluded from this research.

The reader should therefore note that the above statement "*...642 of the 3727 providers constitute 0.04% of the records; the records for these providers represent 100% of the unit standard enrolment records submitted to the NLRD for the providers.*" must further be interpreted to mean that for the remaining 3085 providers, 100% of the records for the specific provider fall into the categories 'No Accreditation', 'No Accreditation (UStd Linked to Lshp)', 'No Accreditation Predicted' or ' ' No Accreditation Predicted (UStd Linked to Lshp)'.

Unlike any of the other categories for this semantic business rule, it is unlikely that unit standard enrolment records that appear in the 'No Accreditation' category do so as a result of data capturing issues related to the enrolment record. Rather the data capturing or data quality issues reside in the lack of an accreditation of a provider to offer a unit standard.

Table L.2.1.1 provides an overview of the records found in this category grouped by submitting ETQE. The table differentiates the number of providers that have the submitting ETQE as their primary ETQE ("Primary ETQE of provider") and the number of providers where the submitting ETQE is not the primary ETQE of the provider ("Not Primary ETQE of provider"). A submitting ETQE may utilize another ETQE's providers (Section 3.8.3.5). It is found the same providers that are not accredited have been utilized by more than one ETQE and as a result the count of provider by submitting ETQE is 4034, whereas there are only 3727 discrete providers in this category.

Table L.2.1.1 ‘No Accreditation’ records by submitting ETQE identifier, count of Not Primary ETQE providers, count of Primary ETQE of provider and % unit standard enrolment records in the category

Submitting ETQE Identifier	Not Primary ETQE of Provider		Primary ETQE of Provider		Total	
	Count of Provider	%Records	Count of Provider	% Records	Count of Provider	%Records
1075	13	0.29%	11	0.43%	24	0.72%
1100	2	0.02%	4	0.49%	6	0.51%
1102	13	0.11%	143	2.28%	156	2.39%
1103	82	0.87%	296	2.17%	378	3.04%
1104	7	0.14%	5	0.10%	12	0.24%
1105	18	0.10%	783	13.88%	801	13.98%
1106	21	0.76%	58	1.46%	79	2.22%
1107	21	0.15%	64	1.09%	85	1.24%
1108	1	0.00%	2	0.00%	3	0.00%
1109	24	0.37%	118	2.48%	142	2.86%
1110	18	0.23%	30	0.33%	48	0.55%
1111	8	0.05%	201	13.72%	209	13.77%
1112	31	0.18%	172	2.16%	203	2.34%
1113	15	0.03%	127	0.33%	142	0.36%
1114	101	2.58%	106	7.44%	207	10.02%
1115	2	0.00%	84	1.19%	86	1.19%
1116	29	0.19%	116	2.26%	145	2.46%
1117	13	0.05%	212	3.74%	225	3.78%
1118	10	0.17%	27	0.34%	37	0.51%
1119	11	0.25%	100	0.09%	111	0.35%
1120	11	0.31%	16	1.81%	27	2.12%
1121	0	0.00%	3	0.02%	3	0.02%
1122	13	0.31%	130	1.21%	143	1.52%
1123	17	0.58%	203	7.60%	220	8.18%
1125	6	0.02%	123	1.12%	129	1.14%
1126	26	1.08%	276	8.52%	302	9.61%
1127	29	4.30%	82	10.59%	111	14.89%
<b>Grand Total</b>	<b>542</b>	<b>13.15%</b>	<b>3492</b>	<b>86.85%</b>	<b>4034</b>	<b>100.00%</b>

The reader should note that in the case of the overall implementation of a provider accreditation (see Section 4.4.3) it is understandable that an ETQE that is not the primary ETQE of the provider may not be aware that a specific provider has not been awarded an active accreditation by the provider’s primary ETQE. The implementation of accreditations to offer a unit standard are however different in that it is at the discretion of the accrediting ETQE to accredit a provider to offer its unit standards.

Analysis of Table L.2.1.1 shows the following notable trends:

- ETQE identifier 1105 has the highest incidence of the number of providers (801) that are not accredited to offer the unit standard. Further analysis shows that 75% of the providers linked to unit standard enrolment records for this ETQE have never been accredited to offer the unit standard.

- ETQE identifier 1103 has the second highest incidence of the number of providers (378) that are not accredited to offer the unit standard. Further analysis shows that 59.53% of the providers linked to unit standard enrolment records for this ETQE have never been accredited to offer the unit standard.
- ETQE identifier 1126 has the third highest incidence of the number of providers (302) that are not accredited to offer the unit standard. Further analysis shows that 51.10% of the providers linked to unit standard enrolment records for this ETQE have never been accredited to offer the unit standard.

### ***L.2.2 Start Before, End Before***

This category indicates that the unit standard enrolment started before the provider was accredited to offer the unit standard and either was achieved or expired before the provider was accredited to offer the unit standard. This category contains 16.28% of all of the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 28 ETQEs are linked to this category. Of these records, 44.17% were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 4935 are linked to this category. Of these 4935 unit standards, 10 unit standards contribute to 11.50% of records in this category. Most notably, although 26 of the 4935 unit standards constitute less than 0.01% of the records; the records for these unit standards represent 100% of the unit standard enrolment records submitted to the NLRD.

Of the 6254 discrete providers in the dataset, 1584 providers are linked to this category. Of these 1584 providers, 10 providers contribute to 24.93% of the records. Most notably, although 48 of the 1584 providers constitute less than 0.01% of the records; the records for these providers represent 100% of the unit standard enrolment records submitted to the NLRD for the providers.

The high densities of records that infringe on this rule, in combination with the high percentage of records in this category for some providers, hints at possible systemic issues in regard to provider accreditations to offer unit standards. The patterns in the data are

however not clearly delineated and a further review of unit standard enrolment that fall into this category needs to be conducted (refer to Appendix L.2.7).

### ***L.2.3 Start Before, End During***

This category indicates that the unit standard enrolment both started before and either was achieved or expired whilst the provider was accredited to offer the unit standard. This category contains 10.51% of all of the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 25 ETQEs are linked to this category. Of these records, 61.94% were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 4646 are linked to this category. Of these 4646 unit standards, 10 unit standards contribute to 20.16% of records in this category. Most notably, although 9 of the 4646 unit standards constitute less than 0.01% of the records; the records for these unit standards represent 100% of the unit standard enrolment records submitted to the NLRD.

Of the 6254 discrete providers in the dataset, 1456 providers are linked to this category. Of these 1456 providers, 10 providers contribute to 33.95% of the records. Most notably, although 17 of the 1456 providers constitute less than 0.01% of the records; the records for these providers represent 100% of the unit standard enrolment records submitted to the NLRD for the providers.

The high densities of records that infringe on this rule, in combination with the high percentage of records in this category for some providers, hints at possible systemic issues in regard to provider accreditations to offer unit standards. The patterns in the data are however not clearly delineated and a further review of unit standard enrolment records that fall into this category was conducted (refer to Appendix L.2.7).

Cross checking the results in this category with the results in the ‘Start Before, End Before’ (Appendix L.2.2) category shows some remarkable similarities. The first of which was that two ETQEs identified as top contributors to the ‘Start Before, End Before’ category are the top contributors to this category. A precursory review also revealed that 40% of the top 10 providers that contributed to the ‘Start Before, End Before’ category are top 10 contributors

in this category. These similarities suggested that a further review conducted on these records should be conducted in conjunction with the records in the 'Start Before, End Before' category.

#### ***L.2.4 Start After, End After***

This category indicates that the unit standard enrolment started after the provider was accredited to offer the unit standard and either was achieved or expired after the provider was no longer accredited to offer the unit standard. This category contains 6.18% of all of the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 25 ETQEs are linked to this category. Of these records, 49.69% were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 3327 are linked to this category. Of these 3327 unit standards, 10 unit standards contribute to 8.06% of records in this category. Most notably, although 44 of the 3327 unit standards constitutes less than 0.01% of the records; the records for these unit standards represent 100% of the unit standard enrolment records submitted to the NLRD for the unit standard.

Of the 6254 discrete providers in the dataset, 1087 providers are linked to this category. Of these 1087 providers, 10 providers contribute to 37.56% of the records. Most notably, although 45 of the 1087 providers only contribute 0.01% of the records; the records for these providers represent 100% of the unit standard enrolment records submitted to the NLRD for the providers.

The high densities of records that infringe on this rule in relation to the number of unit standard enrolment records submitted to the NLRD for the provider hints at possible systemic issues in regard to provider accreditations to offer the unit standard. The patterns in the data are however not clearly delineated and a further review of unit standard enrolment records submitted to the NLRD for these providers needed to be conducted (refer to Appendix L.2.8).

#### ***L.2.5 Start During, End After***

This category indicates that the unit standard enrolment both started during and either was achieved or expired after the provider was accredited to offer the unit standard. This category contains 3.98% of all of the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 23 ETQEs are linked to this category. Of these records, 60.80% were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 2940 are linked to this category. Of these 2940 unit standards, 10 unit standards contribute to 16.36% of records in this category. Most notably, although 3 of the 2940 unit standards constitute less than 0.01% of the records; the records for these unit standards represent 100% of the unit standard enrolment records submitted to the NLRD.

Of the 6254 discrete providers in the dataset, 765 providers are linked to this category. Of these 765 providers, 10 providers contribute to 40.95% of the records. Most notably, although 6 of the 765 providers contribute less than 0.01% of the records; the records for these providers represent 100% of the unit standard enrolment records submitted to the NLRD for the providers.

The high densities of records that infringe on this rule in relation to the number of unit standard enrolment records submitted to the NLRD for the provider hints at possible systemic issues in regard to provider accreditations to offer the unit standard. The patterns in the data are however not clearly delineated and a further review of unit standard enrolment records submitted to the NLRD for these providers needed to be conducted (refer to Appendix L.2.8).

Cross checking the results in this category with the results in the ‘Start After, End After’ (Appendix L.2.4) category shows some similarities. The first of which is that two of the ETQEs identified as top contributors to the ‘Start After, End After’ category is a top contributor to this category. These similarities suggested that a further review conducted on these records should be conducted in conjunction with the records in the ‘Start During, End After’ category.

### ***L.2.6 Start Before, End After***

This category indicates that the unit standard enrolment both started before the provider was accredited to offer the unit standard and either was achieved or expired after the provider was no longer accredited to offer the unit standard. This category contains 0.28% of all of the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 17 ETQEs are linked to this category. Of these records, 64.93% were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 671 are linked to this category. Of these 671 unit standards, 10 unit standards contribute to 26.71% of records in this category.

Of the 6254 discrete providers in the dataset, 120 providers are linked to this category. Of these 120 providers, 10 providers contribute to 79.89% of the records. Most notably, although 1 of the 120 providers contribute less than 0.01% of the records; the records for this provider represent 100% of the unit standard enrolment records submitted to the NLRD for the provider.

The low incidence of records that fall into this category suggests that these records are found in this category as a result of incorrect data capturing at the source of the data.

### ***L.2.7 Start Before, End Before or End During***

As stated in Appendix L.2.2 and L.2.3, the high density of records submitted to the NLRD for the providers in the categories ‘Start Before, End Before’ and ‘Start Before, End During’, in conjunction with intersections in two of the top 3 ranked ETQEs and top 10 ranked providers in these categories, suggested that further analysis of these records as a single data set should be conducted.

Further analysis found that these two initial categories shared 3999 unit standards and 1103 providers. As a result, the categories ‘Start Before, End Before’ and ‘Start Before, End During’ were grouped into a category called ‘Start Before, End Before or End During’ for this analysis. This category indicates that the unit standard enrolment started before the provider was accredited to offer the unit standard and either was achieved or expired before or whilst the provider was accredited to offer the unit standard. As a result of this

consolidation the 'Start Before, End Before or End During' category contains 26.80% of all the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 28 ETQEs are linked to this category. Of these records, 50.83% were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 5560 are linked to this category. Of these 5560 unit standards, 10 unit standards contribute to 13.86% of records in this category. Most notably, although 42 of the 5560 unit standards constitute less than 0.01% of the records; the records for this unit standard represent 100% of the unit standard enrolment records submitted to the NLRD.

Of the 6254 discrete providers in the dataset, 1936 providers are linked to this category. Of these 1936 providers, 10 providers contribute to 25.02% of the records. Most notably, although 84 of the 1936 providers constitute less than 0.01% of the records; the records for these providers represent 100% of the unit standard enrolment records submitted to the NLRD for the providers.

The volume of records found in this consolidated category constituted 8% of the total unit standard enrolment records that form part of the research. The implementation of data mining techniques, beyond exploratory data mining techniques, is needed to provide SAQA with both a succinct description of these records and to attempt to identify any data records that exist in this category as a result of possible incorrect data capturing at the source of the data.

The clustering data mining technique, as described in Appendix I.3, is applied to this data set in an effort to achieve both these aims.

An initial review of the data in this category, in combination with the understanding that the provider linked to a unit standard enrolment record may have a primary ETQE that differs from the ETQE that submitted the unit standard enrolment record to the NLRD, prompted the implementation of a new data field on the data set prior to data mining. The indicator PRIMARY\_ETQE\_DESC was developed as a nominal data value that contains the value 'Primary ETQE of provider' if the ETQE identifier of the ETQE that submitted the unit

standard enrolment record to the NLRD (ETQE\_ID) was the same as the ETQE identifier of the primary ETQE of the provider (PROV\_ETQE\_ID). The same indicator would have the value 'Not Primary ETQE of provider' if the primary ETQE identifier of the provider differed from the ETQE identifier of the ETQE that submitted the unit standard record to the NLRD. It was hoped that the implementation of the PRIMARY\_ETQE\_DESC indicator would allow the data mining algorithm to find discrete patterns in the data related to the utilization of providers that are accredited by an ETQE other than the ETQE that submitted the unit standard enrolment record to the NLRD.

The data mining effort results in the development of 8 data clusters (a technical description of each cluster is provided in Appendix M.2.1) that have the following most dominant characteristics:

1. Cluster 1

This cluster describes more than 25% of the records. The cluster is predominantly described as containing unit standard enrolment records for 128 unit standards as offered by 33 providers all of which have ETQE identifier 1111 as their primary ETQE. The records were submitted to the NLRD by ETQE identifier 1111 and predominantly have a NQF\_LEVEL\_DESC of 'Level 02'.

2. Cluster 2

This cluster describes more than 23% of the records. The cluster is relatively diverse in that it describes records submitted by 11 ETQEs, covering 1060 different unit standards offered by 176 different providers. These records predominantly have a PROV\_ACCRED\_IND\_DESC of 'Start Before, End Before' and UNIT\_STD\_TYPE\_DESC of 'Regular'.

3. Cluster 3

This cluster describes nearly 21% of the records. The cluster is diverse in that it describes records covering 176 different unit standards offered by 269 different providers. These records predominantly have a UNIT\_STD\_TYPE\_DESC of 'Regular-Fundamental'.

4. Cluster 4

This cluster describes only slightly more than 11% of the records. The cluster describes records covering 241 unit standards offered by 140 providers. All of the providers have

ETQE identifiers 1103 or 1105 as their primary ETQE and these records were all submitted to the NLRD by ETQE identifiers 1103 and 1105.

5. Cluster 5

This cluster describes nearly 6% of the records as belonging to 80 unit standards, all of which were predominantly have a ENROL\_TYPE\_DESC of 'Residential Learning (i.e. Contact Mode)'. The records were submitted to the NLRD by ETQE identifier 1106 and the same ETQE is the primary ETQE of the providers that offered these unit standards.

6. Cluster 6

The cluster describes slightly more than 5% of the records. All of the records in this cluster are described as being submitted by ETQEs that are not the primary ETQE of the provider that offered the unit standard.

7. Cluster 7

This cluster describes nearly 5% of the records as being submitted to the NLRD by ETQE identifiers 1031 and 1033. All of the records in this cluster are described as being submitted by ETQEs that are not the primary ETQE of the provider that offered the unit standard.

8. Cluster 8

The cluster describes more than 4% of the records as having been submitted to the NLRD by ETQE identifiers 1106, 1107, 1109, 1112, 1126. The provider accreditation for all of these records started before and ended before the enrolment on these unit standards commenced.

As stated before, this category contains records from 28 different ETQEs. The above description of the 8 clusters generated by the clustering algorithm shows that 4 of these clusters (clusters 1, 4, 7 and 8) each describe records that were submitted to the NLRD by specific ETQEs. Further, the clustering algorithm shows that two of the categories contain records for unit standards that were submitted to the NLRD by ETQEs that are not the primary ETQE of the providers that offered the unit standards.

The clustering algorithm does not generate any clusters with a population of less than 1%. A review of the probabilities allocated to each record in a cluster does however allow for the identification of records that are low enough to be considered anomalous as per the criteria defined in Appendix I.3. These records constitute 2.55% of the records found in this

category, and possibly exist in this category as a result of data capturing problems at the source of the data.

### ***L.2.8 Start During, Start After and End After***

As stated in Appendix L.2.4 and L.2.5, the high density of records submitted to the NLRD for the providers in the categories ‘Start During, End After’ and ‘Start After, End After’, in conjunction with intersections in the top 2 ranked ETQEs in these categories, suggested that further analysis of these records as a single data set should be conducted.

Further analysis found that these two initial categories shared 2371 unit standards and 517 providers. As a result, the categories ‘Start During, End After’ and ‘Start After, End After’ were grouped into a category called ‘Start During, Start After and End After’ for this analysis. This category indicates that the unit standard enrolment started during or after the provider was accredited and either was achieved or expired after the provider was no longer accredited. As a result of this consolidation the ‘Start During, Start After and End After’ category contains 10.16% of all the records that infringe on this semantic business rule.

Of the 29 discrete ETQEs in the dataset, 25 ETQEs are linked to this category. More than 52% of these records were submitted to the NLRD by 3 ETQEs.

Of the 9124 discrete unit standards in the dataset, 3879 are linked to this category. Of these 3879 unit standards, 10 unit standards contribute to 9.54% of records in this category. Most notably, although 50 of the 3879 unit standards only constitute less than 0.01% of the records; the records for these unit standards represent 100% of the unit standard enrolment records submitted to the NLRD.

Of the 6254 discrete providers in the dataset, 1330 providers are linked to this category. Of these 1330 providers, 10 providers contribute to 28.85% of the records. Most notably, although 52 of the 1330 providers only constitute less than 0.01% of the records; the records for these providers represent 100% of the unit standard enrolment records submitted to the NLRD for the providers.

The volume of records found in this consolidated category constitutes 3.03% of the total unit standard enrolment records that form part of the research. The implementation of data

mining techniques, beyond exploratory data mining techniques, is needed to provide SAQA with both a succinct description of these records and to attempt to identify any data records that exist in this category as a result of possible incorrect data capturing at the source of the data.

The clustering data mining technique, as described in Appendix I.3, is applied to this data set in an effort to achieve both these aims.

An initial review of the data in this category, in combination with the understanding that the provider linked to a unit standard enrolment record may have a primary ETQE that differs from the ETQE that submitted the unit standard enrolment record to the NLRD, prompted the implementation of a new data field on the data set prior to data mining. The indicator PRIMARY\_ETQE\_DESC was developed as a nominal data value that contains the value 'Primary ETQE of provider' if the ETQE identifier of the ETQE that submitted the unit standard enrolment record to the NLRD (ETQE\_ID) was the same as the ETQE identifier of the primary ETQE of the provider (PROV\_ETQE\_ID). The same indicator would have the value 'Not Primary ETQE of provider' if the primary ETQE identifier of the provider differed from the ETQE identifier of the ETQE that submitted the unit standard record to the NLRD. It was hoped that the implementation of the PRIMARY\_ETQE\_DESC indicator would allow the data mining algorithm to find discrete patterns in the data related to whether or not the ETQE submitted the data was primary ETQE of the provider that offered the unit standard.

The data mining effort results in the development of 8 data clusters (a technical description of each cluster is provided in Appendix M.2.2) that have the following most dominant characteristics:

- Cluster 1  
This cluster describes nearly 28% of the records as belonging to 75 unit standards offered by 48 providers. These records were submitted to the NLRD by ETQE identifier 1105.
- Cluster 2  
The cluster is relatively diverse and describes nearly than 20% of the records as having been submitted to the NLRD by 10 different ETQEs (ETQE identifiers 1075, 1102, 1103, 1109, 1110, 1111, 1114, 1115, 1125 and 1127). The cluster encompasses

751 unit standards offered by 98 providers. The majority of these records have a PROV\_ACCRED\_IND\_DESC of 'Start After, End After'.

- Cluster 3

This cluster describes nearly 13% of the records as belonging to 172 unit standards offered by 18 providers. These records were submitted to the NLRD by ETQE identifier 1126. The majority of these records have a PROV\_ACCRED\_IND\_DESC of 'Start After, End After'.

- Cluster 4

This cluster describes more than 9% of the records as belonging to 121 unit standards offered by 18 providers. These records were submitted to the NLRD by ETQE identifier 1112.

- Cluster 5

The cluster describes more than 8% of the records as belonging to 26 unit standards that have a SUB\_FIELD\_DESC of 'Language'. The unit standards are offered by 99 providers.

- Cluster 6

The cluster is relatively diverse and describes slightly more than 8% of the records as having been submitted to the NLRD by 6 different ETQEs (ETQE identifiers 1075, 1103, 1106, 1107, 1111 and 1126). The cluster encompasses 337 unit standards offered by 92 providers.

- Cluster 7

This cluster describes nearly 8% of the records as belonging to 559 unit standard offered by 68 providers. These enrolment records were submitted to the NLRD by 9 different ETQEs (ETQE identifiers 1075, 1100, 1103, 1109, 1110, 1111, 1122, 1125 and 1127).

- Cluster 8

The cluster describes nearly 6.5% of the records as belonging to 126 providers that offered 47 unit standards that all have a FIELD\_DESC of 'Physical, Mathematical, Computer and Life Sciences'. These enrolment records were submitted to the NLRD by 13 different ETQEs (ETQE identifiers 1075, 1103, 1105, 1106, 1107, 1109, 1110, 1112, 1114, 1117, 1123, 1125 and 1127).

Of the 8 clusters generated 5 provide a very discrete description of the characteristics of the records found in the cluster. The most notable clusters that are generated for this category are clusters 1, 3, 4, 5 and 8. Each of these clusters points to problems related either to specific unit standards, providers and/or ETQEs. Cluster 1, 3 and 4 suggest systemic issues related to specific ETQEs. Cluster 5 suggests that a specific grouping of unit standards that have a subfield description of 'Language' may have systemic problems, whereas Cluster 8 suggests that a specific grouping of unit standards that have a field description of 'Physical, Mathematical, Computer and Life Sciences' may have systemic problems. None of the clusters seemed to indicate a trend in regard to the utilization of providers whose primary ETQE is other than that of the submitting ETQE.

The clustering algorithm does not generate any clusters with a population of less than 1%. A review of the probabilities allocated to each record in a cluster does however allow for the identification of records that are low enough to be considered anomalous as per the criteria defined in Appendix I.3. These records constitute 5.94% of the records found in this category, and possibly exist in this category as a result of data capturing problems at the source of the data.

### ***L.2.9 Summary of semantic infringements by ETQE***

The preceding sections provide the results of records that infringe on this semantic business rule from the granular perspective of the unit standard enrolment record in relation to the complete dataset. This approach supports the determination of patterns within the data that point to systemic and anomalous problems within the overall dataset, which in turn lends itself to assessing the quality of the data in the data set.

The approach however ignores the diverse nature of ETQEs, and in particular the volume of the records that each ETQE submits to the NLRD. The final step in the analysis of this semantic business rule provides an overview of the percentage of records, calculated as a percentage of the number of records submitted by the ETQE, which infringe on this semantic business rule.

The results are presented as the percentage of records submitted by the ETQE that fall into a category that describes a semantic business rule issue (see Table L.2.9.1):

Table L.2.9.1 % of records submitted by an ETQE that have a category that describes a semantic business rule issue

ETQE Identifier	% Semantic Rule Issue
1116	89.23%
1115	75.36%
1117	74.07%
1100	69.42%
1075	69.05%
1112	66.34%
1114	60.34%
1122	56.22%
1120	40.65%
1125	38.70%
1109	36.92%
1105	28.28%
1126	25.47%
1110	25.18%
1111	20.99%
1103	19.50%
1106	19.34%
1107	17.06%
1102	14.15%
1119	13.61%
1123	12.38%
1127	12.12%
1118	11.63%
1104	10.26%
1113	8.93%
1121	7.95%
1108	4.22%
1124	0.43%

The results clearly illustrate that the infringement of this semantic business rule could be considered systemic at a number of the ETQEs.

### ***L.2.10 Conclusion***

The analysis of unit standard enrolment records in regard to whether the provider was accredited to offer the unit standard for the duration of the learner’s active enrolment highlights the possibility of systemic issues in regard to provider accreditations.

The cluster analysis for the ‘Start Before, End Before or End During’ and ‘Start During, Start After and End After’ categories is able to provide a clear description of the data in the categories. Further, a comparison across the two cluster analyses shows that ETQE

identifier 1105 is featured in both categories. The analysis of the 'No Accreditation' category highlights possible systemic issues in regard to provider accreditations as implemented by ETQE identifiers 1105, 1126 and 1103.

The cluster analysis of both the 'Start Before, End Before or End During' and 'Start During, Start After and End After' categories is able to identify records that may exist in these categories as a result of incorrect data capturing on the unit standard enrolment record. The analysis of the 'Start Before, End After' category in turn allows for the identification of enrolment records that have possibly been captured incorrectly.

Finally, the summary of semantic infringements by ETQE, which shows the percentage of infringements of this semantic business rule calculated as a percentage of the number of unit standard enrolment records submitted to the NLRD by ETQE, shows clear trends of a systemic nature at some ETQEs.

## Appendix M

This appendix provides a technical description of the outputs of data mining activities that were conducted when analysing whether the provider was accredited to offer the qualification for the duration of the learner's active enrolment on the qualification or unit standard. The data mining activities focuses on gaining a better understanding of data records that fall into specific categories of the data field PROV\_ACCRED\_IND (see Appendix E.3.7 and Appendix G.3.7) and the possible identification of anomalous data records in the respective data sets.

This semantic business rule defines that a provider must be accredited to offer the qualification for the duration of the learner's active enrolment and is applicable to qualification and unit standard enrolments. As a result the structure of this appendix has sub sections that focus on the specific data mining activities per specific categories for each of these types of enrolment records.

### *M.1 Qualification enrolment*

#### *M.1.1 Start Before, End Before or End During cluster data mining*

This section provides a technical description of the clusters that were generated by cluster data mining the consolidated data category 'Start Before, End Before or End During' (see Appendix L.1.7) for qualification enrolment records. A detailed description of the manner in which cluster data mining was conducted can be found in Appendix I.3. The results of the generated clustering model were significant because the model was measured as being 97.61% accurate.

The generated clusters show a tight coupling between data fields that describe the ETQEs, qualifications and providers. This is as a result of the organic relationship between qualifications and ETQEs (qualifications are generally implemented by one ETQE only) and providers and ETQEs (providers generally offer qualifications that are implemented by their primary ETQE).

The model did not generate any clusters that contained less than 1% of the records. The detection of anomalous records based on cluster specific probabilities of less than .6 being allocated to the record was conducted. In this manner it was determined that 1.23% of the

records possibly exist in this category as a result of data capturing problems (see Appendix I.3).

The technical description of each of the 8 clusters is provided below. The description is limited to the top 10 attributes of the cluster rule, where each indent of text represents the importance of the attribute. For example:

A line formatted like this represents an importance of 100%

*A line formatted like this represents an importance greater than 75% and less than 100%*

*A line formatted like this represents an importance greater than 50% and less than or equal to 75%*

*A line formatted like this represents an importance less than or equal to 50%*

#### 1. Cluster 1

% of records: 25.38%

Average probability: 0.9913

Rule:

*ASSESSOR\_ID IN ('NULL')*

*AND PROVIDER\_ID IN ('715', '49153', '44779', '42963', '42946', '39897', '37405', '37392', '36737', '36381', '35735', '35671', '31677', '30217', '16993', '1575', '15241', '14920', '14658', '13248', '12894', '12666')*

*AND LEARNERSHIP\_ID IN ('NULL', '894', '892', '888')*

*AND QUALICATION\_ID IN ('71506', '61772', '59293', '59114', '58080', '57954', '50097', '48904', '36250', '35945', '23850', '23673', '23672', '23671', '21810', '20924', '20190')*

*AND ETQE\_ID IN ('1126')*

*AND PROV\_ETQE\_ID IN ('1126', '1031')*

*AND SUBFIELD\_DESC IN ('Personal Care', 'Office Administration', 'Marketing', 'Generic Management', 'Finance, Economics and Accounting', 'Cleaning, Domestic, Hiring, Property and Rescue Services')*

*AND PROV\_PROVINCE\_DESC IN ('Gauteng', 'Eastern Cape')*

*AND FIELD\_DESC IN ('Services', 'Business, Commerce and Management Studies')*

*AND PROVIDER\_TYPE\_DESC IN ('Training', 'Education and Training')*

## 2. Cluster 2

% of records: 15.91%

Average probability: 0.9498

Rule:

*ASSESSOR\_ID IN ('NULL')*

*AND LEARNERSHIP\_ID IN ('NULL', '999', '816', '744', '483', '1139', '1085', '1077', '1072', '1071', '1070', '1069', '1000')*

*AND QUALICATION\_ID IN ('78982', '78981', '65426', '64827', '64826', '63426', '63351', '62826', '62086', '61467', '59343', '58799', '58411', '58223', '57897', '57821', '50302', '49706', '49705', '49623', '49414', '49297', '49148', '49108', '49094', '49070', '48989', '48889', '48781', '24010', '23910', '23270')*

*AND PROVIDER\_ID IN ('715', '5205', '51225', '51199', '51195', '51145', '50489', '50240', '49951', '49947', '49943', '49940', '49726', '49724', '46460', '45504', '45299', '44202', '42946', '41592', '41565', '40194', '40149', '39104', '37727', '35427', '35381', '34674', '34337', '33644', '32184', '31909', '31751', '31750', '29815', '29730', '29600', '29436', '29370', '27090', '27002', '26996', '25367', '25098', '23498', '22294', '22278', '21966', '21961', '21942', '21830', '2151', '1945', '12894', '11253', '11246', '11244', '11242')*

*AND SUBFIELD\_DESC IN ('Wholesale and Retail', 'Transport, Operations and Logistics', 'Promotive Health and Developmental Services', 'Manufacturing and Assembly', 'Information Technology and Computer Sciences', 'Finance, Economics and Accounting')*

*AND ENROL\_TYPE\_DESC IN ('Mixed Mode')*

*AND ETQE\_ID IN ('1127', '1123', '1117', '1114', '1113', '1109', '1103', '1102')*

*AND PROV\_ETQE\_ID IN ('1126', '1123', '1117', '1115', '1114', '1109', '1103', '1102', '1031')*

*AND PROVIDER\_TYPE\_DESC IN ('Education and Training')*

*AND FIELD\_DESC IN ('Services', 'Physical, Mathematical, Computer and Life Sciences', 'Manufacturing, Engineering and Technology', 'Health Sciences and Social Services', 'Business, Commerce and Management Studies')*

## 3. Cluster 3

% of records: 14.20%

Average probability: 0.9439

Rule:

```
ASSESSOR_ID IN ('NULL')
AND LEARNERSHIP_ID IN ('NULL', '801', '794', '523', '377', '337', '321', '320', '305',
'240', '1498', '1476', '1465', '1463', '1462', '1461', '1460', '1459', '1387', '1376')
AND PROVIDER_ID IN ('6970', '672', '662', '592', '51153', '50457', '50456', '48475',
'47584', '46460', '46459', '44224', '44143', '44091', '44005', '43436', '42546', '41566',
'41493', '41440', '39674', '39665', '39043', '38660', '38625', '38604', '38596', '38580',
'38579', '38575', '38574', '38564', '38563', '38561', '38560', '38555', '38554', '38552',
'38551', '38542', '37623', '36888', '36875', '36860', '36858', '36831', '36684', '36480',
'35945', '35942', '35940', '35926', '35925', '35920', '34946', '33741', '32807', '32233',
'31725', '30028', '29980', '29713', '29313', '29273', '29240', '28872', '27531', '26783',
'2657', '26362', '2626', '25366', '23274', '22993', '2247', '2184', '21665', '20642',
'20574', '1916', '18755', '18753', '1578', '15241', '14640', '11691', '11091', '11064')
AND QUALICATION_ID IN ('78981', '71967', '65066', '64827', '64726', '61686',
'61608', '60312', '60311', '60310', '60206', '60186', '59868', '59033', '58968', '58777',
'58756', '58556', '58514', '58284', '58161', '58160', '57820', '57711', '57467', '50389',
'50222', '50040', '49709', '49666', '49597', '49148', '49144', '49094', '49062', '49031',
'49030', '49026', '48989', '48982', '48889', '48800', '48778', '48743', '48590', '48492',
'48490', '24310', '24290', '23695', '23650', '23647', '23644', '23642', '23641', '23492',
'22690', '22507', '21887', '21861', '21832', '21829', '21828', '20307', '20211')
AND SUBFIELD_DESC IN ('Electrical Infrastructure Construction', 'Civil
Engineering Construction', 'Building Construction', 'Public Administration', 'Primary
Agriculture', 'Manufacturing and Assembly', 'Information Technology and Computer
Sciences', 'Finance, Economics and Accounting', 'Fabrication and Extraction',
'Engineering and Related Design')
AND FIELD_DESC IN ('Physical, Mathematical, Computer and Life
Sciences', 'Physical Planning and Construction', 'Manufacturing, Engineering
and Technology', 'Business, Commerce and Management Studies')
AND ETQE_ID IN ('1127', '1125', '1123', '1112', '1111', '1110', '1109', '1107',
'1104', '1103')
AND -9.7 <= END_PROV_ACCRED_IND <= 0
AND QUALICATION_CLASS_DESC IN ('Regular-Unit Stds Based',
'Regular-ELOAC')
AND ENROL_TYPE_DESC IN ('Unknown', 'Mixed Mode')
```

4. Cluster 4

% of records: 13.40%

Average probability: 1.0000

Rule:

SUBFIELD\_DESC IN ('Safety in Society')

*AND LEARNERSHIP\_ID IN ('NULL')*

*AND QUALICATION\_ID IN ('78160', '58594', '50139', '22507')*

*AND ASSESSOR\_ID IN ('NULL')*

*AND PROVIDER\_ID IN ('47651', '47510', '47468', '47358', '47283', '47213', '47082', '46946', '46944', '46935', '46763', '46605', '35542', '32878', '28937', '28922', '23220', '23172', '23163', '21121', '21085', '21017', '21006', '2066', '48295', '48150')*

*AND PROV\_ETQE\_ID IN ('1105')*

*AND ETQE\_ID IN ('1105')*

*AND PROV\_PROVINCE\_DESC IN ('South Africa National')*

*AND ENROL\_TYPE\_DESC IN ('Mixed Mode')*

*AND FIELD\_DESC IN ('Law, Military Science and Security')*

5. Cluster 5

% of records: 12.83%

Average probability: 0.9894

Rule:

SUBFIELD\_DESC IN ('Finance, Economics and Accounting')

*AND PROVIDER\_ID IN ('50009', '37975', '35342')*

*AND LEARNERSHIP\_ID IN ('NULL')*

*AND QUALICATION\_ID IN ('58393', '58392', '36230', '20408', '20375')*

*AND ASSESSOR\_ID IN ('NULL')*

*AND PROV\_ETQE\_ID IN ('1116')*

*AND ETQE\_ID IN ('1116')*

*AND PROV\_PROVINCE\_DESC IN ('Gauteng')*

*AND ENROL\_TYPE\_DESC IN ('Mixed Mode')*

*AND FIELD\_DESC IN ('Business, Commerce and Management Studies')*

6. Cluster 6

% of records: 8.44%

Average probability: 0.9391

Rule:

*ASSESSOR\_ID IN ('NULL')*

*AND SUBFIELD\_DESC IN ('Safety in Society', 'Public Administration', 'Manufacturing and Assembly', 'Finance, Economics and Accounting', 'Cleaning, Domestic, Hiring, Property and Rescue Services')*

*AND LEARNERSHIP\_ID IN ('NULL', '744', '74', '53', '240', '24', '212')*

*AND QUALICATION\_ID IN ('73729', '62826', '60170', '58393', '58392', '49666', '49094', '48982', '48800', '48550', '35970', '24471', '24435', '24290', '23270', '22994', '22507', '21887', '20408', '20203', '20202')*

*AND PROVIDER\_ID IN ('715', '50009', '48775', '48625', '48591', '47584', '47498', '46806', '46460', '44688', '43436', '40627', '39208', '37975', '36921', '36875', '35342', '35341', '29600', '28305', '27078', '26950', '1916', '1907', '1758', '17020', '16932', '1635', '15241', '12247', '11100')*

*AND ENROL\_TYPE\_DESC IN ('Mixed Mode')*

*AND PROV\_ETQE\_ID IN ('NULL', '1120', '1116', '1031')*

*AND PROVIDER\_CLASS\_DESC IN ('Private', 'NULL')*

*AND ETQE\_ID IN ('1127', '1126', '1120', '1116', '1113', '1110', '1109', '1105', '1103')*

*AND -58.2 <= END\_PROV\_ACCRED\_IND <= 0*

## 7. Cluster 7

% of records: 5.61%

Average probability: 0.9997

Rule:

*QUALICATION\_ID IN ('58778')*

*AND LEARNERSHIP\_ID IN ('NULL', '1374')*

*AND PROVIDER\_ID IN ('50578', '50114', '38989', '32693', '28340', '28049', '21337', '21328')*

*AND SUBFIELD\_DESC IN ('Early Childhood Development')*

*AND ETQE\_ID IN ('1106')*

*AND ASSESSOR\_ID IN ('NULL', '8145645', '8145537', '8145178', '7309474', '6056024', '6055425', '4631855', '4282642', '4282588', '3313073', '3312995', '3018856', '3018377', '3008406', '3005384')*  
*AND PROV\_ETQE\_ID IN ('1106', '1033')*  
*AND ENROL\_TYPE\_DESC IN ('Residential Learning (i.e. Contact Mode)')*  
*AND FIELD\_DESC IN ('Education, Training and Development')*  
*AND NQF\_LEVEL\_DESC IN ('Level 4')*

#### 8. Cluster 8

% of records: 4.24%

Average probability: 0.9903

Rule:

*LEARNERSHIP\_ID IN ('NULL')*  
*AND QUALICATION\_ID IN ('73271', '50351', '23135', '23134')*  
*AND PROVIDER\_ID IN ('50114', '49583', '42882', '34125', '32693', '29963', '28514', '28315', '28089', '21337', '21328', '21318')*  
*AND ETQE\_ID IN ('1106')*  
*AND ASSESSOR\_ID IN ('NULL', '9574855', '9464189', '3444826', '3313040', '3024103', '3018856', '3018150', '3011652')*  
*AND SUBFIELD\_DESC IN ('Early Childhood Development', 'Adult Learning')*  
*AND PROV\_ETQE\_ID IN ('1106', '1033')*  
*AND ENROL\_TYPE\_DESC IN ('Residential Learning (i.e. Contact Mode)')*  
*AND FIELD\_DESC IN ('Education, Training and Development')*  
*AND PROVIDER\_CLASS\_DESC IN ('Unknown', 'Public')*

#### ***M.1.2 Start During, Start After and End After cluster data mining***

This section provides a technical description of the clusters that were generated by cluster data mining the consolidated data category ‘Start During, Start After and End After’ (see Appendix L.1.8) for qualification enrolment records. A detailed description of the manner in which cluster data mining was conducted can be found in Appendix I.3. The results of the generated clustering model were significant because the model was measured as being 97.58% accurate.

The generated clusters show a tight coupling between data fields that describe the ETQEs, qualifications and providers. This is as a result of the organic relationship between qualifications and ETQEs (qualifications are generally implemented by one ETQE only) and providers and ETQEs (providers generally offer qualifications that are implemented by their primary ETQE).

The model did not generate any clusters that contained less than 1% of the records. The detection of anomalous records based on cluster specific probabilities of less than .6 being allocated to the record was conducted. In this manner it was determined that 0.81% of the records possibly exist in this category as a result of data capturing problems (see Appendix I.3).

The technical description of each of the 8 clusters is provided below. The description is limited to the top 10 attributes of the cluster rule, where each indent of text represents the importance of the attribute. For example:

A line formatted like this represents an importance of 100%

*A line formatted like this represents an importance greater than 75% and less than 100%*

*A line formatted like this represents an importance greater than 50% and less than or equal to 75%*

*A line formatted like this represents an importance less than or equal to 50%*

#### 1. Cluster 1

% of records: 31.20%

Average probability: 0.9999

Rule:

QUALIFICATION\_ID IN ('48930')

*AND LEARNERSHIP\_ID IN ('NULL')*

*AND PROV\_ETQE\_ID IN ('1079')*

*AND SUBFIELD\_DESC IN ('Finance, Economics and Accounting')*

*AND ETQE\_ID IN ('1079')*

*AND ASSESSOR\_ID IN ('NULL')*

*AND PROVIDER\_ID IN ('47351', '39360', '39315', '39313', '39307', '39278', '39250', '29529', '29524', '26215', '26187', '26179', '26177', '26176', '26171', '26164', '26163', '26162', '26161', '26159', '26157', '26154', '26151', '26147', '26146', '26142', '26141', '26140', '26139', '26134', '26126', '26124', '26120', '26116', '26111', '26110', '26109',*

'26106', '26105', '26096', '26084', '26083', '26082', '26080', '26075', '26074', '26068',  
'26065', '26063', '26059', '26056', '26055', '26053', '26045', '26044', '26040', '26039',  
'26036', '26035', '26032', '26031', '26030', '26022', '26021', '26020', '26017', '26015',  
'26008', '26005', '26004', '25993', '25988', '25975', '25972', '25970', '25969', '25964',  
'25963', '25962', '25961', '25958', '25957', '25954', '25951', '25949', '25946', '25944',  
'25943', '25941', '25940', '25939', '25938', '25931', '25929', '25923', '25922', '25919',  
'25913', '25911', '25910', '25901', '25900', '25892', '25891', '25890', '25885', '25884',  
'25883', '25882', '25870', '25869', '25868', '25867', '25864', '25862', '25861', '25847',  
'25845', '25833', '25832', '25831', '25828', '25826', '25825', '25824', '25823', '25822',  
'25821', '25820', '25815', '25808', '25801', '25799', '25792', '25789', '25785', '25781',  
'25780', '25779', '25778', '25777', '25773', '25771', '25765', '25764', '25763', '25761',  
'25760', '25754', '25748', '25747', '25741', '25740', '25739', '25738', '25737', '25736',  
'25735', '25732', '25713', '25712', '25711', '25710', '25699', '25695', '25694', '25691',  
'25688', '25686', '25685', '25684', '25682', '25679', '25678', '25677', '25675', '25672',  
'25670', '25659', '25656', '25654', '25652', '25647', '25643', '25638', '25636', '25627',  
'25624', '25617', '25614', '25593', '25592', '25588', '25587', '25582', '25580', '25579',  
'25578', '25577', '25576', '25573', '25572', '25571', '25570', '25569', '25566', '25561',  
'25560', '25558', '25553', '25548', '25546', '25544', '25542', '25534', '25533', '25528',  
'25525', '25517', '25515', '25507', '25496', '25495', '25484', '25482', '25472', '25470',  
'25468', '25466', '25462', '25458', '25443', '25442', '25440', '25439', '25437', '25434',  
'25432')

AND ENROL\_TYPE\_DESC IN ('Work Place Learning')

AND FIELD\_DESC IN ('Business, Commerce and Management Studies')

AND NQF\_LEVEL\_DESC IN ('Level 7')

## 2. Cluster 2

% of records: 18.89%

Average probability: 0.9906

Rule:

*PROVIDER\_ID* IN ('796', '48792', '41493', '32230', '32209', '28156', '25429', '25426',  
'23362', '1905', '18758', '1726', '1578', '14814')

AND *QUALICATION\_ID* IN ('57934', '57625', '49946', '49709', '49708', '49666',  
'48550')

AND *ASSESSOR\_ID* IN ('NULL')

*AND LEARNERSHIP\_ID IN ('NULL', '53', '528', '523')*  
*AND SUBFIELD\_DESC IN ('Human Resources', 'Finance, Economics and Accounting')*  
*AND ENROL\_TYPE\_DESC IN ('Mixed Mode')*  
*AND ETQE\_ID IN ('1127', '1120', '1075')*  
*AND FIELD\_DESC IN ('Business, Commerce and Management Studies')*  
*AND QUALICATION\_CLASS\_DESC IN ('Regular-Provider-Stds Base', 'Regular-ELOAC')*  
*AND PROV\_ETQE\_ID IN ('NULL', '1127', '1120', '1075', '1031')*

### 3. Cluster 3

% of records: 10.75%

Average probability: 0.9325

Rule:

*LEARNERSHIP\_ID IN ('NULL', '801', '196', '1176')*  
*AND PROVIDER\_ID IN ('796', '48559', '46926', '41011', '40960', '40925', '39612', '39205', '39001', '38989', '38346', '37747', '36682', '35961', '35920', '35551', '35516', '32919', '32151', '32049', '32014', '31764', '29804', '29781', '29273', '28531', '27114', '27104', '24960', '2349', '2159', '2132', '20861', '20860', '20772', '2071', '1974', '18580', '14396')*  
*AND ASSESSOR\_ID IN ('NULL')*  
*AND QUALICATION\_ID IN ('66791', '65426', '61566', '60207', '60006', '59768', '59406', '58968', '58594', '50389', '50139', '49623', '49614', '49148', '49106', '49070', '48992', '48989', '48937', '48492', '24290', '24214', '24010', '23171', '22688', '20830', '20369', '20168', '14128')*  
*AND ENROL\_TYPE\_DESC IN ('Mixed Mode')*  
*AND 0 <= START\_PROV\_ACCRED\_IND <= 9.6*  
*AND FIELD\_DESC IN ('Physical Planning and Construction', 'Law, Military Science and Security', 'Health Sciences and Social Services', 'Business, Commerce and Management Studies', 'Agriculture and Nature Conservation')*  
*AND PROV\_ETQE\_ID IN ('1125', '1122', '1117', '1116', '1109', '1105', '1104')*  
*AND QUALICATION\_CLASS\_DESC IN ('Regular-Unit Stds Based')*  
*AND SUBFIELD\_DESC IN ('Secondary Agriculture', 'Safety in Society', 'Promotive Health and Developmental Services', 'Primary Agriculture',*

'Physical Planning, Design and Management', 'Nature Conservation', 'Justice in Society', 'Hospitality, Tourism, Travel, Gaming and Leisure', 'Finance, Economics and Accounting', 'Engineering and Related Design', 'Civil Engineering Construction', 'Building Construction')

#### 4. Cluster 4

% of records: 10.50%

Average probability: 0.9469

Rule:

*PROVIDER\_ID IN ('42050', '41517', '41386', '41011', '40926', '40925', '40623', '39084', '38548', '36983', '36953', '36906', '36682', '36061', '35955', '35920', '35319', '33654', '32145', '32062', '32044', '32014', '31724', '30028', '29782', '29712', '29678', '29294', '29273', '29272', '29260', '29182', '27859', '27120', '27114', '27109', '26989', '26632', '26631', '26616', '25366', '25355', '25332', '23274', '22306', '1578', '12666', '11772', '11244', '11088')*

*AND ASSESSOR\_ID IN ('NULL')*

*AND LEARNERSHIP\_ID IN ('NULL', '945', '803', '801', '799', '744', '364', '240', '189', '1242', '1176', '1086')*

*AND QUALICATION\_ID IN ('65426', '65046', '61586', '61566', '60207', '60186', '59406', '59114', '58968', '58802', '58799', '58798', '58756', '57954', '57848', '50324', '49811', '49623', '49144', '49094', '49070', '49030', '48996', '48994', '48989', '48987', '48982', '48828', '48491', '48490', '36230', '24290', '23850', '23695', '23270', '22882', '21907', '20936', '20830', '20730', '14130', '14128', '14127')*

*AND ETQE\_ID IN ('1126', '1125', '1122', '1116', '1112', '1111', '1109', '1108', '1107', '1103')*

*AND PROV\_ETQE\_ID IN ('1126', '1125', '1122', '1116', '1114', '1112', '1111', '1109', '1108')*

*AND ENROL\_TYPE\_DESC IN ('Mixed Mode')*

*AND PROVIDER\_CLASS\_DESC IN ('Private')*

*AND SUBFIELD\_DESC IN ('Visual Arts', 'Public Administration', 'Primary Agriculture', 'Physical Planning, Design and Management', 'Office Administration', 'Nature Conservation', 'Music', 'Manufacturing and Assembly', 'Hospitality, Tourism, Travel, Gaming and Leisure', 'Finance, Economics and*

Accounting', 'Fabrication and Extraction', 'Electrical Infrastructure Construction', 'Civil Engineering Construction', 'Building Construction')  
AND QUALICATION\_CLASS\_DESC IN ('Regular-Unit Stds Based')

5. Cluster 5

% of records: 9.07%

Average probability: 0.9112

Rule:

*PROVIDER\_ID IN ('50627', '5045', '48559', '48255', '46877', '46423', '44729', '43440', '43397', '43294', '43286', '41474', '38989', '38426', '37105', '34674', '34651', '33756', '33754', '33535', '30217', '30139', '29880', '29767', '29731', '29441', '29352', '29277', '28710', '28531', '28282', '2748', '26362', '25095', '23169', '22974', '22903', '21464', '20861', '20603', '20357', '19858', '18229', '13655', '11691', '11228', '11127')*  
*AND ASSESSOR\_ID IN ('NULL')*  
*AND QUALICATION\_ID IN ('73286', '65426', '64866', '64827', '64826', '58799', '58798', '58594', '58362', '58244', '58043', '57848', '57711', '50601', '50351', '50139', '50077', '49706', '49623', '49614', '49031', '49030', '48987', '48835', '48815', '35945', '24290', '24150', '23850', '23695', '23270', '22456', '21887', '21031', '20730', '20525')*  
*AND ETQE\_ID IN ('1126', '1117', '1113', '1111', '1109', '1107', '1105', '1103', '1102')*  
*AND LEARNERSHIP\_ID IN ('NULL', '799', '778', '744', '305', '240', '215', '1450', '1274', '1270', '1269', '1242', '1093')*  
*AND PROV\_ETQE\_ID IN ('NULL', '1126', '1117', '1115', '1111', '1109', '1105', '1103', '1102')*  
*AND PROVIDER\_TYPE\_DESC IN ('Unknown', 'Employer', 'Education and Training', 'Education')*  
*AND ENROL\_STATUS\_DESC IN ('Enrolled')*  
*AND ENROL\_TYPE\_DESC IN ('Unknown', 'Mixed Mode')*  
*AND FIELD\_DESC IN ('Physical Planning and Construction', 'Manufacturing, Engineering and Technology', 'Law, Military Science and Security', 'Health Sciences and Social Services', 'Business, Commerce and Management Studies')*

6. Cluster 6

% of records: 7.08%

Average probability: 0.9834

Rule:

QUALICATION\_TYPE\_DESC IN ('National Higher Certificate', 'National Diploma')  
AND PROVIDER\_ID IN ('29707', '28297', '28279', '28261', '28244', '28169', '28140',  
'28038', '26362', '1760', '1726', '13655')  
AND QUALICATION\_ID IN ('50351', '49708', '48894', '23135', '23134')  
AND LEARNERSHIP\_ID IN ('NULL')  
AND SUBFIELD\_DESC IN ('Schooling', 'Early Childhood Development', 'Adult  
Learning')  
AND ETQE\_ID IN ('1106')  
AND PROV\_ETQE\_ID IN ('1106', '1031')  
AND ASSESSOR\_ID IN ('NULL', '7339985', '5013557', '3313098', '3025400',  
'3008822')  
AND ENROL\_TYPE\_DESC IN ('Residential Learning (i.e. Contact Mode)')  
AND FIELD\_DESC IN ('Education, Training and Development')

#### 7. Cluster 7

% of records: 6.38%

Average probability: 0.9974

Rule:

QUALICATION\_ID IN ('58778')  
AND PROVIDER\_ID IN ('50125', '48868', '37251', '28446', '28438', '28060', '21389')  
AND LEARNERSHIP\_ID IN ('NULL')  
AND SUBFIELD\_DESC IN ('Early Childhood Development')  
AND ETQE\_ID IN ('1106')  
AND ASSESSOR\_ID IN ('NULL', '8145429', '8145122', '5518282', '5518273',  
'4283216', '4282973', '4282714', '3557318', '3027298', '3018346')  
AND PROV\_ETQE\_ID IN ('1106', '1033', '1031')  
AND PROV\_PROVINCE\_DESC IN ('Northern Cape', 'Limpopo', 'Gauteng', 'Free  
State')  
AND ENROL\_TYPE\_DESC IN ('Residential Learning (i.e. Contact Mode)')  
AND FIELD\_DESC IN ('Education, Training and Development')

#### 8. Cluster 8

% of records: 6.14%

Average probability: 0.9965

Rule:

```
PROVIDER_ID IN ('11092', '11091', '11087')
AND LEARNERSHIP_ID IN ('NULL')
AND QUALICATION_ID IN ('22882', '13719', '13717', '13716', '13715', '13714',
'13713', '13689', '13673', '13670', '13657', '13650')
AND SUBFIELD_DESC IN ('Engineering and Related Design')
AND ETQE_ID IN ('1115')
AND ASSESSOR_ID IN ('NULL', '3020861')
AND PROV_ETQE_ID IN ('NULL', '1115')
AND PROVIDER_CLASS_DESC IN ('Private', 'NULL')
    AND FIELD_DESC IN ('Manufacturing, Engineering and Technology')
    AND PROVIDER_TYPE_DESC IN ('NULL', 'Education')
```

## ***M.2 Unit Standard enrolment***

### ***M.2.1 Start Before, End Before or End During cluster data mining***

This section provides a technical description of the clusters that were generated by cluster data mining the consolidated data category ‘Start Before, End Before or End During’ (see Appendix L.2.7) for unit standard enrolment records. A detailed description of the manner in which cluster data mining was conducted can be found in Appendix I.3. The results of the generated clustering model were significant because the model was measured as being 98.28% accurate.

The generated clusters show a tight coupling between data fields that describe the ETQEs, unit standards and providers. This is as a result of the organic relationship between unit standards and ETQEs and providers and ETQEs (providers generally offer unit standards that are implemented by their primary ETQE).

The model did not generate any clusters that contained less than 1% of the records. The detection of anomalous records based on cluster specific probabilities of less than .6 being allocated to the record was conducted. In this manner it was determined that 2.55% of the

records possibly exist in this category as a result of data capturing problems (see Appendix I.3).

The technical description of each of the 8 clusters is provided below. The description is limited to the top 10 attributes of the cluster rule, where each indent of text represents the importance of the attribute. For example:

A line formatted like this represents an importance of 100%

*A line formatted like this represents an importance greater than 75% and less than 100%*

*A line formatted like this represents an importance greater than 50% and less than or equal to 75%*

*A line formatted like this represents an importance less than or equal to 50%*

#### 1. Cluster 1

% of records: 25.31%

Average probability: 0.9969

Rule:

ASSESSOR\_ID IN ('NULL')

*AND PROVIDER\_ID IN ('25117', '36831', '36833', '36844', '36857', '36858', '38542', '38551', '38554', '38555', '38558', '38560', '38561', '38563', '38574', '38575', '38584', '38590', '38601', '38603', '38604', '38607', '38608', '38610', '38640', '38659', '38660', '38662', '38666', '38667', '38676', '46459', '50456')*

*AND UNIT\_STANDARD\_ID IN ('10568', '10570', '10572', '10573', '110092', '110135', '110138', '110139', '110144', '110234', '110434', '115767', '116454', '116653', '116674', '116687', '119109', '119112', '119129', '119136', '119137', '119584', '12232', '12526', '14128', '15282', '15297', '15301', '15316', '243781', '243789', '243791', '243793', '243794', '243796', '243801', '244376', '244377', '244378', '244380', '244381', '244382', '244383', '244385', '244386', '244387', '244389', '244393', '244395', '244396', '244397', '244398', '244399', '244400', '244401', '244402', '244403', '244405', '244407', '244408', '244409', '244410', '244414', '244415', '244416', '244417', '244418', '244419', '244420', '244422', '244423', '244425', '244426', '244427', '244428', '244429', '244430', '244431', '244432', '244433', '244434', '244435', '244436', '244437', '244438', '244439', '244441', '244442', '244443', '244444', '244446', '244447', '244448', '244449', '244450', '244451', '244456', '244457', '244459', '244460', '244461', '244462', '244463', '244464', '244465', '244466', '244467', '244469', '244470', '244477',*

'244493', '252671', '253042', '253813', '254594', '7524', '7525', '7526', '9616', '9617',  
 '9618', '9619', '9695', '9706', '9707', '9708', '9710', '9717')  
 AND SUBFIELD\_DESC IN ('Engineering and Related Design', 'Fabrication and  
 Extraction')  
 AND PROV\_ETQE\_ID IN ('1111')  
 AND ETQE\_ID IN ('1111')  
 AND FIELD\_DESC IN ('Manufacturing, Engineering and Technology')  
 AND NQF\_LEVEL\_DESC IN ('Level 2')  
 AND PROVIDER\_CLASS\_DESC IN ('Private')  
 AND PROVIDER\_TYPE\_DESC IN ('Training')

- Cluster 2

% of records: 23.16%

Average probability: 0.9747

Rule:

ASSESSOR\_ID IN ('4386256', '4707504', '4783572', '7355866', 'NULL')  
 AND PROVIDER\_ID IN ('10923', '11100', '11236', '11242', '11244', '11246', '11250',  
 '11255', '11262', '11305', '11691', '12666', '12894', '14640', '14795', '15241', '1575',  
 '1578', '17121', '17127', '18511', '18584', '18698', '1914', '1916', '1945', '20584',  
 '2076', '20933', '2151', '2158', '21830', '2184', '21961', '21966', '22236', '22278',  
 '22294', '22299', '22832', '23274', '23498', '24870', '25033', '25366', '25367', '25384',  
 '2626', '26342', '26362', '26982', '26986', '27078', '27090', '27092', '27097', '27125',  
 '27128', '27135', '28872', '29188', '29212', '29251', '29269', '29273', '29275', '29277',  
 '29368', '29600', '29679', '29691', '29709', '29713', '29937', '29980', '31726', '31751',  
 '32169', '32230', '32231', '32233', '32807', '32915', '32933', '33665', '33741', '33754',  
 '33758', '34827', '35378', '35381', '35427', '35675', '35735', '35920', '35940', '36331',  
 '36381', '36642', '36684', '36921', '36926', '37064', '37132', '37206', '37392', '37405',  
 '37623', '37637', '37652', '37706', '37783', '37872', '38142', '38476', '38604', '38959',  
 '39543', '39665', '39890', '39897', '40194', '40195', '41440', '41451', '41474', '41493',  
 '41533', '41565', '41566', '41592', '41926', '42318', '42546', '42963', '42989', '44070',  
 '44091', '44143', '44964', '45021', '45223', '45299', '45504', '46461', '48790', '48791',  
 '48793', '48794', '48809', '49045', '49183', '49215', '49816', '49940', '49943', '49945',  
 '49951', '50009', '50240', '50245', '50456', '50489', '50873', '51138', '51195', '51198',  
 '51199', '51244', '51359', '51771', '52019', '52587', '52716', '592', '715')

AND UNIT\_STANDARD\_ID IN ('11429', '114295', '11430', '114353', '114356',  
'114360', '114363', '114367', '114375', '114379', '114473', '114480', '114747',  
'114753', '114755', '114766', '114770', '114772', '114776', '114783', '114799',  
'114822', '11490', '114904', '114906', '114907', '114908', '114909', '114910', '114911',  
'114912', '114913', '114914', '114915', '114916', '114917', '114919', '114920',  
'114921', '114922', '114923', '114924', '114925', '114926', '114927', '114928',  
'114929', '114936', '114970', '114971', '114972', '114973', '114974', '114976',  
'114977', '114979', '114981', '114983', '114986', '114987', '114988', '114990',  
'114991', '114992', '114993', '114994', '114995', '114996', '114997', '115000',  
'115001', '115002', '115234', '115240', '115375', '115379', '115380', '115382',  
'115384', '115390', '115401', '115403', '115404', '115405', '115409', '115448',  
'115770', '115789', '116173', '116177', '116182', '116184', '116216', '116217',  
'116218', '116219', '116220', '116221', '116222', '116223', '116270', '116357',  
'116362', '116363', '116370', '116375', '116381', '116382', '116406', '116411',  
'116501', '116737', '117008', '117034', '117046', '117128', '117134', '117138',  
'117143', '117144', '117145', '117146', '117147', '117149', '117150', '117151',  
'117152', '117163', '117166', '117172', '117173', '117175', '117188', '117258',  
'117261', '117433', '117775', '10001', '10002', '10003', '10004', '10006', '10019',  
'10026', '10028', '10029', '10030', '10031', '10032', '10033', '10034', '10035', '10036',  
'10037', '10038', '10039', '10040', '10041', '10042', '10043', '10044', '10045', '10046',  
'10047', '10048', '10049', '10050', '10051', '10054', '10055', '10062', '10066', '10069',  
'10071', '10075', '10137', '10138', '10139', '10140', '10141', '10143', '10144', '10146',  
'10147', '10186', '10187', '10188', '10212', '10246', '10250', '10254', '10269', '10270',  
'10271', '10272', '10312', '10330', '10338', '10339', '10340', '10341', '10343', '10344',  
'10345', '10348', '10365', '10366', '10367', '10370', '10371', '10375', '10389', '10396',  
'10398', '10403', '10404', '10405', '10406', '10408', '10505', '10599', '10602', '10604',  
'10615', '10630', '10639', '10729', '10730', '10732', '10733', '10734', '10735', '10972',  
'10990', '10995', '10997', '10998', '11000', '11002', '110061', '110092', '110488',  
'110512', '11252', '11258', '11303', '113869', '113875', '113893', '113894', '113896',  
'113920', '113921', '113924', '113926', '113928', '113929', '113935', '113938',  
'113939', '113940', '113941', '113945', '114059', '114060', '114061', '114062',  
'114063', '114064', '114065', '114066', '114067', '114068', '114069', '114070',  
'114071', '114072', '114073', '114074', '114075', '114076', '114077', '114078',  
'114079', '114080', '114081', '114082', '114083', '114084', '114085', '114086',

'114087', '114088', '114089', '114090', '114091', '114092', '114094', '114199',  
'114200', '114223', '114226', '11423', '114232', '114235', '11424', '114243', '11425',  
'11426', '11427', '11428', '252224', '252225', '252226', '252227', '252228', '252229',  
'252230', '252232', '252233', '252234', '252235', '252236', '252389', '252390',  
'252391', '252392', '252397', '252400', '252402', '252403', '254072', '254076',  
'254077', '254078', '254080', '254152', '254171', '254200', '255514', '255516',  
'255517', '256011', '258172', '258173', '258174', '258175', '258176', '258177',  
'258178', '258179', '258192', '258193', '258232', '258234', '258236', '258238',  
'258892', '258893', '258894', '258895', '258896', '258897', '258898', '258899',  
'258900', '258914', '258915', '259621', '259634', '259636', '259639', '260480',  
'260615', '263373', '263451', '263472', '263473', '263491', '263531', '263551',  
'263993', '376497', '7194', '7199', '7206', '7207', '7209', '7210', '7213', '7218', '7220',  
'7221', '7222', '7239', '7240', '7241', '7242', '7243', '7248', '7253', '7257', '7259',  
'7261', '7264', '7267', '7271', '7275', '7289', '7303', '7305', '7308', '7309', '7311',  
'7318', '7321', '7323', '7325', '7330', '7333', '7338', '7341', '7342', '7344', '7350',  
'7351', '7352', '7353', '7354', '7355', '7356', '7358', '7359', '7362', '7365', '7366',  
'7368', '7369', '7370', '7373', '7374', '7376', '7378', '7379', '7506', '7524', '7525',  
'7526', '7528', '7530', '7626', '7650', '7723', '7802', '7813', '7816', '7871', '7877',  
'7890', '7895', '7897', '7928', '8014', '8017', '8040', '8055', '8056', '8269', '8273',  
'8302', '8305', '8435', '8436', '8437', '8573', '8635', '8664', '8820', '8824', '8959',  
'9003', '9012', '9021', '9059', '9061', '9079', '9080', '9125', '9139', '9142', '9285',  
'9339', '9523', '9545', '9547', '9550', '9856', '9894', '9895', '9896', '9899', '9930',  
'9949', '9958', '9977', '9981', '9982', '9984', '9985', '9986', '9987', '9990', '117776',  
'117779', '117780', '117783', '117785', '117786', '117792', '117795', '117798',  
'117799', '117801', '117802', '117822', '117823', '117824', '117826', '117827',  
'117829', '117831', '117832', '117834', '117846', '117848', '117849', '117854',  
'117855', '117860', '117864', '117867', '117869', '117887', '117894', '117900',  
'117904', '117908', '117909', '117914', '117915', '117916', '117917', '117918',  
'117943', '118022', '118045', '118046', '118050', '118054', '118060', '118062',  
'119150', '119152', '119154', '119155', '119156', '119158', '119163', '119165',  
'119173', '119176', '119183', '119189', '119248', '119277', '119282', '119367',  
'119369', '119576', '119577', '119580', '119582', '119583', '119584', '11961', '119693',  
'119698', '119699', '11971', '119752', '119753', '119755', '119760', '119761', '119763',  
'119765', '119767', '119768', '119769', '119770', '119819', '119911', '119916',

'119921', '119926', '119932', '119973', '119974', '119977', '119978', '119979',  
'120014', '120015', '120022', '120023', '120025', '120026', '120028', '120029',  
'120030', '120031', '120032', '120033', '120034', '120035', '120036', '120037',  
'120039', '120040', '120043', '120092', '120125', '120126', '120127', '120128',  
'120129', '120130', '120131', '120132', '120133', '120134', '120135', '120136',  
'120137', '120138', '120139', '120140', '120141', '120142', '120143', '120144',  
'120145', '120146', '120147', '120148', '120149', '120150', '120151', '120152',  
'120153', '120154', '120155', '120256', '120419', '120433', '120520', '12053', '12055',  
'12056', '12057', '12152', '12157', '12181', '12198', '12236', '123151', '123269',  
'123270', '123271', '123272', '123273', '123274', '123276', '123277', '123278',  
'123279', '12333', '123384', '12340', '12349', '12352', '12450', '12473', '12478',  
'12482', '12483', '12493', '12498', '12500', '12501', '12505', '12529', '12530', '12554',  
'12564', '12565', '12567', '12592', '12667', '12752', '12753', '12754', '12755', '12757',  
'12758', '12759', '12760', '12761', '12762', '12763', '12764', '12766', '12772', '12773',  
'12776', '12779', '12780', '12781', '12899', '12903', '12904', '12917', '12920', '12926',  
'13014', '13016', '13017', '13174', '13176', '13179', '13182', '13184', '13188', '13189',  
'13191', '13193', '13234', '13237', '13239', '13241', '13317', '13319', '13321', '13392',  
'13393', '13396', '13397', '13398', '13413', '13414', '13415', '13416', '13417', '13420',  
'13431', '13432', '13433', '13435', '13437', '13438', '13444', '13446', '13447', '13450',  
'13453', '13456', '13458', '13460', '13720', '13730', '13737', '13889', '13890', '13891',  
'13900', '13902', '13903', '13942', '13957', '13967', '13978', '13979', '13980', '13989',  
'14012', '14013', '14015', '14053', '14054', '14055', '14071', '14076', '14077', '14079',  
'14080', '14082', '14333', '14336', '14431', '14432', '14433', '14434', '14435', '14442',  
'14443', '14445', '14446', '14447', '14461', '14462', '14523', '14569', '14578', '14628',  
'14630', '14637', '14649', '14655', '14658', '14659', '14667', '14674', '14677', '14679',  
'14689', '14899', '14900', '14901', '14902', '14903', '14904', '14905', '14906', '14907',  
'14908', '14909', '14910', '14911', '14912', '14915', '14920', '14996', '15001', '15005',  
'15008', '15014', '15025', '15051', '15071', '15076', '15078', '15081', '15085', '15088',  
'15098', '15099', '15103', '15105', '15106', '15140', '15186', '15189', '15202', '15238',  
'15244', '15353', '230015', '230038', '230087', '230088', '230090', '230091', '230092',  
'230094', '230095', '242571', '242572', '242573', '242574', '242575', '242576',  
'242577', '242578', '242579', '242580', '242581', '242582', '242583', '242584',  
'242585', '242586', '242587', '242588', '242589', '242590', '242591', '242592',  
'242593', '242594', '242595', '242596', '242597', '242598', '242599', '242600',

'242601', '242602', '242603', '242604', '242605', '242606', '242607', '242608',  
 '242609', '242610', '242611', '242612', '242613', '242614', '242615', '242616',  
 '242617', '242618', '242619', '242620', '242621', '242622', '242623', '242624',  
 '242625', '242626', '242627', '242628', '242629', '242630', '242631', '242632',  
 '242633', '242634', '242635', '242636', '242672', '242685', '242802', '242804',  
 '242867', '242869', '242870', '242871', '242873', '242874', '242875', '242876',  
 '242877', '242879', '242880', '242881', '242882', '242883', '242885', '242887',  
 '242891', '243035', '243206', '243210', '243211', '243215', '243223', '243224',  
 '243682', '243683', '243689', '243690', '243693', '243695', '243696', '243697',  
 '243698', '243729', '243821', '243823', '243826', '243827', '243964', '244078',  
 '244090', '244095', '244105', '244139', '244140', '244143', '244153', '244160',  
 '244161', '244185', '244300', '244304', '244509', '244510', '244513', '244515',  
 '244516', '246460', '246476', '246481', '246483', '246488', '246489', '246490',  
 '246552', '246750', '246751', '246752', '246753', '246754', '246755', '246756',  
 '252037', '252038', '252039', '252042', '252043', '252044', '252046', '252048',  
 '252049', '252051', '252052', '252054', '252057', '252058', '252059', '252060',  
 '252061', '252207', '252208', '252209', '252210', '252211', '252212', '252214',  
 '252215', '252216', '252217', '252218', '252219', '252220', '252221', '252222',  
 '252223')

AND SUBFIELD\_DESC IN ('Building Construction', 'Civil Engineering Construction',  
 'Cleaning, Domestic, Hiring, Property and Rescue Services', 'Curative Health', 'Finance,  
 Economics and Accounting', 'Generic Management', 'Hospitality, Tourism, Travel, Gaming  
 and Leisure', 'Human Resources', 'Information Technology and Computer Sciences',  
 'Manufacturing and Assembly', 'Marketing', 'Preventive Health', 'Primary Agriculture', 'Public  
 Administration', 'Transport, Operations and Logistics', 'Wholesale and Retail')

AND FIELD\_DESC IN ('Business, Commerce and Management Studies', 'Health Sciences and  
 Social Services', 'Manufacturing, Engineering and Technology', 'Physical Planning and  
 Construction', 'Services')

AND PROV\_PROVINCE\_DESC IN ('Gauteng', 'Kwazulu/Natal', 'Western Cape')

AND PROV\_ETQE\_ID IN ('1031', '1102', '1109', '1112', '1114', '1117', '1120', '1123', '1125',  
 '1126', '1127')

AND PROV\_ACCRED\_IND\_DESC IN ('Start Before, End Before')

AND PROVIDER\_CLASS\_DESC IN ('Mixed: Public and Private', 'Private')

AND UNIT\_STD\_TYPE\_DESC IN ('Regular')

- Cluster 3

% of records: 20.86%

Average probability: 0.9880

Rule:

```
ASSESSOR_ID IN ('16129194', '3013505', '3445519', '4707500', '4783536', '6130511',  
'8934272', 'NULL')  
AND UNIT_STANDARD_ID IN ('10152', '10156', '10157', '110040', '110081',  
'113983', '114093', '114600', '114606', '114609', '114613', '114653', '114958',  
'115104', '115109', '115110', '115118', '115122', '115376', '115408', '115838',  
'116537', '116544', '116550', '116551', '116947', '116948', '116949', '116952',  
'116953', '116954', '116957', '116959', '116960', '116962', '117016', '117884',  
'117919', '117940', '117941', '117942', '117944', '117945', '119095', '119381',  
'119385', '119390', '119393', '119471', '119473', '119474', '119476', '119477',  
'119479', '119480', '119481', '119482', '119483', '119484', '119486', '119488',  
'119489', '119648', '119649', '119652', '119653', '119654', '119656', '119657',  
'119683', '119685', '119687', '119690', '119691', '120317', '120325', '120389',  
'120390', '120396', '120401', '120402', '12170', '12171', '12172', '12434', '12461',  
'12479', '12486', '12487', '12488', '13186', '13928', '13929', '13931', '13932', '13933',  
'13934', '13935', '13936', '13946', '13947', '13948', '13949', '13951', '13953', '13958',  
'13960', '13964', '13965', '13966', '13970', '14355', '14358', '14359', '14361', '14364',  
'14365', '14366', '14367', '14368', '14369', '14370', '14376', '14673', '14676', '14682',  
'14684', '14925', '14926', '14927', '14929', '14930', '14932', '14934', '14935', '14938',  
'14941', '14943', '14964', '15109', '15251', '15253', '242828', '242833', '242836',  
'242838', '244595', '244601', '252053', '7464', '7465', '7466', '7467', '7468', '7469',  
'7473', '7478', '7480', '7481', '7485', '7486', '7497', '7543', '7564', '7583', '7584',  
'7585', '7587', '7588', '7590', '7592', '8121', '8572', '8979', '8980', '8981', '8982',  
'8983', '8984', '8985', '8986', '8987', '8989', '8990', '8991', '8992', '8993', '8996',  
'9024', '9025', '9026', '9027', '9029', '9030', '9032', '9033', '9319', '9320')  
AND PROVIDER_ID IN ('10196', '10923', '11088', '11100', '11185', '11200', '11228',  
'11240', '11242', '11244', '11246', '11250', '11255', '11262', '11305', '11468', '11772',  
'11990', '12489', '12592', '12666', '14450', '14568', '14640', '14795', '14928', '15241',  
'1575', '1578', '16107', '1680', '18511', '18698', '1898', '1914', '1916', '1945', '20584',  
'2084', '21028', '21121', '2151', '2158', '2160', '21830', '2184', '21961', '21966', '2206',  
'22278', '22294', '22745', '22831', '22832', '22911', '22993', '23152', '23220', '23498',  
'24841', '24850', '24870', '25033', '25097', '25098', '25366', '25384', '2626', '26304',
```

'26342', '26362', '26377', '26432', '2657', '2682', '27063', '27078', '27090', '27093',  
 '27097', '27128', '27135', '28305', '28360', '28872', '29212', '29273', '29275', '29277',  
 '29313', '29368', '29370', '29371', '29732', '29835', '29888', '29980', '30127', '30217',  
 '30794', '31670', '31705', '31724', '31725', '31726', '31750', '31766', '32020', '32167',  
 '32169', '32197', '32264', '32520', '32695', '32848', '32885', '32918', '32919', '32933',  
 '33644', '34827', '34946', '35378', '35381', '35404', '35405', '35493', '35675', '35683',  
 '35735', '35816', '35920', '35925', '35926', '35940', '36189', '36372', '36381', '36467',  
 '36684', '36831', '36858', '36921', '37064', '37303', '37378', '37379', '37392', '37405',  
 '37406', '37623', '37637', '37662', '37688', '37694', '37726', '37783', '37872', '38456',  
 '38476', '38542', '38548', '38551', '38554', '38555', '38560', '38561', '38563', '38568',  
 '38570', '38574', '38575', '38578', '38590', '38596', '38604', '38608', '38611', '38633',  
 '38660', '38662', '38676', '38989', '39011', '39027', '39068', '39665', '39834', '39897',  
 '39919', '39923', '39934', '40194', '41440', '41451', '41474', '41493', '41565', '41566',  
 '41592', '41926', '42182', '42522', '42546', '42946', '42963', '42987', '42989', '44091',  
 '44143', '44224', '44322', '45223', '45299', '45504', '46459', '46461', '46624', '46658',  
 '47409', '47446', '47510', '47620', '47630', '47651', '48679', '48790', '48791', '48794',  
 '48809', '49045', '49183', '49215', '49342', '49724', '49940', '49943', '50009', '5004',  
 '50240', '50245', '50456', '50489', '50873', '51082', '51119', '51129', '51138', '51144',  
 '51153', '51158', '51191', '51195', '51199', '51209', '51213', '51222', '51225', '51249',  
 '51250', '51499', '51588', '51590', '51771', '51820', '52019', '5205', '52716', '715')

AND SUBFIELD\_DESC IN ('Generic Management', 'Information Technology and  
 Computer Sciences', 'Language', 'Mathematical Sciences', 'Office Administration',  
 'People/Human-Centred Development', 'Preventive Health')

AND FIELD\_DESC IN ('Business, Commerce and Management Studies', 'Communication  
 Studies and Language', 'Physical, Mathematical, Computer and Life Sciences')

AND UNIT\_STD\_TYPE\_DESC IN ('Regular-Fundamental')

AND PROV\_ETQE\_ID IN ('1031', '1102', '1103', '1105', '1111', '1114', '1115', '1116',  
 '1117', '1120', '1123', '1125', '1126')

AND ENROL\_TYPE\_DESC IN ('Mixed Mode', 'RPL for Unknown Purpose',  
 'Unknown', 'Work Place Learning')

AND PROV\_PROVINCE\_DESC IN ('Eastern Cape', 'Gauteng', 'Kwazulu/Natal',  
 'Limpopo', 'South Africa National', 'Western Cape')

- Cluster 4

% of records: 11.01%

Average probability: 0.9713

Rule:

```
ASSESSOR_ID IN ('3021924', 'NULL')
AND UNIT_STANDARD_ID IN ('10024', '10246', '10269', '10765', '10767', '10771',
'10773', '10998', '11002', '11219', '11303', '113869', '113926', '113941', '114243',
'114996', '115075', '11511', '11512', '11513', '11514', '11515', '11516', '11517',
'11518', '11519', '11520', '11522', '11525', '11526', '11530', '116146', '116252',
'116551', '116611', '116731', '116737', '117722', '117850', '117888', '117894',
'117917', '119156', '119179', '119186', '119359', '119666', '119667', '119668',
'119669', '119752', '119755', '119756', '119760', '119761', '119762', '119763',
'119765', '119767', '119768', '119769', '119770', '11998', '12002', '12007', '120493',
'120494', '120495', '120496', '120497', '120498', '120499', '120500', '120501',
'120502', '120503', '120504', '120505', '120506', '120508', '120509', '120511',
'120512', '12236', '12242', '12345', '123527', '123528', '123529', '123530', '123531',
'123532', '123536', '12482', '12483', '12493', '12500', '12501', '12505', '12908',
'13035', '13078', '13079', '13189', '13237', '13275', '13299', '13929', '13953', '14462',
'14597', '14611', '14624', '14635', '14639', '14640', '14679', '14688', '15113', '15116',
'15117', '15118', '15120', '15122', '15124', '15126', '15127', '15130', '15131', '15133',
'15134', '15135', '230038', '230039', '230040', '230041', '230042', '230043', '230045',
'230046', '230482', '242829', '242842', '242847', '242850', '243084', '243085',
'243128', '243131', '243951', '244065', '244066', '244068', '244069', '244070',
'244073', '244074', '244125', '244137', '244139', '244140', '244143', '244145',
'244148', '244152', '244153', '244160', '244161', '244180', '244185', '244187',
'244193', '244194', '244196', '244198', '244199', '244201', '244206', '244305',
'244352', '244595', '244703', '244707', '244708', '246710', '246711', '253442',
'253457', '253991', '253993', '253995', '253996', '253997', '253998', '253999',
'254000', '254002', '254003', '254004', '254005', '254007', '254009', '254010',
'256071', '259614', '259621', '259641', '259739', '259754', '259779', '259954',
'259955', '259956', '260454', '260654', '260655', '260696', '260734', '260735',
'260736', '260738', '260740', '261676', '261680', '261681', '335871', '335872',
'335876', '335877', '335881', '335913', '335914', '335917', '376497', '7876', '7900',
'8063', '8617', '9981', '9982', '9986')
AND PROVIDER_ID IN ('23172', '23220', '25033', '25095', '25097', '25098', '25136',
'25152', '25177', '2626', '26499', '2657', '2682', '28872', '28873', '28921', '28960',
```

'30122', '30127', '30216', '32848', '32849', '32851', '32885', '35493', '35816', '36101',  
'36189', '36191', '37456', '37508', '37529', '38273', '39833', '39834', '41525', '41566',  
'44899', '45299', '45504', '46508', '46514', '46581', '46598', '46624', '46637', '46648',  
'46658', '46666', '46733', '46735', '46736', '46764', '46783', '46813', '46907', '46927',  
'46931', '46940', '46942', '46954', '46955', '46958', '47070', '47109', '47215', '47240',  
'47278', '47325', '47329', '47350', '47368', '47390', '47409', '47412', '47446', '47480',  
'47488', '47497', '47510', '47512', '47525', '47534', '47554', '47570', '47586', '47620',  
'47627', '47630', '47651', '47698', '47742', '47762', '47798', '47813', '47967', '47999',  
'48053', '48150', '48169', '48244', '48247', '48253', '48293', '48295', '5004', '51075',  
'51771', '51819', '51820', '5205', '10196', '10923', '17142', '2066', '20725', '20743',  
'20772', '20818', '20834', '20846', '20860', '20861', '21009', '21016', '21017', '21121',  
'21188', '21942', '21961', '21966', '22273', '22294', '22911', '22993', '23142', '23152',  
'23169', '23170', '23171')

*AND SUBFIELD\_DESC IN ('Building Construction', 'Finance, Economics and  
Accounting', 'Manufacturing and Assembly', 'Safety in Society', 'Transport, Operations  
and Logistics')*

*AND PROV\_ETQE\_ID IN ('1103', '1105')*

*AND PROV\_PROVINCE\_DESC IN ('South Africa National')*

*AND ETQE\_ID IN ('1103', '1105')*

*AND ENROL\_TYPE\_DESC IN ('Mixed Mode')*

*AND PROVIDER\_CLASS\_DESC IN ('Mixed: Public and Private')*

*AND PROVIDER\_TYPE\_DESC IN ('Education and Training')*

- Cluster 5

% of records: 5.73%

Average probability: 0.9892

Rule:

*UNIT\_STANDARD\_ID IN ('10305', '10306', '10311', '10312', '110075', '110077',  
'110078', '114493', '114600', '114602', '114607', '114609', '114610', '114613',  
'114941', '114955', '114959', '115770', '115772', '115806', '115807', '117882',  
'117891', '117894', '117912', '119474', '119476', '119479', '119482', '119484',  
'119486', '119488', '119489', '119678', '119679', '120053', '123411', '123413',  
'123414', '123415', '13660', '13870', '13871', '13872', '13873', '13942', '14461',  
'14599', '14817', '15238', '15239', '15245', '242829', '242833', '242836', '244273',*

'244274', '244276', '244277', '244479', '244485', '244486', '244489', '244492',  
'244495', '244497', '244498', '244501', '244502', '244588', '244627', '7417', '7425',  
'7426', '7427', '7485', '7995', '8664', '9032', '9033')

AND PROVIDER\_ID IN ('14607', '28025', '28046', '28049', '28052', '28060', '28077',  
'28089', '28133', '28140', '28146', '28162', '28186', '28189', '28205', '28225', '28234',  
'28244', '28259', '28273', '28275', '28289', '28293', '28305', '28307', '28311', '28328',  
'28340', '28347', '28387', '28438', '28504', '29967', '32693', '32705', '35590', '36220',  
'36242', '37425', '39919', '49583', '50114', '50125', '50578')

AND ASSESSOR\_ID IN ('3004701', '3007815', '3008495', '3008567', '3009223',  
'3009227', '3011147', '3014782', '3015053', '3015208', '3018377', '3018856',  
'3021267', '3021501', '3021502', '3022113', '3023994', '3024132', '3024534',  
'3024873', '3027229', '3029008', '3029212', '3030037', '3030552', '3030624',  
'3031555', '3031754', '3312964', '3313026', '3313040', '3313051', '3313066',  
'3313073', '3313107', '3313363', '3557354', '4282671', '4282752', '4282949',  
'4282973', '4283241', '4631855', '4631893', '4631906', '4632187', '4632205',  
'5013225', '5013555', '5013688', '5013700', '5013724', '5518085', '5518102',  
'5518257', '5518273', '5518294', '5518382', '5518383', '5518384', '5518385',  
'6055536', '6055807', '6055911', '6056024', '7309219', '7309236', '7309349',  
'7309474', '7309547', '8145155', '8145178', '8145317', '8145490', '9050564',  
'9050578', '9463777', '9464189', '9464227', '9464228', '9464322', '9574801',  
'9574840', '9574855', 'NULL')

AND PROV\_ETQE\_ID IN ('1106')

AND ETQE\_ID IN ('1106')

AND SUBFIELD\_DESC IN ('Adult Learning', 'Early Childhood Development',  
'Generic Management', 'Higher Education and Training', 'Hospitality, Tourism,  
Travel, Gaming and Leisure', 'Human Resources', 'Language', 'Mathematical  
Sciences', 'People/Human-Centred Development')

AND ENROL\_TYPE\_DESC IN ('Residential Learning (i.e. Contact Mode)')

AND PROVIDER\_CLASS\_DESC IN ('Unknown')

AND PROVIDER\_TYPE\_DESC IN ('Education and Training')

AND PROV\_PROVINCE\_DESC IN ('Eastern Cape', 'Free State', 'Gauteng', 'Kwazulu/Natal',  
'Limpopo', 'Mpumalanga', 'Western Cape')

- Cluster 6

% of records: 5.08%

Average probability: 0.9985

Rule:

PROVIDER\_TYPE\_DESC IN ('NULL')

AND PRIMARY\_ETQE\_DESC IN ('Not Primary ETQE of provider')

AND PROVIDER\_CLASS\_DESC IN ('NULL')

AND ASSESSOR\_ID IN ('5664362', '5664432', '5664633', '5664641', '5664693',  
'5664943', 'NULL')

AND PROVIDER\_ID IN ('12247', '14814', '1635', '16932', '1696', '17020', '1758',  
'20724', '26376', '28720', '36875', '38597', '38625', '39205', '39208', '41226', '43436',  
'44305', '44320', '44688', '46645', '46714', '47330', '47498', '47584', '48231', '48625')

AND UNIT\_STANDARD\_ID IN ('10001', '10023', '10024', '10028', '10036', '10037',  
'10044', '10246', '10366', '10404', '10405', '10406', '10568', '10570', '10572', '10573',  
'10584', '10765', '10767', '10771', '10773', '10920', '10923', '10926', '10929', '10932',  
'110020', '110092', '110135', '114236', '114600', '114601', '114603', '114604',  
'114605', '114606', '114607', '114608', '114609', '114610', '114611', '114612',  
'114613', '114615', '114635', '11490', '114906', '114908', '114911', '114913', '114920',  
'114923', '11513', '11514', '11515', '11516', '11517', '11518', '11519', '11522', '11526',  
'116356', '116357', '116358', '116359', '116360', '116361', '116362', '116363',  
'116364', '116365', '116368', '116370', '116374', '116375', '116377', '116378',  
'116379', '116380', '116381', '116454', '116653', '116687', '116949', '11715', '117722',  
'119109', '119112', '119136', '119348', '119351', '119358', '119360', '119365',  
'119367', '119368', '119369', '119370', '119482', '119484', '119666', '119667',  
'119668', '119669', '119780', '120317', '120322', '12170', '12198', '12345', '123532',  
'12434', '12461', '12479', '12486', '12488', '12501', '12554', '12564', '12565', '12778',  
'12892', '12916', '12917', '12920', '12922', '12924', '12925', '12926', '12927', '12929',  
'12930', '12931', '12933', '12935', '13174', '13176', '13179', '13182', '13189', '13191',  
'13694', '13889', '13902', '13903', '13957', '13962', '13964', '13969', '13988', '13989',  
'13990', '14012', '14071', '14097', '14599', '14817', '14899', '14900', '14901', '14902',  
'14903', '14905', '14906', '14907', '14908', '14909', '14910', '14911', '14912', '15076',  
'15080', '15082', '15087', '15090', '15094', '15096', '15098', '15099', '15103', '15105',  
'15106', '15199', '15297', '15316', '243697', '243729', '244450', '244470', '258172',  
'259957', '259959', '259960', '259961', '259962', '7464', '7465', '7466', '7467', '7468',  
'7473', '7478', '7480', '7485', '7486', '7497', '7520', '7524', '7525', '7526', '7528',

'7530', '7564', '7638', '7650', '7808', '8262', '8265', '8273', '8280', '8284', '8290',  
'8292', '8295', '8297', '8298', '8302', '8305', '8578', '8617', '8979', '8980', '8981',  
'8984', '8985', '8986', '8987', '8990', '8991', '8992', '8993', '8996', '9024', '9026',  
'9027', '9029', '9030', '9032', '9033', '9374', '9617', '9618', '9695', '9706', '9707',  
'9708', '9710', '9717', '9866', '9977', '9981', '9982', '9984')

*AND PROV\_PROVINCE\_DESC IN ('NULL')*

*AND SUBFIELD\_DESC IN ('Building Construction', 'Cleaning, Domestic, Hiring, Property  
and Rescue Services', 'Fabrication and Extraction', 'Finance, Economics and Accounting',  
'Generic Management', 'Language', 'Manufacturing and Assembly', 'Mathematical Sciences',  
'Public Administration', 'Safety in Society', 'Wholesale and Retail')*

*AND ENROL\_TYPE\_DESC IN ('Mixed Mode', 'Unknown')*

- Cluster 7

% of records: 4.75%

Average probability: 0.9769

Rule:

*PRIMARY\_ETQE\_DESC IN ('Not Primary ETQE of provider')*

*AND PROVIDER\_ID IN ('21318', '21328', '21332', '21336', '21337', '21342', '21375',  
'21665', '21780', '32264', '33812', '36467', '48475', '48477', '48478', '48534', '48558',  
'48591', '592', '663', '672')*

*AND ASSESSOR\_ID IN ('3002101', '3008406', '3011448', '3013505', '3024384',  
'3027008', '3027166', '3027363', '3031299', '3032933', '3040050', '3313267',  
'4282640', '4282641', '4282714', '4282924', '4631724', '4631972', '4632259',  
'4632260', '4632261', '4632262', '4783551', '4783571', '4783572', '4783603',  
'5013309', '5013828', '5518246', '5518529', '5518570', '5664357', '5664419',  
'5664556', '6056024', '7309240', '7363907', '8145591', '9464452', '9574572', 'NULL')*

*AND UNIT\_STANDARD\_ID IN ('123413', '123414', '12461', '12478', '12480', '12482',  
'12483', '12493', '12859', '13237', '13660', '13864', '13865', '13866', '13867', '13868',  
'13869', '13870', '13871', '13872', '13873', '13942', '14071', '14477', '14586', '14599',  
'14676', '14678', '14681', '14899', '14900', '14901', '14902', '14903', '14904', '14905',  
'14906', '14907', '14908', '14909', '14910', '14911', '14912', '15117', '15234', '15235',  
'15238', '15249', '242829', '242833', '242836', '243084', '244068', '244180', '244273',  
'244274', '244275', '244276', '244277', '244479', '244485', '244486', '244489',  
'244492', '244495', '244497', '244498', '244501', '244502', '244587', '259956',*

'260654', '260735', '7417', '7425', '7426', '7427', '7465', '7466', '7467', '7468', '7470',  
 '7473', '7478', '7480', '7481', '7485', '7486', '7541', '7543', '7545', '7547', '7551',  
 '7552', '7995', '8664', '8979', '8980', '8981', '8984', '8985', '8986', '8987', '8991',  
 '8992', '8993', '8996', '9024', '9025', '9026', '9027', '9029', '9030', '9032', '9033',  
 '9285', '9339', '9460', '9550', '9897', '9952', '9977', '9981', '9982', '9984', '9985',  
 '10001', '10023', '10305', '10306', '10307', '10311', '10312', '113941', '113983',  
 '114600', '114602', '114607', '114609', '114610', '114613', '114941', '114955',  
 '114959', '114976', '114991', '115770', '115772', '115776', '116356', '116357',  
 '116358', '116359', '116360', '116361', '116362', '116363', '116364', '116365',  
 '116368', '116370', '116374', '116375', '116377', '116378', '116379', '116380',  
 '116381', '116737', '116949', '117173', '117882', '117888', '117891', '117912',  
 '119074', '119076', '119348', '119351', '119358', '119360', '119365', '119367',  
 '119368', '119369', '119370', '119379', '119381', '119390', '119471', '119473',  
 '119474', '119476', '119477', '119479', '119480', '119482', '119484', '119486',  
 '119488', '119489', '119651', '119652', '119653', '119657', '119678', '119679',  
 '119683', '119685', '119687', '119689', '119761', '119770', '120053', '12236')

AND PROV\_ETQE\_ID IN ('1031', '1033')

AND PROVIDER\_CLASS\_DESC IN ('Public')

AND SUBFIELD\_DESC IN ('Adult Learning', 'Building Construction', 'Early Childhood  
 Development', 'Generic Management', 'Higher Education and Training', 'Language',  
 'Manufacturing and Assembly', 'Mathematical Sciences', 'People/Human-Centred  
 Development', 'Public Administration')

AND ETQE\_ID IN ('1103', '1106', '1109', '1110')

AND PROVIDER\_TYPE\_DESC IN ('Education')

AND PROV\_PROVINCE\_DESC IN ('Gauteng', 'Mpumalanga', 'North West', 'Western Cape')

- Cluster 8

% of records: 4.10%

Average probability: 0.9236

Rule:

UNIT\_STANDARD\_ID IN ('10017', '10018', '10019', '10023', '10024', '10148',  
 '10163', '10188', '10248', '10305', '10306', '10307', '10311', '10312', '10371', '110061',  
 '110070', '110071', '110073', '110075', '110076', '110077', '110078', '110492', '11258',  
 '11303', '114235', '114493', '114495', '114509', '114600', '114602', '114609', '114613',  
 '114906', '114910', '114911', '114913', '114920', '114923', '114927', '114941',

'114955', '114959', '115450', '115770', '115776', '115806', '115807', '115809',  
'116081', '116180', '116181', '116356', '116358', '116359', '116364', '116365',  
'116370', '116374', '116378', '117034', '117173', '117182', '117194', '117202',  
'117203', '117213', '117216', '117873', '117882', '117887', '117888', '117891',  
'117894', '117904', '117943', '11926', '119351', '119353', '119369', '119474', '119476',  
'119478', '119479', '119482', '119484', '119486', '119488', '119489', '119570',  
'119571', '119582', '119678', '119683', '119685', '119686', '119687', '119689',  
'119691', '119729', '119739', '119742', '119743', '119814', '119815', '119817',  
'119819', '120252', '120256', '120396', '120399', '120402', '120418', '120420',  
'120421', '120427', '120433', '120434', '12050', '12155', '12157', '12170', '12332',  
'123378', '123391', '123392', '123411', '123413', '123414', '123415', '123418', '12554',  
'12859', '12927', '13660', '13870', '13871', '13872', '13873', '13942', '13994', '14016',  
'14071', '14241', '14243', '14588', '14598', '14599', '14637', '14649', '14659', '14667',  
'14671', '14674', '14677', '14679', '14686', '14689', '14690', '14691', '14692', '14693',  
'14695', '14696', '14899', '14900', '14901', '14904', '14906', '14908', '14909', '14910',  
'14911', '14912', '15110', '15154', '15234', '15235', '15238', '15239', '15244', '15245',  
'15249', '15258', '242794', '242795', '242796', '242798', '242801', '242805', '242806',  
'242809', '242811', '242827', '242829', '242833', '242834', '242836', '243027',  
'243035', '243036', '243040', '243043', '243045', '243046', '243049', '243050',  
'243210', '243214', '243220', '243221', '243222', '243223', '244105', '244273',  
'244274', '244276', '244277', '244479', '244485', '244486', '244489', '244492',  
'244495', '244497', '244498', '244501', '244502', '244588', '244627', '246489',  
'246596', '246597', '246598', '246599', '246601', '246602', '246750', '246751',  
'246752', '246753', '246754', '246755', '246756', '252058', '259621', '261754',  
'263993', '264277', '264991', '264992', '264993', '264994', '264995', '264996',  
'264997', '264998', '7401', '7404', '7417', '7485', '7802', '7807', '7808', '7880', '7893',  
'8347', '8664', '8783', '8821', '8822', '8824', '8839', '8922', '8940', '8959', '9029',  
'9032', '9033', '9059', '9063', '9064', '9068', '9079', '9080', '9128', '9550', '9899',  
'9930', '9977', '9979', '9981', '9982', '9984', '9985', '9986', '9987', '9990')  
AND PROVIDER\_ID IN ('11244', '11691', '14607', '20574', '20577', '20584', '2184',  
'2192', '2197', '2206', '22831', '22832', '25366', '25384', '26995', '27090', '27092',  
'27097', '27100', '28025', '28046', '28050', '28052', '28060', '28122', '28146', '28161',  
'28162', '28165', '28186', '28189', '28225', '28229', '28244', '28245', '28259', '28265',  
'28273', '28275', '28289', '28298', '28302', '28305', '28307', '28311', '28314', '28315',

'28328', '28338', '28340', '28347', '28397', '28398', '28438', '28514', '28517', '29277',  
'29963', '29967', '29980', '30794', '32098', '32693', '32705', '32933', '34125', '34551',  
'34946', '35675', '35735', '35816', '35920', '36220', '36242', '36467', '37303', '37379',  
'37405', '37425', '37427', '37645', '38226', '39170', '39674', '39919', '39998', '41533',  
'41566', '42096', '42100', '42522', '42875', '42888', '44143', '44747', '50114', '50125',  
'50489', '50578')

AND ASSESSOR\_ID IN ('16123886', '17116927', '17151500', '17152126', '17369109',  
'17374505', '17374564', '17374813', '17374961', '3002391', '3004556', '3005166',  
'3005196', '3005384', '3005520', '3006033', '3008567', '3009227', '3010955',  
'3010964', '3011029', '3011652', '3011961', '3012013', '3012234', '3014998',  
'3015053', '3018377', '3018856', '3020604', '3020705', '3021349', '3021501',  
'3021502', '3021694', '3022113', '3024103', '3024426', '3024552', '3024873',  
'3024887', '3025410', '3027104', '3028350', '3028806', '3028825', '3029008',  
'3029571', '3029674', '3029817', '3030109', '3030624', '3031555', '3031732',  
'3031754', '3032635', '3033025', '3033062', '3033203', '3039613', '3040061',  
'3059514', '3059515', '3059546', '3312930', '3313011', '3313040', '3313055',  
'3313057', '3313058', '3313226', '3313287', '3313288', '3349640', '3444826',  
'3446748', '3482940', '3483001', '3483274', '3483283', '3510763', '3511014',  
'3557298', '3557318', '3575464', '3588732', '4282671', '4282752', '4282944',  
'4282955', '4282970', '4282973', '4283014', '4283068', '4283123', '4283241',  
'4389374', '4631723', '4631751', '4631777', '4631781', '4631854', '4631855',  
'4631864', '4631961', '4631994', '4632014', '4632023', '4632142', '4632187',  
'4632205', '4632223', '4632228', '4632229', '4632230', '4632231', '4632232',  
'4632233', '4632234', '4632236', '4632237', '4632240', '4632253', '4632254',  
'4632255', '4783540', '4783551', '4783571', '4783581', '4783604', '5013426',  
'5013521', '5013535', '5013627', '5013661', '5517836', '5517847', '5517930',  
'5518040', '5518124', '5518273', '5518382', '5518383', '5518385', '5664834',  
'5665008', '6055330', '6055670', '6055807', '6055883', '6055940', '6055989',  
'7309011', '7309236', '7309349', '7309368', '7309556', '7354188', '7355874',  
'7363836', '7363846', '7363907', '8145155', '8145490', '8145645', '8934272',  
'9050086', '9050342', '9050524', '9050564', '9050580', '9050592', '9463971',  
'9464076', '9464227', '9464228', '9464254', '9464341', '9464377', '9464381',  
'9464465', '9574717', '9574855', '9574906', '9575147', 'NULL')

AND PROV\_ETQE\_ID IN ('1106', '1107', '1109', '1112', '1126')

```

AND ETQE_ID IN ('1106', '1107', '1109', '1126')
    AND PROVIDER_TYPE_DESC IN ('Education and Training', 'Employer')
    AND -10.7 <= END_PROV_ACCRED_IND <= 0
    AND SUBFIELD_DESC IN ('Adult Learning', 'Building Construction', 'Civil Engineering
Construction', 'Cleaning, Domestic, Hiring, Property and Rescue Services', 'Early Childhood
Development', 'Generic Management', 'Hospitality, Tourism, Travel, Gaming and Leisure',
'Human Resources', 'Language', 'Manufacturing and Assembly', 'Mathematical Sciences',
'People/Human-Centred Development', 'Primary Agriculture', 'Public Administration',
'Wholesale and Retail')
    AND -22.2 <= START_PROV_ACCRED_IND <= -1
        AND ENROL_STATUS_DESC IN ('Achieved')

```

### ***M.2.2 Start During, Start After and End After cluster data mining***

This section provides a technical description of the clusters that were generated by cluster data mining the consolidated data category ‘Start During, Start After and End After’ (see Appendix L.2.8) for unit standard enrolment records. A detailed description of the manner in which cluster data mining was conducted can be found in Appendix I.3. The results of the generated clustering model were significant because the model was measured as being 96.24% accurate.

The generated clusters show a tight coupling between data fields that describe the ETQEs, unit standards and providers. This is as a result of the organic relationship between unit standards and ETQEs and providers and ETQEs (providers generally offer unit standards that are implemented by their primary ETQE).

The model did not generate any clusters that contained less than 1% of the records. The detection of anomalous records based on cluster specific probabilities of less than .6 being allocated to the record was conducted. In this manner it was determined that 5.94% of the records possibly exist in this category as a result of data capturing problems (see Appendix I.3).

The technical description of each of the 8 clusters is provided below. The description is limited to the top 10 attributes of the cluster rule, where each indent of text represents the importance of the attribute. For example:

A line formatted like this represents an importance of 100%

*A line formatted like this represents an importance greater than 75% and less than 100%*

*A line formatted like this represents an importance greater than 50% and less than or equal to 75%*

*A line formatted like this represents an importance less than or equal to 50%*

- Cluster 1

% of records: 27.81%

Average probability: 0.9916

Rule:

*ASSESSOR\_ID IN ("NULL")*

*AND UNIT\_STANDARD\_ID IN ("113869", "113926", "113941", "114958", "114996", "11522", "11525", "11530", "116121", "116146", "116148", "116150", "116151", "116158", "116160", "116162", "116551", "117722", "119359", "119666", "119667", "119668", "119669", "11998", "12002", "12007", "120327", "120493", "120494", "120495", "120496", "120497", "120498", "120499", "120500", "120501", "120502", "120503", "120504", "120505", "120506", "120508", "120509", "120510", "120511", "120512", "123527", "123528", "123529", "123530", "123531", "123532", "123535", "123536", "12501", "13929", "13953", "14663", "242842", "244193", "244194", "244196", "244198", "244199", "244201", "244206", "244352", "244595", "246710", "246711", "252191", "7473", "9027", "9029", "9030")*

*AND PROVIDER\_ID IN ("17146", "17156", "2066", "20669", "20689", "2071", "20725", "20785", "20798", "20860", "20861", "20964", "20997", "21012", "21014", "21015", "21016", "21017", "21089", "21114", "21121", "21164", "21188", "23092", "23096", "23169", "23170", "23171", "23172", "25152", "25158", "30137", "35501", "35516", "35551", "37488", "37523", "37532", "37540", "38319", "38346", "38348", "38357", "38377", "46514", "46926", "47700", "47762")*

*AND PROV\_ETQE\_ID IN ("1105")*

*AND SUBFIELD\_DESC IN ("Finance, Economics and Accounting", "Mathematical Sciences", "Public Administration", "Safety in Society")*

*AND ETQE\_ID IN ("1105")*

*AND PROV\_PROVINCE\_DESC IN ("South Africa National")*

*AND FIELD\_DESC IN ("Business, Commerce and Management Studies", "Law, Military Science and Security")*

*AND ENROL\_TYPE\_DESC IN ("Mixed Mode")*

*AND PROVIDER\_CLASS\_DESC IN ("Mixed: Public and Private")*

- Cluster 2

% of records: 19.71%

Average probability: 0.9488

Rule:

*ASSESSOR\_ID* IN ("17162273", "17369081", "3017504", "5663537", "5663618", "5664357", "5664961", "8934347", "9015258", "9382784", "NULL")

*AND PROVIDER\_ID* IN ("11088", "11091", "11092", "11098", "11244", "12592", "12666", "12894", "13248", "14640", "14795", "1578", "17119", "1726", "18758", "1915", "19858", "20357", "20933", "21464", "21985", "2206", "22168", "22680", "22687", "25095", "25426", "26362", "26380", "26986", "27104", "28156", "28531", "29238", "29260", "29272", "29273", "29298", "29342", "31530", "31726", "31909", "31998", "32014", "32044", "32072", "32209", "35920", "35940", "36061", "36643", "37105", "37206", "37264", "37392", "37583", "37715", "38548", "38555", "38560", "38640", "39208", "39612", "39637", "39665", "39728", "39735", "40926", "40939", "41386", "41493", "42656", "43922", "43923", "43926", "44003", "44010", "44070", "44087", "44092", "44093", "44095", "44109", "44320", "44728", "44880", "46423", "48559", "48679", "48809", "49279", "49533", "49783", "49792", "51338", "51339", "663", "672")

*AND UNIT\_STANDARD\_ID* IN ("15176", "15186", "15189", "15192", "15194", "15197", "15200", "15201", "15202", "15212", "15231", "15232", "15234", "15235", "15237", "15238", "15239", "15244", "15245", "15246", "15247", "15249", "15282", "230087", "230088", "230090", "230091", "230092", "230094", "230095", "242571", "242572", "242573", "242574", "242575", "242576", "242577", "242578", "242579", "242580", "242581", "242582", "242583", "242584", "242585", "242586", "242587", "242588", "242589", "242590", "242591", "242592", "242593", "242594", "242595", "242596", "242597", "242598", "242599", "242600", "242601", "242602", "242603", "242604", "242605", "242606", "242607", "242608", "242609", "242610", "242611", "242612", "242613", "242614", "242615", "242616", "242617", "242618", "242619", "242620", "242621", "242622", "242623", "242624", "242625", "242626", "242627", "242628", "242629", "242630", "242631", "242632", "242633", "242634", "242635", "242636", "242671", "242672", "242794", "242795", "242796", "242808", "242827", "242828", "242836", "242867", "242869", "242870", "242872", "242876", "242887", "242897", "242917", "242918", "242919", "242920", "243026", "243032", "243036", "243037", "243040", "243043", "243047", "243050", "243080", "243146", "243147",

"243148", "243149", "243150", "243151", "243152", "243153", "243154", "243155",  
"243156", "243157", "243158", "243159", "10002", "10004", "10019", "10020",  
"10025", "10048", "10054", "10055", "10061", "10152", "10157", "10163", "10165",  
"10186", "10187", "10188", "10211", "10225", "10227", "10246", "10250", "10271",  
"10366", "10370", "10371", "10375", "10381", "10385", "10390", "10394", "10395",  
"10404", "10405", "10406", "10408", "10409", "10570", "10731", "10914", "10932",  
"10950", "10995", "10997", "10998", "11000", "11002", "110040", "110070",  
"110097", "110492", "110495", "110498", "110501", "110504", "110507", "110510",  
"110511", "110514", "110515", "110518", "110519", "110520", "110521", "110523",  
"110542", "110545", "11258", "11303", "113835", "113869", "113875", "113894",  
"113926", "113935", "113941", "113972", "113973", "114021", "114223", "114226",  
"114235", "114243", "114290", "114291", "114606", "114896", "114899", "11490",  
"114902", "114903", "114908", "114909", "114913", "114930", "114931", "114935",  
"114937", "114940", "114944", "114945", "114946", "114948", "114958", "114973",  
"114975", "114977", "114983", "114987", "114991", "115002", "115089", "115234",  
"115239", "115240", "11550", "115552", "115847", "116097", "116274", "116292",  
"116356", "116357", "116358", "116359", "116360", "116361", "116362", "116363",  
"116364", "116365", "116368", "116370", "116374", "116375", "116377", "116378",  
"116379", "116380", "116381", "116397", "116573", "116575", "116578", "116580",  
"116581", "116588", "116591", "116594", "116597", "116600", "116608", "116632",  
"116639", "116674", "116944", "117020", "117125", "117133", "117134", "117135",  
"117137", "117138", "117139", "117140", "117143", "117144", "117146", "117148",  
"117149", "117154", "117158", "117166", "117173", "117175", "117188", "117232",  
"117258", "117261", "117433", "117512", "117722", "117850", "117882", "117887",  
"117888", "117894", "117914", "117915", "117917", "117918", "117943", "118022",  
"118029", "118031", "118035", "118036", "118045", "119112", "11923", "11924",  
"11926", "119276", "119278", "119279", "11928", "119281", "119282", "119319",  
"119320", "119322", "119323", "119348", "119349", "119351", "119353", "119357",  
"119358", "119359", "119360", "119365", "119367", "119368", "119369", "119370",  
"119495", "119534", "119570", "119571", "119572", "119573", "119574", "119575",  
"119692", "119693", "119694", "119695", "119697", "119698", "119699", "11971",  
"119754", "119761", "119767", "119770", "119838", "119839", "119846", "119847",  
"119932", "120014", "120015", "120022", "120039", "120092", "120127", "120131",  
"120132", "120133", "120135", "120137", "120138", "120141", "120145", "120146",

"120149", "120152", "120324", "120389", "120402", "120411", "12053", "12065",  
"12075", "12152", "12155", "12156", "12157", "12181", "12216", "12218", "12221",  
"12224", "12225", "12236", "123275", "123276", "123277", "123278", "12333",  
"123378", "123385", "123386", "123388", "123389", "123392", "123434", "123436",  
"123437", "123438", "123473", "123481", "12450", "12472", "12473", "12478",  
"12480", "12482", "12483", "12493", "12498", "12500", "12501", "12554", "12564",  
"12684", "12778", "13119", "13174", "13176", "13179", "13182", "13184", "13186",  
"13188", "13189", "13191", "13193", "13231", "13234", "13237", "13238", "13239",  
"13240", "13241", "13252", "13275", "13299", "13431", "13617", "13678", "13730",  
"13820", "13900", "13902", "13928", "13932", "13934", "13949", "13951", "13957",  
"13958", "13964", "13989", "13994", "14012", "14013", "14015", "14036", "14058",  
"14062", "14066", "14067", "14071", "14073", "14079", "14080", "14127", "14241",  
"14242", "14243", "14332", "14333", "14359", "14376", "14431", "14432", "14433",  
"14434", "14435", "14442", "14443", "14444", "14445", "14446", "14447", "14462",  
"14523", "14534", "14535", "14537", "14538", "14540", "14542", "14548", "14550",  
"14552", "14568", "14597", "14626", "14649", "14671", "14674", "14679", "14689",  
"14690", "14691", "14693", "14696", "14793", "14794", "14795", "14796", "14797",  
"14798", "14800", "14804", "14805", "14808", "14809", "14810", "14811", "14812",  
"14813", "14814", "14816", "14819", "14821", "14822", "14825", "14827", "14899",  
"14900", "14901", "14904", "14906", "14908", "14910", "14911", "14912", "15008",  
"15011", "15012", "15025", "15154", "243160", "243161", "243162", "243163",  
"243164", "243165", "243166", "243167", "243168", "243169", "243170", "243171",  
"243172", "243173", "243242", "243682", "243683", "243768", "243774", "244078",  
"244462", "244470", "244508", "244509", "244510", "244511", "244512", "244513",  
"244514", "244515", "244516", "244519", "244521", "244524", "244617", "246454",  
"246457", "246458", "246459", "246460", "246462", "246463", "246465", "246467",  
"246476", "246477", "246478", "246480", "246481", "246483", "246485", "246486",  
"246488", "246489", "246490", "246503", "246552", "251979", "251981", "251983",  
"251984", "252410", "252421", "252529", "254114", "254116", "254118", "254120",  
"254131", "254132", "254133", "254134", "254135", "254138", "254140", "254142",  
"254151", "255491", "255992", "255993", "255994", "255995", "255996", "255997",  
"255998", "255999", "256000", "256001", "256002", "256003", "256004", "259621",  
"260177", "260655", "260734", "260736", "261674", "261675", "261676", "261678",  
"261679", "261680", "261681", "261682", "261694", "261695", "261696", "261697",

"261714", "261734", "264415", "265017", "265018", "337076", "337080", "7192",  
"7254", "7506", "7524", "7525", "7526", "7528", "7530", "7802", "7810", "7811",  
"7816", "7817", "7829", "7835", "7865", "7899", "8014", "8055", "8056", "8347",  
"8437", "8664", "8665", "8679", "8680", "8782", "8823", "8887", "9014", "9059",  
"9071", "9241", "9259", "9285", "9339", "9550", "9706", "9707", "9710", "9717",  
"9898", "9899", "9977", "9981", "9982", "9984", "9985", "9986", "9988", "9990",  
"9997", "9999")

AND SUBFIELD\_DESC IN ("Adult Learning", "Building Construction", "Civil Engineering  
Construction", "Engineering and Related Design", "Environmental Sciences", "Fabrication and  
Extraction", "Finance, Economics and Accounting", "Generic Management", "Hospitality,  
Tourism, Travel, Gaming and Leisure", "Human Resources", "Manufacturing and Assembly",  
"Marketing", "Office Administration", "Public Administration", "Transport, Operations and  
Logistics")

AND PROV\_ETQE\_ID IN ("1031", "1075", "1102", "1103", "1109", "1110", "1111", "1114",  
"1115", "1118", "1125", "1126", "1127", "NULL")

AND PROV\_ACCRED\_IND\_DESC IN ("Start After, End After")

AND UNIT\_STD\_TYPE\_DESC IN ("Regular")

AND FIELD\_DESC IN ("Business, Commerce and Management Studies",  
"Manufacturing, Engineering and Technology", "Physical Planning and  
Construction", "Physical, Mathematical, Computer and Life Sciences", "Services")

AND ETQE\_ID IN ("1075", "1102", "1103", "1109", "1110", "1111", "1114",  
"1115", "1125", "1127")

AND ENROL\_TYPE\_DESC IN ("Mixed Mode", "Work Place Learning")

- Cluster 3

% of records: 12.68%

Average probability: 0.9749

Rule:

ASSESSOR\_ID IN ("NULL")

AND PROVIDER\_ID IN ("11772", "12489", "12507", "12666", "12888", "12894",  
"13128", "14450", "14928", "15138", "15442", "22687", "26342", "32919", "37392",  
"37396", "39897", "44779")

AND UNIT\_STANDARD\_ID IN ("10028", "10029", "10030", "10031", "10033",  
"10034", "10035", "10036", "10037", "10038", "10039", "10040", "10041", "10042",  
"10043", "10044", "10054", "10055", "10152", "10187", "10341", "10365", "10366",  
"10367", "10370", "10371", "10375", "110016", "110020", "110026", "110038",

"110040", "110043", "11252", "11258", "114600", "114601", "114602", "114603",  
"114604", "114605", "114606", "114607", "114608", "114609", "114611", "114612",  
"114613", "114615", "114617", "114624", "114635", "119474", "119476", "119479",  
"119486", "119489", "12170", "12198", "12434", "12461", "13435", "13437",  
"13889", "13890", "13891", "13900", "13901", "13902", "13903", "13929", "13931",  
"13932", "13933", "13934", "13935", "13945", "13946", "13947", "13948", "13949",  
"13950", "13951", "13952", "13953", "13954", "13957", "13958", "13960", "13962",  
"13964", "13965", "14356", "14357", "14358", "14359", "14360", "14361", "14363",  
"14365", "14366", "14367", "14368", "14369", "14376", "14569", "14684", "15076",  
"15106", "15231", "15251", "242601", "242610", "242836", "246750", "246751",  
"246752", "246754", "246755", "246756", "7194", "7464", "7466", "7468", "7473",  
"7478", "7480", "7485", "7486", "7497", "7564", "7583", "7584", "7585", "7587",  
"7588", "7590", "7723", "7802", "7807", "7808", "7813", "7877", "8121", "8437",  
"8635", "8979", "8980", "8981", "8982", "8983", "8984", "8985", "8986", "8987",  
"8989", "8990", "8991", "8992", "8993", "9024", "9025", "9026", "9027", "9029",  
"9030", "9032", "9033", "9261", "9550", "9943", "9977")

AND PROV\_ETQE\_ID IN ("1126")

AND ETQE\_ID IN ("1126")

AND ENROL\_TYPE\_DESC IN ("RPL for Unknown Purpose", "Work Place Learning")

AND SUBFIELD\_DESC IN ("Finance, Economics and Accounting", "Generic Management",  
"Hospitality, Tourism, Travel, Gaming and Leisure", "Information Technology and Computer  
Sciences", "Language", "Marketing", "Mathematical Sciences", "Office Administration")

AND FIELD\_DESC IN ("Business, Commerce and Management Studies", "Communication  
Studies and Language", "Physical, Mathematical, Computer and Life Sciences", "Services")

AND PROV\_ACCRED\_IND\_DESC IN ("Start After, End After")

AND 0 <= START\_PROV\_ACCRED\_IND <= 84.6

- Cluster 4

% of records: 9.37%

Average probability: 0.9797

Rule:

ASSESSOR\_ID IN ("3483192", "4356252", "4688803", "NULL")

AND PROVIDER\_ID IN ("21879", "25207", "25317", "25387", "25389", "25391",  
"29674", "29679", "29700", "29724", "31953", "36461", "37827", "37838", "37849",  
"41587", "41606", "50873")

*AND UNIT\_STANDARD\_ID IN ("110039", "115118", "116070", "116074", "116077", "116081", "116082", "116083", "116086", "116087", "116089", "116093", "116094", "116096", "116097", "116098", "116100", "116126", "116128", "116130", "116131", "116132", "116136", "116138", "116139", "116141", "116142", "116143", "116144", "116145", "116165", "116166", "116167", "116170", "116173", "116174", "116175", "116176", "116177", "116178", "116180", "116181", "116182", "116183", "116184", "116185", "116186", "116189", "116191", "116194", "116207", "116208", "116214", "116215", "116216", "116217", "116218", "116219", "116220", "116221", "116222", "116223", "116224", "116225", "116226", "116544", "116655", "116660", "116837", "117093", "119423", "119440", "119465", "119703", "119704", "119705", "119707", "119708", "119709", "119710", "119711", "119712", "119713", "119714", "119715", "119716", "119717", "119718", "119719", "119720", "119721", "119722", "119723", "119724", "119725", "119726", "119727", "119728", "119731", "12220", "12479", "12486", "12487", "12488", "13193", "13373", "14012", "14101", "14684", "7464", "7466", "7468", "7478", "7480", "7481", "8510", "8979", "8980", "8981", "8984", "9616")*

*AND PROV\_ETQE\_ID IN ("1112")*

*AND ETQE\_ID IN ("1112")*

*AND SUBFIELD\_DESC IN ("Horticulture", "Language", "Mathematical Sciences", "Primary Agriculture", "Secondary Agriculture")*

*AND ENROL\_TYPE\_DESC IN ("Mixed Mode")*

*AND PROVIDER\_TYPE\_DESC IN ("Training")*

*AND PROVIDER\_CLASS\_DESC IN ("Private")*

*AND NQF\_LEVEL\_DESC IN ("Level 1", "Level 2")*

- Cluster 5

% of records: 8.19%

Average probability: 0.9559

Rule:

*UNIT\_STANDARD\_ID IN ("114958", "115808", "119471", "119473", "119474", "119476", "119477", "119479", "119480", "119482", "119484", "119486", "119488", "119489", "12170", "8979", "8980", "8981", "8984", "8985", "8986", "8987", "8990", "8991", "8992", "8993")*

*AND ASSESSOR\_ID IN ("16123888", "17162273", "17374751", "17374856", "3013505", "3014866", "3017504", "3039765", "3557298", "4356252", "4631948", "4632161", "4632223", "5663537", "5663618", "6094669", "7309248", "7309500", "7339985", "9015258", "9050705", "9382710", "9463777", "9464322", "9464465", "9502453", "NULL")*

*AND SUBFIELD\_DESC IN ("Language")*

*AND PROVIDER\_ID IN ("11244", "12489", "13654", "1578", "16886", "17119", "1726", "18758", "1915", "19858", "20357", "2066", "20689", "20997", "21121", "21318", "21375", "21389", "21464", "2158", "21879", "2224", "22687", "22883", "25426", "26342", "26380", "27074", "27090", "27104", "28144", "28156", "28413", "28473", "28475", "28531", "28710", "28833", "29238", "29282", "29298", "29679", "29782", "29935", "29937", "29963", "31953", "32014", "32040", "32072", "32098", "32209", "32284", "32285", "32286", "33654", "34485", "34617", "35405", "35551", "35920", "35940", "36643", "36682", "37425", "37488", "37583", "37838", "37849", "38377", "38563", "39612", "39637", "39665", "39998", "41493", "41540", "41587", "42861", "42887", "43319", "43365", "44003", "44010", "44087", "44091", "44092", "44747", "44779", "44880", "46423", "46514", "48559", "49420", "49581", "49792", "50114", "51338", "51339")*

*AND FIELD\_DESC IN ("Communication Studies and Language")*

*AND UNIT\_STD\_TYPE\_DESC IN ("Regular-Fundamental")*

*AND NQF\_LEVEL\_DESC IN ("Level 2", "Level 3", "Level 4")*

*AND ENROL\_TYPE\_DESC IN ("Mixed Mode", "Residential Learning (i.e. Contact Mode)")*

*AND 1 <= END\_PROV\_ACCRED\_IND <= 54.7*

*AND PROVIDER\_CLASS\_DESC IN ("Mixed: Public and Private", "Private", "Public", "Unknown")*

- Cluster 6

% of records: 8.06%

Average probability: 0.8891

Rule:

*PROVIDER\_ID IN ("10053", "10377", "11091", "11094", "11244", "11691", "16886", "20357", "20589", "20689", "21318", "21328", "21375", "21389", "21460", "21464", "2169", "2172", "2183", "21912", "2206", "2220", "2224", "2251", "22831", "22953", "24993", "26995", "27255", "28046", "28049", "28052", "28099", "28133",*

"28144", "28156", "28186", "28244", "28276", "28282", "28293", "28349", "28413",  
"28434", "28435", "28449", "28473", "28475", "28514", "28531", "28710", "28868",  
"29935", "29937", "29963", "32697", "32705", "32728", "34138", "34432", "34485",  
"34574", "35606", "35614", "35628", "35671", "36340", "37334", "37396", "37425",  
"38223", "38597", "38989", "39013", "39998", "42861", "42886", "42887", "42900",  
"44320", "44747", "44769", "44779", "46423", "47894", "48508", "48855", "49581",  
"49633", "50114", "50130", "51304")

AND ASSESSOR\_ID IN ("16123887", "16123888", "17151625", "17374783",  
"18341625", "18341737", "3007815", "3008370", "3008822", "3008834", "3014866",  
"3015639", "3015708", "3018856", "3020705", "3020718", "3021695", "3024160",  
"3024757", "3029008", "3029904", "3030035", "3030928", "3031515", "3039765",  
"3059520", "3300882", "3312964", "3313013", "3313225", "3313372", "3511030",  
"3516049", "3557137", "3557288", "3557298", "3588732", "4282605", "4282924",  
"4632014", "4632161", "4632223", "5013309", "5013724", "5518282", "5648404",  
"5664538", "6055407", "7308990", "7309067", "7309248", "7309500", "7309531",  
"9050524", "9050705", "9463777", "9463856", "9464078", "9464322", "9464465",  
"9574927", "9575168", "9575171", "NULL")

AND UNIT\_STANDARD\_ID IN ("10023", "10024", "10025", "10048", "10054",  
"10055", "10058", "10061", "10075", "10165", "10186", "10187", "10188", "10305",  
"10306", "10311", "10322", "10403", "10568", "10570", "10696", "10701", "10702",  
"10995", "10997", "10998", "11000", "11002", "110135", "110139", "110144",  
"110545", "11303", "113983", "114234", "114243", "114290", "114291", "114383",  
"114609", "114613", "114896", "114899", "114902", "114903", "114909", "114913",  
"114941", "114942", "114955", "114958", "114959", "115109", "115122", "115770",  
"115772", "115776", "115806", "115807", "115847", "116184", "116217", "116219",  
"116220", "116221", "116222", "116454", "116501", "116604", "116653", "116737",  
"116944", "117512", "117722", "117882", "117887", "117888", "117891", "117894",  
"119112", "119136", "11923", "11924", "11926", "11928", "119379", "119390",  
"119580", "119678", "119682", "119761", "119814", "119930", "120053", "120252",  
"120317", "120420", "120421", "120427", "120433", "120434", "12152", "12155",  
"12156", "12157", "12220", "12236", "12256", "12257", "12258", "12260", "12263",  
"12332", "12333", "123374", "123376", "123377", "123378", "123384", "123388",  
"123390", "123392", "123411", "123413", "123414", "123415", "123417", "123418",  
"12446", "12472", "12473", "12474", "12478", "12480", "12482", "12483", "12493",

"12498", "12500", "12501", "12542", "12859", "13014", "13016", "13017", "13174",  
"13179", "13184", "13186", "13188", "13189", "13191", "13193", "13231", "13234",  
"13237", "13238", "13239", "13240", "13251", "13660", "13870", "13871", "13872",  
"13873", "13928", "13934", "13942", "14015", "14113", "14127", "14568", "14586",  
"14599", "14673", "14676", "14678", "14681", "14699", "14700", "14729", "14730",  
"14739", "14800", "14805", "14809", "14822", "14825", "14827", "15071", "15078",  
"15081", "15085", "15088", "15108", "15109", "15130", "15231", "15232", "15234",  
"15235", "15237", "15238", "15244", "15245", "15246", "15247", "15249", "15254",  
"15282", "242794", "242795", "242796", "242797", "242805", "242808", "242809",  
"242829", "242833", "242836", "243026", "243027", "243032", "243034", "243036",  
"243037", "243039", "243040", "243043", "243047", "243048", "243050", "243073",  
"243080", "243081", "243083", "243084", "243085", "243086", "243089", "243092",  
"243093", "244272", "244273", "244274", "244275", "244276", "244277", "244278",  
"244279", "244280", "244359", "244362", "244450", "244462", "244470", "244479",  
"244485", "244486", "244489", "244492", "244495", "244497", "244498", "244501",  
"244502", "244588", "244627", "246750", "246751", "246752", "246753", "246754",  
"246755", "246756", "252058", "256616", "260655", "263993", "264996", "7417",  
"7418", "7420", "7425", "7426", "7427", "7467", "7485", "7626", "7865", "7895",  
"7899", "8664", "8782", "8783", "8820", "8822", "8823", "8824", "8887", "8922",  
"8923", "8940", "8941", "8959", "8960", "8961", "8962", "8963", "9032", "9033",  
"9059", "9063", "9068", "9069", "9070", "9071", "9079", "9080", "9122", "9128",  
"9139", "9142", "9148", "9153", "9285", "9339", "9460", "9543", "9545", "9547",  
"9550", "9706", "9710", "9717", "9895", "9990")

AND PROVIDER\_CLASS\_DESC IN ("NULL", "Public", "Unknown")

AND FIELD\_DESC IN ("Business, Commerce and Management Studies", "Education,  
Training and Development", "Manufacturing, Engineering and Technology",  
"Services")

AND SUBFIELD\_DESC IN ("Adult Learning", "Cleaning, Domestic, Hiring, Property and  
Rescue Services", "Early Childhood Development", "Engineering and Related Design",  
"Fabrication and Extraction", "Generic Management", "Higher Education and Training",  
"Human Resources", "Manufacturing and Assembly", "Mathematical Sciences",  
"People/Human-Centred Development")

AND ETQE\_ID IN ("1075", "1103", "1106", "1107", "1111", "1126")

AND PROV\_ETQE\_ID IN ("1033", "1103", "1106", "1107", "1126", "NULL")

AND PROVIDER\_TYPE\_DESC IN ("Education", "Education and Training", "Employer",  
"NULL")

AND ENROL\_STATUS\_DESC IN ("Achieved")

- Cluster 7

% of records: 7.80%

Average probability: 0.9455

Rule:

ASSESSOR\_ID IN ("16083302", "17162273", "4707504", "4707513", "4783541",  
"4783577", "5664357", "5664448", "5664477", "6094669", "7339985", "7339986",  
"7354826", "7355812", "7355866", "7363810", "7363870", "8218000", "9012063",  
"9015258", "9046687", "9382710", "9502453", "NULL")

AND PROVIDER\_ID IN ("11343", "12223", "1578", "1726", "1898", "1915",  
"20933", "2153", "22680", "22708", "22883", "24841", "25366", "27063", "27074",  
"27090", "27093", "27104", "28833", "29272", "29273", "29282", "29298", "29299",  
"29781", "29782", "32014", "32017", "32040", "32098", "32151", "32209", "34617",  
"34674", "35920", "35940", "36061", "36643", "36682", "37206", "38563", "38608",  
"38637", "39612", "39675", "40939", "41386", "41493", "41540", "43923", "43989",  
"44003", "44010", "44065", "44070", "44087", "44091", "44092", "44114", "44660",  
"44880", "46423", "48559", "49173", "49808", "51339", "52070", "663")

AND UNIT\_STANDARD\_ID IN ("9706", "9861", "9862", "9864", "9865", "9870",  
"9872", "9875", "9876", "9891", "9892", "9899", "9977", "9979", "9981", "9982",  
"9983", "9984", "9985", "9986", "9987", "9988", "9990", "9999", "10001", "10003",  
"10017", "10018", "10019", "10024", "10039", "10048", "10061", "10163", "10165",  
"10186", "10187", "10188", "10225", "10227", "10246", "10251", "10269", "10366",  
"10370", "10371", "10375", "10405", "10584", "10604", "10615", "10630", "10731",  
"10735", "10995", "10997", "10998", "11000", "11002", "110026", "110092",  
"110099", "110112", "110489", "110490", "110492", "110495", "110496", "110498",  
"110501", "110507", "110510", "110514", "110518", "110519", "110520", "110521",  
"110523", "110542", "110545", "11303", "113835", "113894", "113920", "113921",  
"113935", "113941", "113977", "114235", "114243", "114290", "114291", "114480",  
"114495", "114508", "114615", "114890", "114895", "114896", "114899", "11490",  
"114903", "114917", "114923", "114924", "114928", "114929", "114953", "114958",  
"114960", "114974", "114975", "114977", "114981", "114987", "114990", "114991",

"114995", "114996", "115000", "115002", "115233", "115234", "115239", "115240",  
"115241", "115840", "115847", "115872", "116081", "116094", "116097", "116138",  
"116356", "116357", "116358", "116359", "116360", "116361", "116362", "116363",  
"116364", "116365", "116368", "116370", "116374", "116375", "116377", "116378",  
"116379", "116380", "116381", "116513", "116528", "116590", "116674", "116687",  
"116944", "116945", "117020", "117128", "117134", "117138", "117143", "117144",  
"117146", "117149", "117150", "117151", "117163", "117166", "117173", "117175",  
"117188", "117258", "117261", "117407", "117512", "117524", "117775", "117887",  
"117888", "117894", "117900", "11833", "11835", "119112", "11923", "11924",  
"11926", "119277", "11928", "119282", "119319", "119320", "119321", "119322",  
"119323", "119348", "119351", "119353", "119358", "119359", "119360", "119362",  
"119365", "119367", "119368", "119369", "119370", "119534", "119570", "119571",  
"119572", "119573", "119574", "119575", "119693", "119729", "119738", "119930",  
"119971", "120014", "120022", "120028", "120031", "120032", "120033", "120034",  
"120036", "120039", "120044", "120092", "120127", "120135", "120138", "120140",  
"120141", "120144", "120145", "120149", "120153", "120365", "120389", "120402",  
"120406", "120407", "120408", "120409", "120410", "120411", "12050", "12053",  
"12074", "12075", "12152", "12155", "12156", "12157", "12181", "123276",  
"123411", "123453", "123472", "123473", "123475", "123476", "123477", "123479",  
"123481", "12351", "12363", "12368", "12369", "12450", "12478", "12480",  
"12482", "12483", "12500", "12501", "12554", "12778", "12917", "12920", "12925",  
"13119", "13184", "13186", "13219", "13234", "13237", "13238", "13241", "13275",  
"13696", "13900", "13902", "13929", "13932", "13942", "13957", "13958", "13962",  
"13965", "13975", "13979", "13980", "13989", "13990", "14012", "14013", "14015",  
"14053", "14055", "14068", "14071", "14079", "14152", "14353", "14431", "14432",  
"14433", "14434", "14435", "14442", "14443", "14445", "14446", "14447", "14462",  
"14498", "14523", "14551", "14568", "14572", "14578", "14586", "14592", "14597",  
"14626", "14637", "14649", "14659", "14667", "14671", "14677", "14678", "14679",  
"14681", "14689", "14690", "14691", "14797", "14801", "14899", "14900", "14901",  
"14902", "14904", "14905", "14906", "14907", "14908", "14909", "14910", "14911",  
"14912", "15008", "15051", "15154", "15176", "15199", "15231", "15232", "15234",  
"15237", "15238", "15244", "15245", "15246", "15247", "15249", "15282",  
"230087", "230088", "230092", "230094", "230095", "242571", "242583", "242590",  
"242591", "242597", "242598", "242601", "242602", "242606", "242610", "242611",

"242618", "242620", "242621", "242630", "242631", "242672", "242827", "242828",  
 "242838", "242867", "242868", "242869", "242870", "242872", "242875", "242876",  
 "242877", "242883", "242887", "242910", "242911", "242920", "243150", "243161",  
 "243223", "243697", "244078", "244192", "244200", "244402", "244433", "244462",  
 "244509", "244510", "244511", "244512", "244513", "244514", "244515", "244516",  
 "244519", "244524", "244606", "244617", "244628", "246454", "246457", "246458",  
 "246459", "246460", "246462", "246463", "246465", "246467", "246476", "246477",  
 "246478", "246480", "246481", "246483", "246485", "246486", "246488", "246489",  
 "246490", "246503", "246552", "246757", "251978", "252054", "252421", "252529",  
 "252530", "252549", "252550", "254114", "254116", "254120", "254131", "254132",  
 "254134", "254135", "254138", "254140", "254142", "254239", "255491", "255992",  
 "255993", "255994", "255995", "255996", "255997", "255998", "255999", "256000",  
 "256001", "256002", "256003", "256004", "256495", "256496", "256499", "256501",  
 "256513", "256574", "259621", "260614", "261674", "261675", "261676", "261678",  
 "261680", "261681", "261754", "264277", "337077", "337080", "7369", "7506",  
 "7524", "7525", "7526", "7528", "7530", "7676", "7677", "7678", "7680", "7765",  
 "7779", "7803", "7808", "7810", "7811", "7816", "7817", "7819", "7822", "7825",  
 "7826", "7827", "7829", "7835", "7865", "7899", "8056", "8347", "8349", "8363",  
 "8365", "8511", "8572", "8635", "8664", "8665", "8680", "9241", "9259", "9285")

AND PROVIDER\_CLASS\_DESC IN ("Private", "Public")

AND SUBFIELD\_DESC IN ("Adult Learning", "Building Construction", "Civil Engineering  
 Construction", "Cleaning, Domestic, Hiring, Property and Rescue Services", "Curative  
 Health", "Environmental Sciences", "Fabrication and Extraction", "Finance, Economics and  
 Accounting", "Generic Management", "Hospitality, Tourism, Travel, Gaming and Leisure",  
 "Human Resources", "Manufacturing and Assembly", "Marketing", "People/Human-Centred  
 Development", "Preventive Health", "Public Administration", "Wholesale and Retail")

AND 0 <= START\_PROV\_ACCRED\_IND <= 14.1

AND FIELD\_DESC IN ("Business, Commerce and Management Studies", "Manufacturing,  
 Engineering and Technology", "Physical Planning and Construction", "Services")

AND ETQE\_ID IN ("1075", "1100", "1103", "1109", "1110", "1111", "1122", "1125", "1127")

AND PROV\_PROVINCE\_DESC IN ("Eastern Cape", "Gauteng", "Kwazulu/Natal",  
 "Limpopo", "Mpumalanga", "North West", "Undefined", "Western Cape")

AND 1 <= END\_PROV\_ACCRED\_IND <= 36.8

- Cluster 8

% of records: 6.38%

Average probability: 0.9488

Rule:

*UNIT\_STANDARD\_ID* IN ("116551", "116947", "116948", "116949", "116952", "116954", "116957", "116959", "116960", "116962", "117940", "117941", "117942", "119095", "12434", "12461", "13951", "13962", "14101", "242827", "7464", "7465", "7466", "7467", "7468", "7473", "7474", "7478", "7480", "7481", "7485", "7486", "7497", "7564", "7585", "7587", "7589", "8511", "9024", "9025", "9026", "9027", "9029", "9030", "9032", "9033", "9549")

*AND ASSESSOR\_ID* IN ("17162273", "3013505", "3017504", "3018150", "3557298", "4356252", "4632223", "4688803", "4707504", "4707513", "5663537", "5663618", "6094669", "7339985", "7354133", "7354188", "9012073", "9015258", "9050705", "9382825", "9502453", "NULL")

*AND SUBFIELD\_DESC* IN ("Information Technology and Computer Sciences", "Mathematical Sciences")

*AND PROVIDER\_ID* IN ("10053", "10377", "11087", "11088", "11090", "11098", "11244", "11343", "12489", "12666", "12888", "13248", "14795", "1578", "16886", "1696", "17119", "1726", "1915", "19858", "20357", "20574", "20689", "20725", "21121", "21328", "21337", "21375", "21464", "2158", "21879", "2206", "2224", "2251", "22687", "22708", "25207", "25317", "25344", "25389", "25391", "25426", "2622", "26986", "27063", "27090", "27093", "27104", "28473", "28531", "28710", "29238", "29260", "29273", "29282", "29298", "29299", "29679", "29750", "29782", "29814", "29980", "30104", "30204", "31726", "31909", "31953", "32014", "32040", "32044", "32098", "32209", "32284", "32285", "32807", "34485", "34617", "35744", "35920", "35940", "36466", "36682", "37203", "37206", "37425", "37827", "37838", "37849", "37887", "38548", "40942", "40960", "41493", "41540", "41587", "41783", "41856", "41914", "41932", "41941", "42072", "42244", "42351", "42373", "42403", "42861", "43365", "44003", "44010", "44087", "44092", "46423", "46514", "47336", "48508", "48559", "49420", "49581", "49792", "50114", "51212", "51245", "51338", "51339", "5987", "5994")

*AND FIELD\_DESC* IN ("Physical, Mathematical, Computer and Life Sciences")

*AND UNIT\_STD\_TYPE\_DESC* IN ("Regular-Fundamental")

*AND ENROL\_TYPE\_DESC* IN ("Mixed Mode", "Work Place Learning")

*AND PROV\_PROVINCE\_DESC* IN ("Eastern Cape", "Gauteng", "Kwazulu/Natal", "Limpopo", "Mpumalanga", "South Africa National", "Western Cape")

*AND ETQE\_ID IN ("1075", "1103", "1105", "1106", "1107", "1109", "1110", "1112", "1114",  
"1117", "1123", "1125", "1127")*

*AND 1 <= END\_PROV\_ACCRED\_IND <= 54.7*

## Appendix N

This appendix provides a technical description of the outputs of data mining activities that were conducted when analysing whether the intrinsic relationship between the completion of a learnership and achievement of its related qualification has been upheld. The data mining activities focuses on gaining a better understanding of data records that fall into specific categories of the data field QENROL\_IND (see Section 4.7.1) and the possible identification of anomalous data records in the respective data sets.

This semantic business rule defines that the intrinsic relationship between the completion of a learnership and achievement of its related qualification must be upheld and is applicable to learnership enrolment records only.

### *N.1 No Qual Enrolment*

This section provides a technical description of the clusters that were generated by cluster data mining the data category ‘No Qual Enrolment’ (see Section 4.8.1) for learnership enrolment records. A detailed description of the manner in which cluster data mining was conducted can be found in Appendix I.3.

The results of the generated clustering model are significant because the model was measured as being 97.29% accurate. All of the clusters show a tight coupling between data fields that describe the ETQEs, learnerships and providers. This is as a result of the organic relationship between learnerships and ETQEs (learnerships are generally implemented by one ETQE only) and providers and ETQEs (providers generally offer learnerships that are implemented by their primary ETQE).

The model did not generate any clusters that contained less than 1% of the records. The detection of anomalous records based on cluster specific probabilities of less than .6 being allocated to the record was conducted. In this manner it was determined that 1.36% of the records possibly exist in this category as a result of data capturing problems at the source of the data.

The technical description of each of the 8 clusters is provided below. The description is limited to the top 10 attributes of the cluster rule, where each indent of text represents the importance of the attribute. For example:

A line formatted like this represents an importance of 100%

*A line formatted like this represents an importance greater than 75% and less than 100%*

A line formatted like this represents an importance greater than 50% and less than or equal to 75%

*A line formatted like this represents an importance less than or equal to 50%*

1. Cluster 1

% of records: 35.29%

Average probability: 0.9997

Rule:

PROVIDER\_ID IN ('47993', '47992')

*AND LEARNERSHIP\_ID IN ('1554')*

*AND ETQE\_ID IN ('1105')*

*AND LSHP\_ETQE\_ID IN ('1105')*

AND NQF\_LEVEL\_DESC IN ('Level 4')

*AND 56.8 <= END\_DATE\_IND <= 115*

*AND 54.3 <= START\_DATE\_IND <= 102.6*

*AND ENROL\_STATUS\_DESC IN ('Achieved')*

2. Cluster 2

% of records: 18.99%

Average probability: 0.9732

Rule:

*LEARNERSHIP\_ID IN ('884', '880', '878')*

*AND PROVIDER\_ID IN ('49153', '41226', '39919', '39874', '37405', '26386', '25367', '22771', '1575', '14795', '14658', '12894', '12310', '11772')*

*AND ETQE\_ID IN ('1126')*

*AND LSHP\_ETQE\_ID IN ('1126')*

AND NQF\_LEVEL\_DESC IN ('Level 4')

*AND 38.2 <= START\_DATE\_IND <= 134.8*

*AND 37.4 <= END\_DATE\_IND <= 153.8*

*AND ENROL\_STATUS\_DESC IN ('Enrolled', 'Achieved')*

3. Cluster 3

% of records: 12.04%

Average probability: 0.8871

Rule:

```
LEARNERSHIP_ID IN ('900', '893', '883', '822', '795', '643', '519', '462', '461', '456',
'440', '405', '39', '387', '385', '312', '309', '292', '288', '284', '276', '24', '236', '233',
'231', '230', '22', '196', '1579', '1573', '1572', '1374', '1325', '1097', '1031')
AND ETQE_ID IN ('1127', '1125', '1122', '1117', '1116', '1113', '1109', '1106')
AND PROVIDER_ID IN ('669', '662', '595', '50330', '50323', '50204', '50196', '49754',
'49723', '49720', '49719', '49655', '49530', '49520', '49364', '49255', '48789', '45502',
'44688', '43921', '43246', '43232', '43230', '43195', '41542', '40942', '40939', '40133',
'39735', '38968', '38965', '38570', '38560', '38008', '38007', '37970', '37941', '37939',
'37704', '37405', '37062', '35341', '35015', '33692', '32268', '32266', '32256', '32014',
'29795', '29782', '29238', '29198', '29196', '28255', '28186', '28163', '28100', '28049',
'28030', '27161', '27135', '27128', '27125', '27114', '27098', '27090', '27074', '27070',
'27059', '27005', '26941', '26365', '26273', '25387', '24857', '24803', '24798', '24778',
'24712', '24642', '24641', '24630', '2335', '22910', '22745', '21879', '2160', '21318',
'1913', '18758', '16388', '1575', '26510')
AND LSHP_ETQE_ID IN ('1126', '1125', '1122', '1117', '1116', '1113', '1109', '1106')
AND NQF_LEVEL_DESC IN ('Level 7', 'Level 6', 'Level 4', 'Level 3', 'Level 1')
AND 56.8 <= END_DATE_IND <= 212
AND 38.2 <= START_DATE_IND <= 167
AND ENROL_STATUS_DESC IN ('Enrolled', 'Achieved')
```

#### 4. Cluster 4

% of records: 11.83%

Average probability: 0.9897

Rule:

```
LEARNERSHIP_ID IN ('483', '474')
AND PROVIDER_ID IN ('641', '45502', '42591', '42411', '42335', '42312', '42308',
'42279', '42217', '42194', '42163', '42066', '42058', '42016', '41965', '41926', '24977',
'23286', '18518', '18511', '14640')
AND ETQE_ID IN ('1123')
AND LSHP_ETQE_ID IN ('1104')
AND 56.8 <= END_DATE_IND <= 192.6
AND NQF_LEVEL_DESC IN ('Level 5', 'Level 3')
AND 54.3 <= START_DATE_IND <= 167
AND ENROL_STATUS_DESC IN ('Enrolled', 'Achieved')
```

5. Cluster 5

% of records: 6.51%

Average probability: 0.9962

Rule:

LEARNERSHIP\_ID IN ('364')  
AND PROVIDER\_ID IN ('2224', '2222', '2168')  
AND ETQE\_ID IN ('1107')  
AND LSHP\_ETQE\_ID IN ('1107')  
AND NQF\_LEVEL\_DESC IN ('Level 2')  
AND 22.1 <= START\_DATE\_IND <= 86.5  
AND 37.4 <= END\_DATE\_IND <= 95.6  
AND ENROL\_STATUS\_DESC IN ('Enrolled', 'Discontinued', 'Achieved')

6. Cluster 6

% of records: 5.86%

Average probability: 0.9939

Rule:

LEARNERSHIP\_ID IN ('99', '112', '110', '109', '108', '100')  
AND PROVIDER\_ID IN ('41566', '26234', '26233', '26232', '21927', '11101', '11098',  
'11092', '11091', '11088', '11087', '11082', '11078', '11077', '11059')  
AND ETQE\_ID IN ('1115')  
AND LSHP\_ETQE\_ID IN ('1115')  
AND NQF\_LEVEL\_DESC IN ('Level 6', 'Level 4', 'Level 3', 'Level 1')  
AND ENROL\_STATUS\_DESC IN ('Enrolled')  
AND 22.1 <= START\_DATE\_IND <= 134.8  
AND 37.4 <= END\_DATE\_IND <= 153.8

7. Cluster 7

% of records: 5.52%

Average probability: 0.8831

Rule:

LEARNERSHIP\_ID IN ('959', '894', '893', '888', '803', '771', '770', '715', '714', '670',  
'668', '667', '646', '645', '628', '406', '386', '364', '339', '320', '240', '189', '187', '185',  
'184', '146', '1303', '1069')

*AND PROVIDER\_ID IN ('715', '6297', '5205', '48591', '48370', '43921', '43443', '43426', '43422', '43403', '42946', '42335', '42058', '40939', '39897', '38676', '38665', '38660', '38611', '38608', '38605', '38580', '38570', '38563', '38560', '38554', '37354', '36858', '36831', '35381', '33692', '31949', '31944', '29815', '29321', '28872', '2883', '27111', '27107', '27104', '27098', '27090', '27089', '27086', '27041', '2682', '2620', '25387', '25308', '25217', '25207', '24849', '22910', '22771', '2197', '21780', '21626', '21342', '20584', '1943', '18710', '18695', '16388', '1575', '12894', '12489', '11244', '10968', '10812')*

*AND ETQE\_ID IN ('1126', '1121', '1120', '1114', '1113', '1112', '1111', '1109', '1103')*

*AND LSHP\_ETQE\_ID IN ('1126', '1113', '1111', '1109', '1107', '1103')*

*AND NQF\_LEVEL\_DESC IN ('Level 3', 'Level 2')*

*AND ENROL\_STATUS\_DESC IN ('Enrolled')*

*AND 56.8 <= END\_DATE\_IND <= 153.8*

*AND 38.2 <= START\_DATE\_IND <= 134.8*

## 8. Cluster 8

% of records: 3.96%

Average probability: 0.9898

Rule:

*PROVIDER\_ID IN ('29600', '1907', '1906', '1898', '18758')*

*AND LEARNERSHIP\_ID IN ('80', '79', '77', '76', '75', '74', '72', '70', '69', '56', '55', '54', '53', '52', '49', '44', '43')*

*AND ETQE\_ID IN ('1120')*

*AND LSHP\_ETQE\_ID IN ('1120')*

*AND NQF\_LEVEL\_DESC IN ('Level 6', 'Level 5')*

*AND ENROL\_STATUS\_DESC IN ('Enrolled')*

*AND 22.1 <= START\_DATE\_IND <= 118.7*

*AND 37.4 <= END\_DATE\_IND <= 153.8*

### ***N.2 Lshp Enrolled, Qual Achieved (Derived)***

This section provides a technical description of the clusters that were generated by cluster data mining the data category 'Lshp Enrolled, Qual Achieved (Derived)' (see Section 4.8.2) for learnership enrolment records. A detailed description of the manner in which cluster data mining was conducted can be found in Appendix I.3.

The results of the generated clustering model are significant because the model was measured as being 95.71% accurate. All of the clusters show a tight coupling between data fields that describe the ETQEs and learnerships. This is as a result of the organic relationship between learnerships and ETQEs (learnerships are generally implemented by one ETQE only).

The model did not generate any clusters that contained less than 1% of the records. The detection of anomalous records based on cluster specific probabilities of less than .6 being allocated to the record was conducted. In this manner it was determined that 1.32% of the records possibly exist in this category as a result of data capturing problems at the source of the data.

The technical description of each of the 8 clusters is provided below. The description is limited to the top 10 attributes of the cluster rule, where each indent of text represents the importance of the attribute. For example:

A line formatted like this represents an importance of 100%

*A line formatted like this represents an importance greater than 75% and less than 100%*

*A line formatted like this represents an importance greater than 50% and less than or equal to 75%*

*A line formatted like this represents an importance less than or equal to 50%*

#### 1. Cluster 1

% of records: 33.37%

Average probability: 0.9420

Rule:

*LEARNERSHIP\_ID IN ('96', '951', '945', '944', '943', '939', '899', '895', '884', '60', '55', '528', '523', '513', '461', '460', '39', '292', '284', '231', '1571', '1031')*

*AND NQF\_LEVEL\_DESC IN ('Level 4')*

*AND LSHP\_ETQE\_ID IN ('1127', '1126', '1120', '1119', '1117', '1115', '1109')*

*AND ETQE\_ID IN ('1127', '1126', '1120', '1119', '1117', '1115', '1109')*

*AND 62 <= END\_DATE\_IND <= 158*

*AND 48 <= START\_DATE\_IND <= 138*

#### 2. Cluster 2

% of records: 21.86%

Average probability: 0.9853

Rule:

*LEARNERSHIP\_ID IN ('774', '770', '739', '720', '708', '1450', '1415', '1407', '1273')*  
*AND ETQE\_ID IN ('1103')*  
*AND LSHP\_ETQE\_ID IN ('1103')*  
*AND NQF\_LEVEL\_DESC IN ('Level 4', 'Level 2')*  
*AND 46 <= END\_DATE\_IND <= 142*  
*AND 48 <= START\_DATE\_IND <= 123*

3. Cluster 3

% of records: 17.25%

Average probability: 0.9283

Rule:

*LEARNERSHIP\_ID IN ('98', '961', '959', '95', '93', '894', '893', '771', '740', '709', '518',  
'483', '465', '1327', '1269', '1241', '101')*  
*AND LSHP\_ETQE\_ID IN ('1126', '1115', '1104', '1103', '1102')*  
*AND NQF\_LEVEL\_DESC IN ('Level 3')*  
*AND ETQE\_ID IN ('1127', '1126', '1123', '1115', '1114', '1103', '1102')*  
*AND 62 <= END\_DATE\_IND <= 174*  
*AND 48 <= START\_DATE\_IND <= 153*

4. Cluster 4

% of records: 10.13%

Average probability: 0.9633

Rule:

*LEARNERSHIP\_ID IN ('781', '1529')*  
*AND ETQE\_ID IN ('1105')*  
*AND LSHP\_ETQE\_ID IN ('1005')*  
*AND NQF\_LEVEL\_DESC IN ('Level 3')*  
*AND 63 <= START\_DATE\_IND <= 123*  
*AND 62 <= END\_DATE\_IND <= 142*

5. Cluster 5

% of records: 7.42%

Average probability: 0.9774

Rule:

*LEARNERSHIP\_ID* IN ('941', '938', '932', '90', '892', '89', '888', '88', '320', '314', '138',  
'1222', '1187', '111', '1070', '1068')  
*AND LSHP\_ETQE\_ID* IN ('1126', '1119', '1115', '1114', '1113')  
*AND ETQE\_ID* IN ('1120', '1119', '1115', '1114', '1113')  
*AND NQF\_LEVEL\_DESC* IN ('Level 2')  
*AND 46 <= END\_DATE\_IND <= 174*  
*AND 48 <= START\_DATE\_IND <= 153*

6. Cluster 6

% of records: 3.75%

Average probability: 0.9898

Rule:

*NQF\_LEVEL\_DESC* IN ('Level 5')  
*AND LEARNERSHIP\_ID* IN ('922', '900', '511', '290', '285', '277', '276', '1133', '1002')  
*AND ETQE\_ID* IN ('1127', '1122', '1120', '1116', '1106')  
*AND LSHP\_ETQE\_ID* IN ('1127', '1126', '1106', '1105')  
*AND 78 <= END\_DATE\_IND <= 142*  
*AND 48 <= START\_DATE\_IND <= 123*

7. Cluster 7

% of records: 3.64%

Average probability: 0.9874

Rule:

*LEARNERSHIP\_ID* IN ('668', '664', '661', '618', '608', '1537', '1477', '1476', '1475',  
'1471', '1466', '1465', '1464', '1463', '1459')  
*AND ETQE\_ID* IN ('1111')  
*AND LSHP\_ETQE\_ID* IN ('1111')  
*AND 46 <= END\_DATE\_IND <= 190*  
*AND NQF\_LEVEL\_DESC* IN ('Level 3', 'Level 2')  
*AND 48 <= START\_DATE\_IND <= 168*

8. Cluster 8

% of records: 2.59%

Average probability: 0.9315

Rule:

NQF\_LEVEL\_DESC IN ('Level 1')  
AND LEARNERSHIP\_ID IN ('405', '289', '110')  
AND ETQE\_ID IN ('1115', '1113')  
AND LSHP\_ETQE\_ID IN ('1115', '1113')  
AND 46 <= END\_DATE\_IND <= 142  
AND 33 <= START\_DATE\_IND <= 138

## Appendix O

This appendix provides a technical description of the outputs of data mining activities that were conducted when analysing whether, in the case where the learner has achieved the qualification, and the qualification is a unit standards based qualification, the learner has achieved the correct number and mix of credits for the qualification. The data mining activities focuses on gaining a better understanding of data records that fall into specific categories of the data fields USTD\_CORE\_IND, USTD\_ELEC\_IND, USTD\_FUND\_IND and USTD\_CREDIT\_IND (see Section Appem0) and the possible identification of anomalous data records in the respective data sets.

This semantic business rule defines that in the case where the learner has achieved the qualification, and the qualification is a unit standards based qualification, the learner has achieved the correct number and mix of credits for the qualification.

### *0.1 Insufficient Unit Standard Credits Achieved*

This section provides a technical description of the clusters that were generated by cluster data mining the all qualification enrolment records where:

- the learner has achieved the qualification,
- the qualification is a unit standards based qualification,
- the learner has achieved less than the required number credits for the qualification (see Section 4.10.1).

A detailed description of the manner in which cluster data mining was conducted can be found in Appendix I.3.

The results of the generated clustering model are significant because the model was measured as being 95.51% accurate. The generated clusters show a tight coupling between data fields that describe the ETQEs, qualifications and providers. This is as a result of the organic relationship between qualifications and ETQEs (qualifications are generally implemented by one ETQE only) and providers and ETQEs (providers generally offer qualifications that are implemented by their primary ETQE).

The model did not generate any clusters that contained less than 1% of the records in this group. The detection of anomalous records based on cluster specific probabilities of less than .6 being allocated to the record was conducted. In this manner it was determined that 1.25%

of the records in this group possibly exist in this group as a result of data capturing problems at the source of the data.

The technical description of each of the 8 clusters is provided below. The description is limited to the top 10 attributes of the cluster rule, where each indent of text represents the importance of the attribute. For example:

A line formatted like this represents an importance of 100%

*A line formatted like this represents an importance greater than 75% and less than 100%*

*A line formatted like this represents an importance greater than 50% and less than or equal to 75%*

*A line formatted like this represents an importance less than or equal to 50%*

### 1. Cluster 1

% of records: 20.37%

Average probability: 0.9729

Rule:

*PROVIDER\_ID IN ('5994', '5368', '5205', '5093', '50489', '50456', '5004', '44687', '44322', '44308', '43921', '43400', '43395', '41566', '4040', '39834', '38660', '38654', '38614', '38611', '38602', '38580', '38578', '38570', '38563', '38561', '38560', '38555', '38554', '38551', '36858', '36831', '36827', '35421', '33812', '33692', '32807', '32662', '32523', '32520', '29980', '29277', '29273', '29272', '2924', '28903', '28899', '28872', '28849', '28845', '2883', '28810', '28714', '28531', '2657', '2656', '26362', '2626', '2620', '25099', '25098', '25097', '25095', '25081', '25073', '25062', '25033', '22993', '22831', '2251', '2224', '2219', '2206', '21942', '21887', '2181', '2168', '21665', '20599', '20589', '20574', '20357', '19930', '19798', '1943', '17766', '17531', '17490', '1276', '11228', '11077', '10923', '10418', '10196', '10162')*

*AND SUBFIELD\_DESC IN ('Manufacturing and Assembly', 'Fabrication and Extraction')*

*AND ASSESSOR\_ID IN ('NULL', '6029282', '5657630', '4651893', '4651820', '4651812', '4651663', '4275597', '4275591', '3571204', '3308441', '3308402', '3308394', '3308361', '3308330', '3308314', '3296504', '3296503', '3296475', '3296462', '3296447', '3296383', '3033337', '3033264', '3033192', '3033146', '3032515', '3032209', '3032191', '3031914', '3031680', '3031550', '3031064', '3030878', '3030575', '3029719', '3029488', '3029390', '3029151', '3029069', '3028745', '3028609', '3028393', '3028346', '3028194', '3028138', '3028047',*

'3027999', '3027669', '3027621', '3027397', '3027126', '3027050', '3025292',  
'3025242', '3024596', '3024350', '3024201', '3024002', '3023860', '3023828',  
'3022092', '3021901', '3021895', '3021566', '3021188', '3021165', '3020879',  
'3018627', '3018477', '3017892', '3017839', '3015642', '3015331', '3015228',  
'3014604', '3014393', '3014198', '3012257', '3012015', '3011326', '3008731',  
'3008593', '3008202', '3008039', '3005702', '3005689', '3005091', '3002628',  
'3002435', '3002419')

AND LEARNERSHIP\_ID IN ('NULL', '778', '776', '775', '774', '754', '753', '752', '749',  
'744', '742', '740', '739', '733', '732', '723', '718', '717', '708', '706', '705', '703', '702',  
'700', '697', '696', '693', '692', '685', '668', '377', '369', '365', '350', '340', '337', '336',  
'306', '305', '1499', '1498', '1483', '1470', '1465', '1463', '1461', '1460', '1459', '1388',  
'1376', '1297', '1274', '1271', '1267')

AND QUALIFICATION\_ID IN ('78961', '71967', '66791', '65226', '65206', '65066',  
'64827', '64826', '62886', '61566', '60311', '60310', '59868', '59322', '59033', '58777',  
'58756', '58556', '58551', '58531', '58514', '58284', '58244', '58043', '57897', '57711',  
'49760', '49706', '49467', '49466', '49108', '49062', '49035', '49031', '49030', '48980',  
'48976', '36171', '24511', '24473', '24472', '24231', '23695', '23694', '23295', '23294',  
'23291', '23290', '23271', '23270', '22887', '22886', '22884', '22882', '22877', '22876',  
'22875', '22788', '22787', '21830', '21829', '21032', '21031', '21022', '21021', '20736',  
'20674', '20671', '20524', '20214', '20211')

AND FIELD\_DESC IN ('Manufacturing, Engineering and Technology')

AND ETQE\_ID IN ('1111', '1107', '1103')

AND NQF\_LEVEL\_DESC IN ('Level 3', 'Level 2', 'Level 1')

AND ENROL\_TYPE\_DESC IN ('Residential Learning (i.e. Contact Mode)',  
'Mixed Mode')

AND USTD\_MIX\_IND\_DESC IN ('Insufficient Credits Achieved, Insufficient  
Core Credits, Insufficient Fundamental Credits, Insufficient Elective Credits',  
'Insufficient Credits Achieved, Insufficient Core Credits, Insufficient  
Fundamental Credits, Elective Credits OK', 'Insufficient Credits Achieved,  
Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK',  
'Insufficient Credits Achieved, Core Credits OK, Insufficient Fundamental  
Credits, Insufficient Elective Credits', 'Insufficient Credits Achieved, Core  
Credits OK, Insufficient Fundamental Credits, Elective Credits OK',

*'Insufficient Credits Achieved, Core Credits OK, Fundamental Credits OK, Insufficient Elective Credits')*

2. Cluster 2

% of records: 18.90%

Average probability: 0.9454

Rule:

*ASSESSOR\_ID IN ('NULL')*

*AND LEARNERSHIP\_ID IN ('NULL', '942', '941', '892', '888', '74', '322', '321', '320', '240')*

*AND QUALIFICATION\_ID IN ('73226', '63487', '61772', '60207', '60186', '59966', '59406', '58968', '58223', '57840', '57821', '57625', '50242', '49709', '49705', '49665', '49106', '49099', '49069', '49028', '48994', '48992', '48982', '48865', '48688', '48590', '48512', '48492', '48491', '48490', '24290', '24190', '24150', '23990', '23870', '23850', '23672', '23492', '23470', '23391', '22994', '22690', '22687', '21870', '21810', '20830', '20432', '20203', '20201', '20194', '17191', '14871', '14868', '14674', '14136', '14132', '14129', '14127', '13733')*

*AND PROVIDER\_ID IN ('796', '672', '592', '51338', '44688', '44653', '44198', '44005', '43921', '43153', '42989', '41493', '41386', '41275', '41258', '41214', '41137', '40960', '40942', '40939', '40926', '40925', '40919', '40906', '40900', '40875', '40410', '40360', '39897', '39490', '38989', '38608', '37616', '37405', '37222', '37171', '36875', '36684', '36682', '36681', '36072', '35961', '35956', '35955', '35943', '35942', '35940', '35926', '35925', '35920', '33758', '33692', '32245', '32240', '32211', '31751', '31725', '31705', '30217', '29709', '29680', '29676', '29324', '29313', '29299', '29298', '29297', '29282', '29265', '29257', '29251', '29250', '29242', '29240', '29238', '29237', '29212', '28156', '28000', '27566', '27128', '27125', '27114', '27104', '27090', '27059', '26508', '26481', '25389', '25387', '25384', '25356', '25337', '25321', '25317', '25259', '25217', '25207', '23498', '22771', '22745', '22719', '21975', '21964', '21961', '21954', '21948', '21945', '21942', '21887', '2076', '1976', '1916', '1914', '1909', '1907', '1898', '18759', '18758', '18755', '18710', '18695', '18584', '18574', '18518', '17127', '17121', '17119', '17114', '1641', '1578', '15241', '13961', '12888', '12666', '12489', '11990', '11259')*

*AND SUBFIELD\_DESC IN ('Transport, Operations and Logistics', 'Secondary Agriculture', 'Public Administration', 'Primary Agriculture', 'Personal Care', 'Nature Conservation', 'Human Resources', 'Hospitality, Tourism, Travel, Gaming and*

*Leisure', 'Forestry and Wood Technology', 'Finance, Economics and Accounting', 'Electrical Infrastructure Construction', 'Civil Engineering Construction', 'Building Construction')*

*AND ETQE\_ID IN ('1127', '1126', '1125', '1122', '1120', '1112', '1110', '1109', '1102', '1075')*

*AND FIELD\_DESC IN ('Services', 'Physical Planning and Construction', 'Business, Commerce and Management Studies', 'Agriculture and Nature Conservation')*

*AND ENROL\_TYPE\_DESC IN ('Work Place Learning', 'Unknown', 'Mixed Mode')*

*AND NQF\_LEVEL\_DESC IN ('Level 5', 'Level 3', 'Level 2')*

*AND USTD\_MIX\_IND\_DESC IN ('Insufficient Credits Achieved, Insufficient Core Credits, Insufficient Fundamental Credits, Insufficient Elective Credits', 'Insufficient Credits Achieved, Insufficient Core Credits, Insufficient Fundamental Credits, Elective Credits OK', 'Insufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK', 'Insufficient Credits Achieved, Core Credits OK, Insufficient Fundamental Credits, Insufficient Elective Credits', 'Insufficient Credits Achieved, Core Credits OK, Insufficient Fundamental Credits, Elective Credits OK', 'Insufficient Credits Achieved, Core Credits OK, Fundamental Credits OK, Insufficient Elective Credits')*

### 3. Cluster 3

% of records: 13.18%

Average probability: 0.9632

Rule:

*LEARNERSHIP\_ID IN ('NULL')*

*AND ASSESSOR\_ID IN ('NULL')*

*AND QUALIFICATION\_ID IN ('24214', '22507', '20513')*

*AND PROVIDER\_ID IN ('48293', '48150', '47993', '47992', '47651', '47584', '47498', '47468', '46946', '46927', '46496', '31696', '21121', '21017', '21014', '20861', '20860', '20691')*

*AND SUBFIELD\_DESC IN ('Safety in Society')*

*AND ETQE\_ID IN ('1105')*

*AND FIELD\_DESC IN ('Law, Military Science and Security')*

*AND USTD\_MIX\_IND\_DESC IN ('Insufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK', 'Insufficient Credits Achieved, Core Credits OK, Fundamental Credits OK, Insufficient Elective Credits (Qual Linked to Lshp)', 'Insufficient Credits Achieved, Core Credits OK, Fundamental Credits OK, Insufficient Elective Credits')*

*AND ENROL\_TYPE\_DESC IN ('Mixed Mode')*

*AND NQF\_LEVEL\_DESC IN ('Level 5', 'Level 4')*

#### 4. Cluster 4

% of records: 12.66%

Average probability: 0.9492

Rule:

*LEARNERSHIP\_ID IN ('NULL')*

*AND QUALIFICATION\_ID IN ('49623')*

*AND ASSESSOR\_ID IN ('NULL')*

*AND PROVIDER\_ID IN ('49726', '49724', '39104', '38989', '37712', '37696', '37662', '37650', '37638', '34618', '34617', '32184', '32170', '32167', '31751', '30037', '29845', '29831', '29371', '29370', '29369', '29368', '24960', '2158', '2143', '2140')*

*AND SUBFIELD\_DESC IN ('Promotive Health and Developmental Services')*

*AND ETQE\_ID IN ('1117')*

*AND FIELD\_DESC IN ('Health Sciences and Social Services')*

*AND NQF\_LEVEL\_DESC IN ('Level 1')*

*AND USTD\_MIX\_IND\_DESC IN ('Insufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Insufficient Elective Credits', 'Insufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK')*

*AND ENROL\_TYPE\_DESC IN ('Mixed Mode')*

#### 5. Cluster 5

% of records: 11.53%

Average probability: 0.9119

Rule:

*ASSESSOR\_ID IN ('NULL', '3015109')*

AND LEARNERSHIP\_ID IN ('NULL', '945', '943', '822', '735', '1093', '1036')  
 AND QUALIFICATION\_ID IN ('78982', '78981', '74647', '71566', '61726', '61686',  
 '61467', '60312', '60206', '59114', '58799', '57934', '57841', '50389', '50097', '49946',  
 '49852', '49708', '49148', '49110', '49038', '49026', '36453', '35945', '24471', '24311',  
 '24310', '23970', '23673', '21813', '21020', '20924', '20911', '20202', '20172', '14133',  
 '14130', '14128')  
 AND PROVIDER\_ID IN ('18710', '1760', '17122', '17121', '17119', '1575', '15241',  
 '14640', '13248', '13157', '13127', '11772', '11691', '796', '715', '51225', '51199',  
 '51145', '5093', '48720', '48559', '45504', '44653', '44210', '44202', '44143', '44010',  
 '44003', '42401', '42312', '42217', '41566', '41335', '41275', '41268', '41258', '41214',  
 '41198', '41192', '41137', '41072', '41011', '41001', '40960', '40949', '40943', '40942',  
 '40939', '40900', '40899', '40844', '40623', '40567', '40360', '40269', '39043', '38580',  
 '37645', '37623', '37392', '37289', '37142', '37102', '36372', '35920', '35735', '33758',  
 '33689', '32250', '32240', '32230', '32211', '31991', '31764', '31751', '31750', '29730',  
 '29370', '28156', '27090', '26995', '26380', '26362', '26342', '25429', '25369', '25217',  
 '24850', '23498', '23274', '22911', '22771', '22717', '21938', '21887', '2162', '21561',  
 '1971', '1967', '1943', '1916', '1915', '1914', '1896', '18758')  
 AND NQF\_LEVEL\_DESC IN ('Level 4')  
 AND SUBFIELD\_DESC IN ('Wholesale and Retail', 'Public Administration',  
 'Promotive Health and Developmental Services', 'Primary Agriculture', 'Office  
 Administration', 'Marketing', 'Information Technology and Computer Sciences',  
 'Human Resources', 'Hospitality, Tourism, Travel, Gaming and Leisure', 'Generic  
 Management', 'Finance, Economics and Accounting', 'Engineering and Related  
 Design', 'Communication Studies', 'Cleaning, Domestic, Hiring, Property and Rescue  
 Services', 'Civil Engineering Construction')  
 AND ENROL\_TYPE\_DESC IN ('Work Place Learning', 'Mixed Mode')  
 AND FIELD\_DESC IN ('Services', 'Physical, Mathematical, Computer and Life  
 Sciences', 'Health Sciences and Social Services', 'Business, Commerce and  
 Management Studies')  
 AND ETQE\_ID IN ('1127', '1126', '1125', '1123', '1122', '1120', '1117', '1112',  
 '1110', '1103', '1075')  
 AND QUALIFICATION\_TYPE\_DESC IN ('National Certificate', 'Further Ed  
 and Training Cert')

6. Cluster 6

% of records: 10.29%

Average probability: 0.9815

Rule:

LEARNERSHIP\_ID IN ('NULL')

AND ASSESSOR\_ID IN ('NULL')

AND PROVIDER\_ID IN ('46767', '27879', '27859', '27594', '27538', '27491', '27465',  
'27445', '27363', '24987', '22719', '22311', '22306', '21561', '18583', '18581', '18580',  
'18577', '18576', '18575', '18574', '18573', '18572', '18567', '17115', '17114', '17110',  
'17109', '17107', '17106', '1641')

AND QUALIFICATION\_ID IN ('21517', '20432', '20172', '17191', '14133', '14132',  
'14130', '14128', '14127')

AND SUBFIELD\_DESC IN ('Hospitality, Tourism, Travel, Gaming and Leisure')

AND ETQE\_ID IN ('1119')

AND FIELD\_DESC IN ('Services')

AND ENROL\_TYPE\_DESC IN ('Unknown')

AND NQF\_LEVEL\_DESC IN ('Level 4', 'Level 2')

AND USTD\_MIX\_IND\_DESC IN ('Insufficient Credits Achieved, Insufficient  
Core Credits, Insufficient Fundamental Credits, Insufficient Elective Credits',  
'Insufficient Credits Achieved, Insufficient Core Credits, Insufficient Fundamental  
Credits, Elective Credits OK', 'Insufficient Credits Achieved, Insufficient Core  
Credits, Fundamental Credits OK, Elective Credits OK')

7. Cluster 7

% of records: 7.60%

Average probability: 0.9521

Rule:

LEARNERSHIP\_ID IN ('NULL')

AND QUALIFICATION\_ID IN ('73271', '58778', '50351', '50350', '24214', '23135',  
'23134', '23133', '23131', '23112', '20513', '20176')

AND PROVIDER\_ID IN ('50114', '37425', '35606', '35590', '34574', '34102', '28504',  
'28475', '28324', '28289', '28273', '28244', '28225', '28215', '28205', '28189', '28165',  
'28162', '28161', '28156', '28151', '28144', '28140', '28089', '28076', '28065', '28063',

'28052', '28049', '28046', '28033', '26399', '23169', '21778', '21450', '21328', '21318',  
'21016', '20861', '20860', '2084', '20834', '2065', '1760', '1590', '15499')  
AND SUBFIELD\_DESC IN ('Safety in Society', 'Early Childhood Development',  
'Adult Learning')  
AND ASSESSOR\_ID IN ('NULL', '9050502', '8145173', '7309474', '6055425',  
'5518260', '5518027', '5035206', '5013832', '5013348', '5013286', '4632154',  
'4631889', '4631775', '4631738', '4282957', '4282945', '4282924', '4282714',  
'4282588', '3557374', '3557335', '3557298', '3557258', '3313386', '3313130',  
'3313117', '3313047', '3312995', '3039842', '3039765', '3039613', '3033203',  
'3033062', '3032951', '3032923', '3032883', '3032779', '3032608', '3032482',  
'3032225', '3032145', '3031940', '3031865', '3031732', '3031549', '3031375',  
'3031106', '3030624', '3030552', '3030415', '3030200', '3030163', '3030035',  
'3029923', '3029892', '3029730', '3029647', '3029623', '3029299', '3029008',  
'3028581', '3028388', '3028286', '3027392', '3027298', '3027229', '3027166',  
'3027165', '3027008', '3025400', '3024892', '3023875', '3023816', '3022113',  
'3022085', '3021903', '3021694', '3021461', '3021436', '3020705', '3020604',  
'3018917', '3018881', '3018871', '3018856', '3018377', '3018346', '3018255',  
'3018071', '3017863', '3017502', '3017369', '3017339', '3015708', '3015695',  
'3015648', '3015635', '3015632', '3015595', '3015442', '3015072', '3014866',  
'3014782', '3014150', '3011708', '3011688', '3011603', '3011205', '3011049',  
'3009255', '3009227', '3008999', '3008822', '3008723', '3008537', '3008471',  
'3008406', '3007815', '3005950', '3005505', '3005384', '3005373', '3002908',  
'3002886', '3002844', '3002650', '3002433', '3002153', '3002121', '3002101')  
AND ETQE\_ID IN ('1106', '1105')  
AND FIELD\_DESC IN ('Law, Military Science and Security', 'Education,  
Training and Development')  
AND ENROL\_TYPE\_DESC IN ('Residential Learning (i.e. Contact Mode)')  
AND QUALIFICATION\_TYPE\_DESC IN ('National Diploma', 'National  
Certificate', 'Further Ed and Training Cert')  
AND NQF\_LEVEL\_DESC IN ('Level 5', 'Level 4')

8. Cluster 8

% of records: 5.46%

Average probability: 0.9595

Rule:

LEARNERSHIP\_ID IN ('NULL')

AND QUALIFICATION\_ID IN ('49623', '49102', '23391', '23210')

AND ASSESSOR\_ID IN ('NULL')

AND PROVIDER\_ID IN ('29371', '29370', '29369', '29368', '27146', '25380', '25376',  
'2162', '2160', '2158', '2151', '2140', '21375', '2135', '18518', '1641')

AND ETQE\_ID IN ('1117')

AND SUBFIELD\_DESC IN ('Promotive Health and Developmental Services',  
'Curative Health')

AND FIELD\_DESC IN ('Health Sciences and Social Services')

AND NQF\_LEVEL\_DESC IN ('Level 1')

AND USTD\_MIX\_IND\_DESC IN ('Insufficient Credits Achieved, Insufficient  
Core Credits, Fundamental Credits OK, Insufficient Elective Credits', 'Insufficient  
Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective  
Credits OK')

AND ENROL\_TYPE\_DESC IN ('Unknown')

## ***O.2 No Unit Standard Credits Achieved***

This section provides a technical description of the clusters that were generated by cluster data mining the all qualification enrolment records where:

- the learner has achieved the qualification,
- the qualification is a unit standards based qualification,
- the learner has not achieved any credits for the qualification (see Section 4.10.2).

A detailed description of the manner in which cluster data mining was conducted can be found in Appendix I.3.

The results of the generated clustering model are significant because the model was measured as being 97.38% accurate. The generated clusters show a tight coupling between data fields that describe the ETQEs, qualifications and providers. This is as a result of the organic relationship between qualifications and ETQEs (qualifications are generally implemented by one ETQE only) and providers and ETQEs (providers generally offer qualifications that are implemented by their primary ETQE).

The model did not generate any clusters that contained less than 1% of the records in this group. The detection of anomalous records based on cluster specific probabilities of less than .6 being allocated to the record was conducted. In this manner it was determined that 1.76% of the records in this group possibly exist in this group as a result of data capturing problems at the source of the data.

The technical description of each of the 8 clusters is provided below. The description is limited to the top 10 attributes of the cluster rule, where each indent of text represents the importance of the attribute. For example:

A line formatted like this represents an importance of 100%

*A line formatted like this represents an importance greater than 75% and less than 100%*

*A line formatted like this represents an importance greater than 50% and less than or equal to 75%*

*A line formatted like this represents an importance less than or equal to 50%*

#### 1. Cluster 1

% of records: 32.05%

Average probability: 0.9809

Rule:

*LEARNERSHIP\_ID IN ('NULL', '778', '775', '752', '749', '744', '742', '732', '717', '702', '692', '483', '320', '306')*

*AND PROVIDER\_ID IN ('5994', '5368', '5205', '51249', '51225', '51195', '51191', '51153', '51138', '5093', '50489', '5004', '45504', '43467', '43400', '43399', '43395', '42591', '42515', '42512', '42418', '42411', '42401', '42335', '42320', '42312', '42308', '42279', '42228', '42217', '42194', '42163', '42156', '42119', '42090', '42059', '42058', '42001', '41992', '41987', '41952', '41926', '41840', '41830', '41811', '41719', '38611', '36888', '35940', '35608', '35436', '35421', '34337', '33692', '32523', '3053', '29620', '29282', '29274', '29265', '28899', '28872', '2883', '28810', '28298', '2748', '2657', '26373', '2626', '2620', '25098', '25097', '25062', '25033', '23470', '23318', '22993', '21954', '21948', '21887', '2168', '21561', '21391', '20357', '1943', '17568', '17495', '17465', '11244', '10923', '10418', '10196', '10195', '10162')*

*AND QUALIFICATION\_ID IN ('79963', '71967', '66226', '64827', '64826', '64806', '58968', '58551', '58532', '58514', '58244', '49706', '49467', '49108', '49094', '49035', '48976', '48492', '48490', '36171', '24473', '24472', '23694', '23294', '23290', '23271', '23270', '22886', '22787', '21830', '21031', '21021', '20905', '20736', '13733')*

*AND SUBFIELD\_DESC IN ('Manufacturing and Assembly', 'Information Technology and Computer Sciences')*

*AND ENROL\_TYPE\_DESC IN ('Mixed Mode')*

*AND ETQE\_ID IN ('1125', '1123', '1115', '1113', '1103')*

*AND FIELD\_DESC IN ('Physical, Mathematical, Computer and Life Sciences', 'Manufacturing, Engineering and Technology')*

*AND QUALIFICATION\_TYPE\_DESC IN ('National Certificate')*

*AND NQF\_LEVEL\_DESC IN ('Level 3', 'Level 2', 'Level 1')*

*AND ENROL\_STATUS\_DESC IN ('Achieved')*

## 2. Cluster 2

% of records: 30.74%

Average probability: 0.9933

Rule:

*LEARNERSHIP\_ID IN ('NULL')*

*AND QUALIFICATION\_ID IN ('58778', '23134', '23133', '20855', '20513', '13757', '13756')*

*AND PROVIDER\_ID IN ('46927', '28506', '28307', '28273', '28244', '28242', '28215', '28205', '28189', '28161', '28156', '28153', '28151', '28140', '28138', '28076', '28052', '28051', '28050', '28049', '23172', '23170', '23169', '21389', '21328', '21318', '21017', '21016', '21015', '21014', '21013', '2064')*

*AND SUBFIELD\_DESC IN ('Safety in Society', 'Early Childhood Development')*

*AND ENROL\_TYPE\_DESC IN ('Residential Learning (i.e. Contact Mode)')*

*AND ETQE\_ID IN ('1106', '1105')*

*AND FIELD\_DESC IN ('Law, Military Science and Security', 'Education, Training and Development')*

*AND NQF\_LEVEL\_DESC IN ('Level 5', 'Level 4')*

*AND ENROL\_STATUS\_DESC IN ('Non-Endorsed Achievement', 'Achieved')*

*AND QUALIFICATION\_CLASS\_DESC IN ('Regular-Unit Stds Based')*

## 3. Cluster 3

% of records: 10.51%

Average probability: 0.9824

Rule:

*QUALIFICATION\_ID IN ('73286', '72027', '23671', '21810')*  
*AND LEARNERSHIP\_ID IN ('NULL', '894')*  
*AND PROVIDER\_ID IN ('715', '49153', '42946', '39897', '38226', '37396', '37395',*  
*'37392', '29212', '22771', '22685', '18575', '1575', '15241', '14450', '12666', '12507',*  
*'12489', '11990', '11772', '11691')*  
*AND ETQE\_ID IN ('1126')*  
*AND SUBFIELD\_DESC IN ('Marketing', 'Generic Management')*  
*AND FIELD\_DESC IN ('Business, Commerce and Management Studies')*  
*AND ENROL\_TYPE\_DESC IN ('Work Place Learning', 'RPL for Unknown*  
*Purpose')*  
*AND NQF\_LEVEL\_DESC IN ('Level 3', 'Level 2')*  
*AND QUALIFICATION\_CLASS\_DESC IN ('Regular-Provider-Stds Base')*  
*AND QUALIFICATION\_TYPE\_DESC IN ('National Certificate')*

#### 4. Cluster 4

% of records: 8.17%

Average probability: 0.9576

Rule:

*LEARNERSHIP\_ID IN ('NULL', '39')*  
*AND QUALIFICATION\_ID IN ('61686', '59807', '58556', '58555', '58393', '58392',*  
*'57625', '50389', '50080', '49946', '49709', '49708', '49038', '48982', '48800', '36230',*  
*'35945', '24311', '23990', '23391', '21908', '21907', '14125')*  
*AND PROVIDER\_ID IN ('51339', '51338', '49715', '49496', '49420', '49350', '49342',*  
*'48792', '44653', '42058', '41493', '41386', '41137', '39606', '37975', '36921', '36875',*  
*'35945', '35920', '35319', '34337', '33692', '33689', '32250', '32240', '32219', '32211',*  
*'28156', '27491', '2626', '25429', '24912', '24850', '24847', '24841', '23243', '22771',*  
*'21887', '21561', '21300', '1971', '1943', '1907', '18758', '18710', '18581', '17127',*  
*'17115', '1641', '1575', '15241', '14920', '10923')*  
*AND SUBFIELD\_DESC IN ('Sport', 'Office Administration', 'Manufacturing and*  
*Assembly', 'Human Resources', 'Hospitality, Tourism, Travel, Gaming and Leisure',*  
*'Finance, Economics and Accounting')*  
*AND ETQE\_ID IN ('1127', '1126', '1122', '1119', '1116', '1115', '1110', '1103',*  
*'1075')*

AND QUALIFICATION\_TYPE\_DESC IN ('National Diploma', 'Further Ed and Training Cert')

AND NQF\_LEVEL\_DESC IN ('Level 5', 'Level 4')

AND FIELD\_DESC IN ('Manufacturing, Engineering and Technology', 'Culture and Arts', 'Business, Commerce and Management Studies')

*AND ENROL\_STATUS\_DESC IN ('Achieved')*

*AND ENROL\_TYPE\_DESC IN ('Work Place Learning', 'Unknown', 'Residential Learning (i.e. Contact Mode)', 'RPL for Unknown Purpose', 'Mixed Mode')*

## 5. Cluster 5

% of records: 7.50%

Average probability: 0.9064

Rule:

*LEARNERSHIP\_ID IN ('NULL', '599', '596')*

*AND QUALIFICATION\_ID IN ('61772', '58995', '58161', '58160', '49665', '49142', '49140', '49136', '48828', '48823', '48760', '48688', '48680', '35970', '35944', '23970', '23850', '21808', '20924', '20205', '20190', '20169', '14130', '14128')*

*AND PROVIDER\_ID IN ('716', '5128', '44779', '44653', '44210', '41214', '37407', '37405', '36381', '35671', '33812', '33327', '31672', '29663', '29475', '29461', '28189', '26818', '26799', '26793', '26783', '26780', '26731', '26714', '26698', '26691', '26633', '26632', '26367', '22911', '22771', '21948', '21561', '18710', '18581', '17115', '16889', '16008', '1575', '14891', '14532', '13571', '12894', '12730', '12666', '12489', '12282', '12032', '11691', '11464', '11441')*

*AND ETQE\_ID IN ('1126', '1108')*

*AND SUBFIELD\_DESC IN ('Visual Arts', 'Personal Care', 'Office Administration', 'Music', 'Marketing', 'Information Studies', 'Hospitality, Tourism, Travel, Gaming and Leisure', 'Generic Management', 'Finance, Economics and Accounting', 'Communication Studies', 'Cleaning, Domestic, Hiring, Property and Rescue Services')*

*AND FIELD\_DESC IN ('Services', 'Culture and Arts', 'Business, Commerce and Management Studies')*

*AND QUALIFICATION\_CLASS\_DESC IN ('Regular-Unit Stds Based')*

*AND QUALIFICATION\_TYPE\_DESC IN ('National Certificate')*

*AND ENROL\_TYPE\_DESC IN ('Work Place Learning', 'Unknown', 'RPL for Unknown Purpose', 'Mixed Mode')*

*AND NQF\_LEVEL\_DESC IN ('Level 5', 'Level 4', 'Level 2', 'Level 1')*

6. Cluster 6

% of records: 4.61%

Average probability: 0.9388

Rule:

*PROVIDER\_ID IN ('716', '33565', '33307', '33283', '29461', '26843', '26824', '26812', '26799', '26747', '26746', '26697', '26691', '26683', '26633', '26626', '26621', '26618', '26616', '23274', '21912', '18762')*

*AND QUALIFICATION\_ID IN ('50496', '49144', '49137', '48835', '48829', '48826', '48825', '48823', '48686')*

*AND LEARNERSHIP\_ID IN ('NULL', '602', '596', '593', '592', '591')*

*AND SUBFIELD\_DESC IN ('Visual Arts', 'Performing Arts', 'Marketing', 'Cultural Studies')*

*AND ETQE\_ID IN ('1108')*

*AND NQF\_LEVEL\_DESC IN ('Level 4')*

*AND FIELD\_DESC IN ('Culture and Arts', 'Business, Commerce and Management Studies')*

*AND QUALIFICATION\_CLASS\_DESC IN ('Regular-Unit Stds Based')*

*AND QUALIFICATION\_TYPE\_DESC IN ('Further Ed and Training Cert')*

*AND ENROL\_TYPE\_DESC IN ('Work Place Learning', 'Unknown', 'Mixed Mode')*

7. Cluster 7

% of records: 3.83%

Average probability: 0.9301

Rule:

*QUALIFICATION\_ID IN ('49685', '49136', '49031', '49030', '48994', '48989', '48988', '48987', '48828', '48823')*

*AND LEARNERSHIP\_ID IN ('NULL', '599', '596')*

*AND PROVIDER\_ID IN ('716', '46784', '38989', '36480', '33609', '33584', '33564', '33327', '33274', '31972', '31953', '29720', '29715', '29713', '29712', '29690', '29679',*

'29676', '29670', '29478', '29475', '27155', '27154', '26623', '26618', '25389', '25387',  
 '25384', '25373', '25371', '25369', '25367', '25339', '25317', '25310', '25270', '25207',  
 '25199', '24849', '21879', '1276', '11228', '11127')  
 AND SUBFIELD\_DESC IN ('Visual Arts', 'Primary Agriculture', 'Horticulture',  
 'Fabrication and Extraction')  
 AND FIELD\_DESC IN ('Culture and Arts', 'Agriculture and Nature Conservation')  
 AND ENROL\_TYPE\_DESC IN ('Unknown')  
 AND ETQE\_ID IN ('1112', '1111', '1108')  
 AND NQF\_LEVEL\_DESC IN ('Level 2', 'Level 1')  
 AND QUALIFICATION\_CLASS\_DESC IN ('Regular-Unit Stds Based')  
 AND QUALIFICATION\_TYPE\_DESC IN ('National Certificate')

#### 8. Cluster 8

% of records: 2.59%

Average probability: 0.9847

Rule:

*PROVIDER\_ID* IN ('37392', '35671', '22771', '15241', '14532')  
 AND *QUALIFICATION\_ID* IN ('72028', '59114', '48683', '24471')  
 AND *LEARNERSHIP\_ID* IN ('NULL', '908')  
 AND *SUBFIELD\_DESC* IN ('Consumer Services', 'Cleaning, Domestic, Hiring,  
 Property and Rescue Services')  
 AND *ETQE\_ID* IN ('1126')  
 AND *FIELD\_DESC* IN ('Services')  
 AND *NQF\_LEVEL\_DESC* IN ('Level 4')  
 AND *ENROL\_TYPE\_DESC* IN ('Unknown', 'RPL for Unknown Purpose')  
 AND *QUALIFICATION\_CLASS\_DESC* IN ('Regular-Unit Stds Based')  
 AND *QUALIFICATION\_TYPE\_DESC* IN ('Further Ed and Training Cert')

### ***O.3 Incorrect Mix of Unit Standard Credits Achieved***

This section provides a technical description of the clusters that were generated by cluster data mining the all qualification enrolment records where:

- the learner has achieved the qualification,
- the qualification is a unit standards based qualification,

- the learner has achieved the required number of credits for the qualification, however the number of credits derived from core, fundamental or elective unit standards is incorrect (see Section 4.10.3).

A detailed description of the manner in which cluster data mining was conducted can be found in Appendix I.3.

The results of the generated clustering model are significant because the model was measured as being 96.28% accurate. The generated clusters show a tight coupling between data fields that describe the ETQEs, qualifications and providers. This is as a result of the organic relationship between qualifications and ETQEs (qualifications are generally implemented by one ETQE only) and providers and ETQEs (providers generally offer qualifications that are implemented by their primary ETQE).

The model did not generate any clusters that contained less than 1% of the records in this group. The detection of anomalous records based on cluster specific probabilities of less than .6 being allocated to the record was conducted. In this manner it was determined that 3.73% of the records in this group possibly exist in this group as a result of data capturing problems at the source of the data.

The technical description of each of the 8 clusters is provided below. The description is limited to the top 10 attributes of the cluster rule, where each indent of text represents the importance of the attribute. For example:

A line formatted like this represents an importance of 100%

*A line formatted like this represents an importance greater than 75% and less than 100%*

*A line formatted like this represents an importance greater than 50% and less than or equal to 75%*

*A line formatted like this represents an importance less than or equal to 50%*

#### 1. Cluster 1

% of records: 38.86%

Average probability: 0.9592

Rule:

*ASSESSOR\_ID IN ('NULL', '3032191', '3028609', '3028349', '3028194', '3028159', '3022092')*

*AND LEARNERSHIP\_ID IN ('NULL', '829', '752', '749', '717', '696', '693', '692', '668', '420', '377', '350', '341', '337', '1480', '1477', '1476', '1474', '1470', '1465', '1464', '1463', '1459', '1376', '1277', '1242', '1093')*  
*AND QUALIFICATION\_ID IN ('65466', '59868', '59566', '59322', '58799', '58798', '58777', '58756', '58284', '57711', '49108', '49031', '49030', '48865', '36171', '24472', '23290', '21829', '21031', '21022', '21021', '20671', '20524', '20215', '20211')*  
*AND PROVIDER\_ID IN ('715', '50456', '46459', '46423', '44687', '44471', '44322', '44308', '44224', '43231', '41576', '41566', '41565', '41482', '38667', '38666', '38660', '38610', '38602', '38592', '38591', '38580', '38570', '38563', '38562', '38560', '38555', '38554', '38551', '36858', '36844', '36842', '36838', '36831', '34946', '33654', '32807', '29980', '29275', '2883', '28810', '2748', '27123', '27074', '2657', '25337', '25062', '22993', '22831', '2251', '2231', '2219', '2206', '21942', '2168', '20574', '1945', '1943', '18695', '17531', '12894', '11228', '11127')*  
*AND FIELD\_DESC IN ('Manufacturing, Engineering and Technology')*  
*AND SUBFIELD\_DESC IN ('Secondary Agriculture', 'Manufacturing and Assembly', 'Fabrication and Extraction')*  
*AND ENROL\_TYPE\_DESC IN ('Mixed Mode')*  
*AND ETQE\_ID IN ('1112', '1111', '1107', '1103')*  
*AND NQF\_LEVEL\_DESC IN ('Level 3', 'Level 2')*  
*AND QUALIFICATION\_TYPE\_DESC IN ('National Certificate')*

## 2. Cluster 2

% of records: 15.43%

Average probability: 0.9485

Rule:

*QUALIFICATION\_ID IN ('58778', '57820', '48889', '48680', '23870', '23133', '20190')*  
*AND LEARNERSHIP\_ID IN ('NULL', '906')*  
*AND PROVIDER\_ID IN ('45504', '42565', '42427', '42401', '42217', '42096', '37154', '28523', '28076', '28063', '23321', '23274', '22771', '22719', '21887', '21328', '18695', '16527', '15241', '14891', '14713', '14532', '14487', '12666', '12486', '12288', '11464')*  
*AND ASSESSOR\_ID IN ('NULL', '3031395', '3026513', '3018346', '3014106', '3007230')*

*AND SUBFIELD\_DESC IN ('Personal Care', 'Information Technology and Computer Sciences', 'Early Childhood Development', 'Consumer Services', 'Cleaning, Domestic, Hiring, Property and Rescue Services')*

*AND ETQE\_ID IN ('1126', '1123', '1106')*

*AND FIELD\_DESC IN ('Services', 'Physical, Mathematical, Computer and Life Sciences')*

*AND USTD\_MIX\_IND\_DESC IN ('Sufficient Credits Achieved, Core Credits OK, Insufficient Fundamental Credits, Elective Credits OK')*

*AND ENROL\_TYPE\_DESC IN ('Work Place Learning', 'Unknown', 'RPL for Unknown Purpose', 'Mixed Mode')*

*AND QUALIFICATION\_TYPE\_DESC IN ('National Certificate')*

### 3. Cluster 3

% of records: 15.14%

Average probability: 0.9550

Rule:

*LEARNERSHIP\_ID IN ('NULL', '240', '215', '189', '1066')*

*AND ASSESSOR\_ID IN ('NULL')*

*AND QUALIFICATION\_ID IN ('49685', '24290', '24190', '24150', '20936', '20831', '20830')*

*AND PROVIDER\_ID IN ('48558', '43153', '33758', '33756', '33692', '32039', '32014', '31991', '31975', '29782', '29720', '29700', '29282', '27135', '27128', '27125', '27119', '27118', '27114', '27107', '27100', '27097', '27090', '27074', '27063', '27059', '27008', '26995', '26986', '25389', '25387', '25366', '22236', '21879', '14640')*

*AND ETQE\_ID IN ('1112', '1109')*

*AND SUBFIELD\_DESC IN ('Horticulture', 'Civil Engineering Construction', 'Building Construction')*

*AND FIELD\_DESC IN ('Physical Planning and Construction', 'Agriculture and Nature Conservation')*

*AND ENROL\_TYPE\_DESC IN ('Mixed Mode')*

*AND QUALIFICATION\_TYPE\_DESC IN ('National Certificate')*

*AND NQF\_LEVEL\_DESC IN ('Level 3', 'Level 2', 'Level 1')*

### 4. Cluster 4

% of records: 8.13%

Average probability: 0.9816

Rule:

LEARNERSHIP\_ID IN ('NULL')

*AND PROVIDER\_ID IN ('641', '35381', '34102', '28156', '28076', '28065', '23470', '23286', '23274', '21836', '21016', '18518', '18517', '18511', '11253', '11250', '11242', '11241')*

*AND QUALIFICATION\_ID IN ('50351', '50349', '48590', '20513', '13736')*

*AND ASSESSOR\_ID IN ('NULL', '3032951', '3030481', '3029623', '3018607', '3014919', '3005373', '3002844')*

*AND ETQE\_ID IN ('1114', '1106', '1105', '1104')*

*AND SUBFIELD\_DESC IN ('Wholesale and Retail', 'Safety in Society', 'Information Technology and Computer Sciences', 'Adult Learning')*

*AND USTD\_MIX\_IND\_DESC IN ('Sufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK')*

*AND FIELD\_DESC IN ('Services', 'Physical, Mathematical, Computer and Life Sciences', 'Law, Military Science and Security', 'Education, Training and Development')*

*AND NQF\_LEVEL\_DESC IN ('Level 5', 'Level 4')*

*AND QUALIFICATION\_TYPE\_DESC IN ('National Certificate')*

## 5. Cluster 5

% of records: 7.43%

Average probability: 0.9965

Rule:

LEARNERSHIP\_ID IN ('NULL')

*AND QUALIFICATION\_ID IN ('78981')*

*AND PROVIDER\_ID IN ('45504', '42546', '42320', '42228', '42217', '42058')*

*AND ASSESSOR\_ID IN ('NULL')*

*AND SUBFIELD\_DESC IN ('Information Technology and Computer Sciences')*

*AND ETQE\_ID IN ('1123')*

*AND FIELD\_DESC IN ('Physical, Mathematical, Computer and Life Sciences')*

*AND NQF\_LEVEL\_DESC IN ('Level 4')*

*AND ENROL\_TYPE\_DESC IN ('Mixed Mode')*

AND USTD\_MIX\_IND\_DESC IN ('Sufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK')

6. Cluster 6

% of records: 6.13%

Average probability: 0.9769

Rule:

*ASSESSOR\_ID IN ('NULL')*

*AND LEARNERSHIP\_ID IN ('NULL', '504', '491')*

*AND PROVIDER\_ID IN ('796', '48792', '48375', '46460', '44114', '39616', '39271', '37680', '37392', '37133', '37132', '36921', '36875', '32245', '32240', '32210', '28116', '25429', '24850', '22771', '2076', '1976', '1914', '1913', '18710', '18699', '17122', '17121', '17119', '15241', '12888', '12655')*

*AND QUALIFICATION\_ID IN ('58393', '58392', '50498', '49666', '49373', '49106', '49038', '48982', '48510', '23850', '21813', '21745', '20790', '20202', '20194')*

*AND SUBFIELD\_DESC IN ('Public Administration', 'Finance, Economics and Accounting')*

*AND ETQE\_ID IN ('1127', '1126', '1120', '1116', '1110')*

*AND FIELD\_DESC IN ('Business, Commerce and Management Studies')*

*AND NQF\_LEVEL\_DESC IN ('Level 6', 'Level 4', 'Level 3')*

*AND ENROL\_TYPE\_DESC IN ('Unknown', 'Mixed Mode')*

*AND ENROL\_STATUS\_DESC IN ('Achieved')*

7. Cluster 7

% of records: 5.04%

Average probability: 0.9405

Rule:

*ASSESSOR\_ID IN ('NULL')*

*AND QUALIFICATION\_ID IN ('49100', '49065', '48992', '48989', '14871', '14868')*

*AND LEARNERSHIP\_ID IN ('NULL', '828', '805', '802', '801')*

*AND PROVIDER\_ID IN ('49754', '43153', '41603', '41598', '41586', '41576', '39205', '37894', '36480', '34991', '31965', '31953', '29712', '29691', '29679', '29670', '27161', '26986', '25391', '25389', '25384', '25382', '25339', '25321', '25317', '25259', '25219', '25207', '25199', '25194', '21874')*

*AND SUBFIELD\_DESC IN ('Primary Agriculture')*  
*AND ETQE\_ID IN ('1112')*  
*AND FIELD\_DESC IN ('Agriculture and Nature Conservation')*  
*AND ENROL\_TYPE\_DESC IN ('Mixed Mode')*  
*AND USTD\_MIX\_IND\_DESC IN ('Sufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK')*  
*AND QUALIFICATION\_TYPE\_DESC IN ('National Certificate')*

8. Cluster 8

% of records: 3.84%

Average probability: 0.9967

Rule:

*LEARNERSHIP\_ID IN ('NULL')*  
*AND ASSESSOR\_ID IN ('NULL')*  
*AND QUALIFICATION\_ID IN ('61746', '58594', '50139', '49099', '24214', '22508')*  
*AND PROVIDER\_ID IN ('48150', '47993', '47992', '46847', '46496', '43153', '38989', '21121', '21006', '2084', '2066', '2065', '18584')*  
*AND SUBFIELD\_DESC IN ('Safety in Society', 'Forestry and Wood Technology')*  
*AND ETQE\_ID IN ('1113', '1105')*  
*AND FIELD\_DESC IN ('Law, Military Science and Security')*  
*AND ENROL\_TYPE\_DESC IN ('Mixed Mode')*  
*AND QUALIFICATION\_TYPE\_DESC IN ('National Diploma', 'National Certificate')*  
*AND USTD\_MIX\_IND\_DESC IN ('Sufficient Credits Achieved, Insufficient Core Credits, Fundamental Credits OK, Elective Credits OK')*

## Appendix P

The analysis of the learnership, qualification and unit standard data conducted in Chapter 4 did highlight a number of records that infringed on the semantic business rules defined in Section 3.6.2. This appendix provides specific recommendations in regard to learnership, qualification and unit standard data records that infringed the semantic business rules.

The recommendations are based on the results from each specific semantic business rule. As a result the structure of this appendix has sub sections are aligned to the structure of Chapter 4.

### ***P.1 ETQE Accreditation***

#### ***P.1.1 Learnership enrolments***

1. All records that fall into the category ‘Start Before, End During’ (see Section 4.2.1) that are not enrolment records that have been post-loaded as a result of the discovery of missing records after an ETQE amalgamation should be referred back to the submitting ETQE for confirmation that the data captured on the learnership enrolment record is correct.
2. All records that fall into the category ‘Start During, End After’ (see Section 4.2.1) should be referred back to the submitting ETQE for confirmation that the data on the learnership enrolment record is correct.

#### ***P.1.2 Qualification enrolments***

All records that fall into the category ‘Start Before, End During’, ‘Start Before, End Before’ and ‘Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End Before’ (see Section 4.2.2) that are not enrolment records that have been post-loaded as a result of the discovery of missing records after an ETQE amalgamation should be referred back to the submitting ETQE for confirmation that the data captured on the qualification enrolment record is correct.

#### ***P.1.3 Unit Standard enrolments***

All records that fall into the category ‘Start Before, End Before’, ‘Start Before, End During’, ‘Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End During’, ‘Submitting ETQE: Start Before, End During, Other ETQE: Start Before, End During’ or ‘Submitting ETQE: Start Before, End Before, Other ETQE: Start Before, End Before’ (see Section 4.2.3) that are not enrolment records that have been post-loaded as a

result of the discovery of missing records after an ETQE amalgamation should be referred back to the submitting ETQE for confirmation that the data captured on the qualification enrolment record is correct.

## ***P.2 ETQE accreditation to quality assure the qualification or unit standard***

### ***P.2.1 Qualification enrolments***

1. The accuracy of the ETQE accreditations to quality assure each of the qualifications that are linked to enrolment records that infringe on this business rule should be confirmed by SAQA.
2. All enrolment records, where SAQA has found that the ETQE accreditation to quality assure the qualification record is correct, should be referred back to the submitting ETQE for confirmation that the data captured on the qualification enrolment record is correct.
3. Specific note should be taken by SAQA in regard to the high incidence of these types of enrolment records for ETQE identifiers 1104 and 1103.

### ***P.2.2 Unit Standard enrolments***

1. The accuracy of the ETQE accreditations to quality assure each of the qualifications that are linked to unit standard enrolment records that infringe on this business rule should be confirmed by SAQA.
2. All enrolment records, where SAQA has found that the ETQE accreditation to quality assure the qualification record is correct, should be referred back to the submitting ETQE for confirmation that the data captured on the unit standard enrolment record is correct.
3. Specific note should be taken by SAQA in regard to the high incidence of these types of enrolment records for ETQE identifiers 1100 and 1104.

## ***P.3 Provider accreditation***

### ***P.3.1 Learnership enrolments***

1. No Accreditation
  - a. All of the providers that are referenced in learnership enrolment records in this category (see Appendix J.1.2) should be referred back to the primary ETQE of the provider. Further, any ETQE's that are not the primary ETQE of the provider, that

are referencing these providers in their submissions to the NLRD, should be notified that these providers are not accredited.

- b. Specific note should be taken by SAQA in regard to the high incidence of these types of providers for ETQE identifiers 1119 and 1115.

## 2. Start Before, End After

The provider that is referenced in this category (see Appendix J.1.6) should be referred back to the primary ETQE of the provider. The primary ETQE should correct the details in the provider record for the provider and resubmit the record to the NLRD.

## 3. Start Before, End Before or End During

- a. All providers that fall into this category (see Appendix J.1.7) as a result of the learnership enrolment record being submitted to the NLRD by their primary ETQE should be referred back to the submitting ETQE. All providers that fall into this category as a result of the learnership enrolment record being submitted to the NLRD by an ETQE other than the primary ETQE should be referred back to the primary ETQE of the provider.

Further, any ETQEs that are not the primary ETQE of the provider, that are referencing these providers in their submissions to the NLRD, should be made aware of the actual accreditation statuses of these providers.

- b. The records that were found to have a cluster probability low enough to be considered anomalous (see Appendix J.1.7) should be referred back to the submitting ETQE for confirmation that the data captured on the learnership enrolment record is correct.

## 4. Start During, Start After and End After

- a. The seemingly systemic problem around the implementation of the learnerships with learnership identifiers 53, 24 and 460 (see Appendix J.1.8.1 and J.1.8.2) should be referred back to the implementing ETQE of these learnerships.
- b. Any other providers that fall into this category (see Appendix J.1.8) as a result of the learnership enrolment record being submitted to the NLRD by their primary ETQE should be referred back to the submitting ETQE. All providers that fall into this category as a result of the learnership enrolment record being submitted to the

NLRD by an ETQE other than the primary ETQE should be referred back to the primary ETQE of the provider.

Further, any ETQEs that are not the primary ETQE of the provider, that are referencing these providers in their submissions to the NLRD, should be made aware of the actual accreditation statuses of these providers.

- c. The records that were found to have a cluster probability low enough to be considered anomalous (see Appendix J.1.8) should be referred back to the submitting ETQE for confirmation that the data captured on the learnership enrolment record is correct.

### ***P.3.2 Qualification enrolments***

#### **1. No Accreditation**

- a. All of the providers that are referenced in qualification enrolment records in this category (see Appendix J.2.3) should be referred back to the primary ETQE of the provider. Further, any ETQE's that are not the primary ETQE of the provider, that are referencing these providers in their submissions to the NLRD, should be notified that these providers are not accredited.
- b. Specific note should be taken by SAQA in regard to the high incidence of these types of providers for ETQE identifiers 1113, 1119 and 1115.

#### **2. Start Before, End After**

All records that fall into the categories 'Start Before, End After' (see Appendix J.2.6) should be referred back to the submitting ETQE for confirmation that the data on the qualification enrolment record is correct.

#### **3. Start Before, End Before or End During**

- a. All providers that fall into this category (see Appendix J.2.7) as a result of the qualification enrolment record being submitted to the NLRD by their primary ETQE should be referred back to the submitting ETQE. All providers that fall into this category as a result of the qualification enrolment record being submitted to the NLRD by an ETQE other than the primary ETQE should be referred back to the primary ETQE of the provider.

Further, any ETQEs that are not the primary ETQE of the provider, that are referencing these providers in their submissions to the NLRD, should be made aware of the actual accreditation statuses of these providers.

- b. The records that were found to have a cluster probability low enough to be considered anomalous (see Appendix J.2.7) should be referred back to the submitting ETQE for confirmation that the data captured on the qualification enrolment record is correct.
- c. Specific note should be taken by SAQA in regard to the high incidence of these types of records for ETQE identifiers 1105, 1106, 1111, 1116 and 1126.

#### 4. Start During, Start After and End After

- a. The seemingly systemic problems around the offering of the following qualification identifiers must be further investigated by SAQA:
  - i. Qualification identifier 50139, 58594 and 60006 by provider identifiers 2071, 20725, 20772, 30139, 35516, 35551, 38426, 46877 and 46926.
  - ii. Qualification identifier 48550 by provider identifier 1905.
  - iii. Qualification identifiers 73729, 20409 and 20408.
  - iv. Qualification identifiers 24010 and 49623 by provider identifiers 2159, 37747, 38989 and 39001.
- b. Any other providers that fall into this category (see Appendix J.2.8) as a result of the qualification enrolment record being submitted to the NLRD by their primary ETQE should be referred back to the submitting ETQE. All providers that fall into this category as a result of the qualification enrolment record being submitted to the NLRD by an ETQE other than the primary ETQE should be referred back to the primary ETQE of the provider.

Further, any ETQEs that are not the primary ETQE of the provider, that are referencing these providers in their submissions to the NLRD, should be made aware of the actual accreditation statuses of these providers.

- c. The records that were found to have a cluster probability low enough to be considered anomalous (see Appendix J.2.8) should be referred back to the submitting ETQE for confirmation that the data captured on the qualification enrolment record is correct.

### ***P.3.3 Unit Standard enrolments***

#### **1. No Accreditation**

- a. All of the providers that are referenced in unit standard enrolment records in this category (see Appendix J.3.4) should be referred back to the primary ETQE of the provider. Further, any ETQE's that are not the primary ETQE of the provider, that are referencing these providers in their submissions to the NLRD, should be notified that these providers are not accredited.
- b. Specific note should be taken by SAQA in regard to the high incidence of these types of providers for ETQE identifiers 1105 and 1105.

#### **2. Start Before, End After**

All records that fall into the categories 'Start Before, End After' (see Appendix J.3.6) should be referred back to the submitting ETQE for confirmation that the data on the unit standard enrolment record is correct.

#### **3. Start Before, End Before or End During**

- a. All providers that fall into this category (see Appendix J.3.7) as a result of the unit standard enrolment record being submitted to the NLRD by their primary ETQE should be referred back to the submitting ETQE. All providers that fall into this category as a result of the unit standard enrolment record being submitted to the NLRD by an ETQE other than the primary ETQE should be referred back to the primary ETQE of the provider.

Further, any ETQEs that are not the primary ETQE of the provider, that are referencing these providers in their submissions to the NLRD, should be made aware of the actual accreditation statuses of these providers.

- b. The records that were found to have a cluster probability low enough to be considered anomalous (see Appendix J.3.7) should be referred back to the submitting ETQE for confirmation that the data captured on the unit standard enrolment record is correct.
- c. Specific note should be taken by SAQA in regard to the high incidence of these types of records for ETQE identifiers 1111, 1127, 1105, 1116 and 1126.

#### **4. Start During, Start After and End After**

- a. All providers that fall into this category (see Appendix J.3.8) as a result of the unit standard enrolment record being submitted to the NLRD by their primary ETQE should be referred back to the submitting ETQE. All providers that fall into this category as a result of the unit standard enrolment record being submitted to the NLRD by an ETQE other than the primary ETQE should be referred back to the primary ETQE of the provider.

Further, any ETQEs that are not the primary ETQE of the provider, that are referencing these providers in their submissions to the NLRD, should be made aware of the actual accreditation statuses of these providers.

- b. The records that were found to have a cluster probability low enough to be considered anomalous (see Appendix J.3.8) should be referred back to the submitting ETQE for confirmation that the data captured on the unit standard enrolment record is correct.
- c. Specific note should be taken by SAQA in regard to the high incidence of these types of records for ETQE identifiers 1105, 1100, 1126, 1127 and 1115.

#### ***P.4 Provider accreditation to offer the qualification or unit standard***

##### ***P.4.1 Qualification enrolments***

###### **1. No Accreditation**

- a. All of the providers that are referenced in qualification enrolment records in this category (see Appendix L.1.1) should be referred back to the submitting ETQE of the enrolment record.
- b. Specific note should be taken by SAQA in regard to the high incidence of these types of providers for ETQE identifiers 1116, 1126 and 1103.

###### **2. Start Before, End After**

All records that fall into the categories 'Start Before, End After' (see Appendix L.1.6) should be referred back to the submitting ETQE for confirmation that the data on the qualification enrolment record is correct.

###### **3. Start Before, End Before or End During**

- a. All providers that fall into this category (see Appendix L.1.7) should be referred back to the submitting ETQE.

- b. The records that were found to have a cluster probability low enough to be considered anomalous (see Appendix L.1.7) should be referred back to the submitting ETQE for confirmation that the data captured on the qualification enrolment record is correct.
  - c. Specific note should be taken by SAQA in regard to the high incidence of these types of providers for ETQE identifiers 1105, 1106, 1116 and 1126.
4. Start During, Start After and End After
- a. All providers that fall into this category (see Appendix L.1.8) should be referred back to the submitting ETQE.
  - b. The records that were found to have a cluster probability low enough to be considered anomalous (see Appendix L.1.8) should be referred back to the submitting ETQE for confirmation that the data captured on the qualification enrolment record is correct.
  - c. Specific note should be taken by SAQA in regard to the high incidence of these types of providers for ETQE identifiers 1079, 1106 and 1115.

#### ***P.4.2 Unit Standard enrolments***

1. No Accreditation
  - a. All of the providers that are referenced in unit standard enrolment records in this category (see Appendix L.2.1) should be referred back to the submitting ETQE of the enrolment record.
  - b. Specific note should be taken by SAQA in regard to the high incidence of these types of providers for ETQE identifiers 1105, 1103 and 1126.
2. Start Before, End After

All records that fall into the categories ‘Start Before, End After’ (see Appendix L.2.6) should be referred back to the submitting ETQE for confirmation that the data on the unit standard enrolment record is correct.
3. Start Before, End Before or End During
  - a. All providers that fall into this category (see Appendix L.2.7) should be referred back to the submitting ETQE.

- b. The records that were found to have a cluster probability low enough to be considered anomalous (see Appendix L.2.7) should be referred back to the submitting ETQE for confirmation that the data captured on the unit standard enrolment record is correct.
  - c. Specific note should be taken by SAQA in regard to the high incidence of these types of providers for ETQE identifiers 1111, 1106, 1105 and 1126.
4. Start During, Start After and End After
- a. All providers that fall into this category (see Appendix L.2.8) should be referred back to the submitting ETQE.
  - b. The records that were found to have a cluster probability low enough to be considered anomalous (see Appendix L.2.8) should be referred back to the submitting ETQE for confirmation that the data captured on the unit standard enrolment record is correct.
  - c. Specific note should be taken by SAQA in regard to the high incidence of these types of providers for ETQE identifiers 1105, 1126, 1112 and 1075.

### ***P.5 Assessor registration***

1. Learnership enrolments
- a. All records that fall into the categories ‘Lshp Completed After Assessor Registration’ and ‘Lshp Completed Before Assessor Registration’ (see Section 4.6.1) should be referred back to the submitting ETQE for confirmation that the data on the learnership enrolment record is correct.
  - b. All of the assessors that are referenced in learnership enrolment records that fall into the category ‘No Registration’ (see Section 4.6.1) should be referred back to the submitting ETQE.

Specific note should be taken by SAQA in regard to the high percentage of ‘No Registration’ assessors (see Section 4.6.1) in relation to the overall number of assessors registered by ETQE identifier 1116.

2. Qualification enrolments
- a. All records that fall into the categories ‘Qual Achieved After Assessor Registration’ and ‘Qual Achieved Before Assessor Registration’ (see Section

4.6.2) should be referred back to the submitting ETQE for confirmation that the data on the qualification enrolment record is correct.

- b. All of the assessors that are referenced in qualification enrolment records that fall into the category 'No Registration' (see Section 4.6.2) should be referred back to the submitting ETQE.

Specific note should be taken by SAQA in regard to the high percentage of 'No Registration' assessors (see Section 4.6.2) in relation to the overall number of assessors registered by ETQE identifier 1115.

### 3. Unit Standard enrolments

- c. All records that fall into the categories 'Ustd Achieved After Assessor Registration' and 'Ustd Achieved Before Assessor Registration' (see Section 4.6.3) should be referred back to the submitting ETQE for confirmation that the data on the unit standard enrolment record is correct.
- d. All of the assessors that are referenced in unit standard enrolment records that fall into the category 'No Registration' (see Section 4.6.3) should be referred back to the submitting ETQE.

Specific note should be taken by SAQA in regard to the high percentage of 'No Registration' assessors (see Section 4.6.3) in relation to the overall number of assessors registered by ETQE identifier 1116.

## ***P.6 Assessor registration to assess the qualification or unit standard***

### ***P.6.1 Qualification enrolments***

1. All of the assessors that are referenced in qualification enrolment records that fall into the category 'No Registration' (see Section 5.7.2) should be referred back to the submitting ETQE.

Specific note should be taken by SAQA in regard to the high percentage of 'No Registration' assessors (see Section 5.7.2) in relation to the overall number of assessors registered by ETQE identifiers 1105 and 1115.

2. All records that fall into the categories 'Qual Achieved Before Assessor Registration' and 'Qual Achieved After Assessor Registration' (see Section 5.7.2) should be

referred back to the submitting ETQE for confirmation that the data on the qualification enrolment record is correct.

#### ***P.6.2 Unit Standard enrolments***

1. All of the assessors that are referenced in unit standard enrolment records that fall into the category 'No Registration' (see Section 4.7.2) should be referred back to the submitting ETQE.
2. All records that fall into the categories 'UStd Achieved Before Assessor Registration' and 'UStd Achieved After Assessor Registration' (see Section 4.7.2) should be referred back to the submitting ETQE for confirmation that the data on the qualification enrolment record is correct.

#### ***P.7 Correlation between learnerships and their associated qualification***

1. No Qual Enrolment
  - a. The accuracy of the learnership-qualification relationship should be confirmed by SAQA for all learnerships where 100% of the learnership's enrolment records fall into this category (see Section 4.8.1).
  - b. Once the accuracy of the learnership-qualification relationship has been confirmed by SAQA (as per item a. above), these learnership enrolment records should be referred back to the submitting ETQE.
  - c. All remaining learnership enrolment records that do not have an associated qualification enrolment record should be referred back to the submitting ETQE (see Section 4.8.1).
  - d. The records that were found to have a cluster probability low enough to be considered anomalous (see Section 4.8.1) should be referred back to the submitting ETQE for confirmation that the data captured on the learnership enrolment record is correct.
  - e. Specific note should be taken by SAQA in regard to the high percentage of 'No Qual Enrolment' records (see Section 4.8.1) in relation to the overall number of learnerships and referenced providers, as described in Clusters 6 and 8, for ETQE identifiers 1115 and 1120.
2. Lshp Enrolled, Qual Achieved (Derived)

- a. The accuracy of the learnership-qualification relationship should be confirmed by SAQA for all learnerships where qualification enrolments exist with a learnership identifier other than the learnership identifier of the learnership enrolment record (see Section 4.8.2).
  - b. Once the accuracy of the learnership-qualification relationship has been confirmed by SAQA (as per item a. above), these learnership enrolment records should be referred back to the submitting ETQE (see Section 4.8.2).
  - c. All records where the learnership identifier on the associated qualification enrolment record is NULL should be referred back to the submitting ETQE (see Section 4.8.2).
  - d. The records that were found to have a cluster probability low enough to be considered anomalous (see Section 4.8.2) should be referred back to the submitting ETQE for confirmation that the data captured on the learnership enrolment record is correct.
  - e. Specific note should be taken by SAQA in regard to the high percentage of 'Lshp Enrolled, Qual Achieved (Derived)' records (see Section 4.8.2) in relation to the overall number of learnerships, as described in Clusters 2, 4 and 7, for ETQE identifiers 1103, 1105 and 1111.
3. Lshp Enrolled, Qual Achieved
- a. All of the records that fall into this category should be referred back to the submitting ETQE (see Section 4.8.3).
  - b. Specific note should be taken by SAQA in regard to the high percentage of records in this category that have been submitted to the NLRD by ETQE identifier 1115 (see Section 4.8.3).
4. Lshp Completed, Qual Enrolled
- a. All of the associated qualification enrolment records for the learnership enrolment records that fall into this category should be referred back to the submitting ETQE (see Section 4.8.4).
  - b. Specific note should be taken by SAQA in regard to the high percentage of records in this category that have been submitted to the NLRD by ETQE identifier 1126 (see Section 4.8.4).

5. Lshp Completed, Qual Enrolled (Derived)
  - a. The accuracy of the learnership-qualification relationship should be confirmed by SAQA for all learnerships where qualification enrolments exist with a learnership identifier other than the learnership identifier of the learnership enrolment record (see Section 4.8.5).
  - b. Once the accuracy of the learnership-qualification relationship has been confirmed by SAQA (as per item a. the associated qualification enrolment records of these learnership enrolment records should be referred back to the submitting ETQE (see Section 4.8.5).
  - c. All associated qualification enrolment records for learnership records, where the learnership identifier on the associated qualification enrolment record is NULL, should be referred back to the submitting ETQE (see Section 4.8.5).
  
6. Lshp Completed Before Qual (Derived)
  - a. The accuracy of the learnership-qualification relationship should be confirmed by SAQA for all learnerships where qualification enrolments exist with a learnership identifier other than the learnership identifier of the learnership enrolment record (see Section 4.8.6).
  - b. Once the accuracy of the learnership-qualification relationship has been confirmed by SAQA (as per item a. above), both the learnership enrolment record and their associated qualification enrolment records should be referred back to the submitting ETQE (see Section 4.8.6).
  - c. Both the learnership enrolment records and the associated qualification enrolment records, where the learnership identifier on the associated qualification enrolment record is NULL, should be referred back to the submitting ETQE (see Section 4.8.6).
  
7. Lshp Completed Before Qual
  - a. All of the learnership enrolment records and their associated qualification enrolment records that fall into this category should be referred back to the submitting ETQE (see Section 4.8.7).
  - b. Specific note should be taken by SAQA in regard to the high percentage of records in this category that have been submitted to the NLRD by ETQE identifier 1111 (see Section 4.8.7).

8. Lshp Completed After Qual (Derived)

- a. The accuracy of the learnership-qualification relationship should be confirmed by SAQA for all learnerships where qualification enrolments exist with a learnership identifier other than the learnership identifier of the learnership enrolment record (see Section 4.8.8).
- b. Once the accuracy of the learnership-qualification relationship has been confirmed by SAQA (as per item a. above), both the learnership enrolment record and their associated qualification enrolment records should be referred back to the submitting ETQE (see Section 4.8.8).
- c. Both the learnership enrolment records and the associated qualification enrolment records, where the learnership identifier on the associated qualification enrolment record is NULL, should be referred back to the submitting ETQE (see Section 4.8.8).

9. Lshp Completed After Qual

- a. All of the records that fall into this category should be referred back to the submitting ETQE (see Section 4.8.9).
- b. Specific note should be taken by SAQA in regard to the high percentage of records in this category that have been submitted to the NLRD by ETQE identifier 1115 (see Section 4.8.9).

10. Summary of semantic infringements by ETQE

Specific note should be taken by SAQA in regard to the high percentage, calculated as a percentage of the number of records submitted by the ETQE, for most ETQEs (see Section 4.8.10). Further, the implementation of specific interventions to address the highest percentages should be considered.

***P.8 Qualification/Unit Standard Registration***

***P.8.1 Qualification enrolments***

1. Start After, End After

- a. SAQA must confirm with the 4 ETQEs that have records in this category that they are aware that the qualifications in this category are no longer registered.

- b. If SAQA finds that the registration period of these qualifications is not disputed then the records in this category should be referred back to the submitting ETQE for confirmation that the data on the qualification enrolment record is correct.
2. Start After, End During
  - a. SAQA must confirm with the 5 ETQEs that have records in this category that they are aware that the qualifications in this category are no longer registered.
  - b. If SAQA finds that the registration period of these qualifications is not disputed then the records in this category should be referred back to the submitting ETQE for confirmation that the data on the qualification enrolment record is correct.
3. Start Before, End During

SAQA should refer the qualification enrolment records that have a start date which precedes the qualification registration start date by more than one year back to the submitting ETQE for confirmation that the data on the qualification enrolment record is correct.
4. Start During, End After

SAQA should refer the qualification enrolment records that have an end date which succeeds the last date of achievement for the qualification by more than one year back to the submitting ETQE for confirmation that the data on the qualification enrolment record is correct.
5. Start Before, End before

SAQA should refer all of these qualification enrolment records back to the submitting ETQE for confirmation that the data on the qualification enrolment record is correct.

### ***P.8.2 Unit Standard enrolments***

1. Start Before, End Before

SAQA should refer the unit standard enrolment records that have an start date which precedes the unit standard registration start date by more than one year back to the submitting ETQE for confirmation that the data on the unit standard enrolment record is correct.
2. Start After, End During

SAQA should refer the unit standard enrolment records that have an start date which succeeds the last date of enrolment for the unit standard by more than one year back to the submitting ETQE for confirmation that the data on the unit standard enrolment record is correct.

3. Start During, End After

SAQA should refer the unit standard enrolment records that have an end date which succeeds the last date of achievement for the unit standard by more than one year back to the submitting ETQE for confirmation that the data on the unit standard enrolment record is correct.

4. Start Before, End During

SAQA should refer the unit standard enrolment records that have a start date which precedes the unit standard registration start date by more than one year back to the submitting ETQE for confirmation that the data on the qualification enrolment record is correct.

5. Start After, End After

SAQA should refer all of these unit standard enrolment records back to the submitting ETQE for confirmation that the data on the qualification enrolment record is correct.

6. Start Before, End After

SAQA should refer all of these unit standard enrolment records back to the submitting ETQE for confirmation that the data on the qualification enrolment record is correct.

***P.9 Unit Standard based qualification achievements***

1. Insufficient Unit Standard Credits Achieved

- a. The accuracy of the qualification unit standard relationship should be confirmed by SAQA for qualification identifiers 49623, 49102, 23391, 23210, 24214, 22507 and 20513.
- b. Once the accuracy of the qualification unit standard relationship has been confirmed by SAQA (as per item a. above), all of the qualification enrolment records found in this group should be referred back to the submitting ETQE.
- c. The records that were found to have a cluster probability low enough to be considered anomalous (see Section 4.10.1) should be referred back to the submitting ETQE for confirmation that the data captured on the qualification enrolment record is correct.
- d. Specific note should be taken by SAQA in regard to the seemingly systemic issues in regard to the implementation of unit standard based qualifications for ETQE identifiers 1103, 1105, 1106, 1107, 1111 and 1119.

2. No Unit Standard Credits Achieved

- a. All of the qualification enrolment records found in this group should be referred back to the submitting ETQE.
  - b. The records that were found to have a cluster probability low enough to be considered anomalous (see Section 4.10.2) should be referred back to the submitting ETQE for confirmation that the data captured on the qualification enrolment record is correct.
  - c. Specific note should be taken by SAQA in regard to the seemingly systemic issues in regard to the implementation of unit standard based qualifications for ETQE identifiers 1105, 1106, 1108, 1111, 1112 and 1126.
3. Incorrect Mix of Unit Standard Credits Achieved
- a. All of the qualification enrolment records found in this group should be referred back to the submitting ETQE.
  - b. The records that were found to have a cluster probability low enough to be considered anomalous (see Section 4.10.3) should be referred back to the submitting ETQE for confirmation that the data captured on the qualification enrolment record is correct.
  - c. Specific note should be taken by SAQA in regard to the seemingly systemic issues in regard to the implementation of unit standard based qualifications for ETQE identifiers 1105, 1106, 1112, 1123 and 1126.

#### ***P.10 Data quality affinity***

##### **1. Learnership enrolments**

Specific note should be taken by SAQA in regard to the loading of missing learnership enrolment records for learnership identifier 1554 by ETQE identifier 1105. The ETQE has not submitted the related qualification enrolment records for this learnership (Section 4.11.1).

##### **2. Qualification enrolments**

Specific note should be taken by SAQA in regard to the loading of missing qualification enrolment records for qualification identifiers 20513, 17227 and 10471 by ETQE identifier 1105. The ETQE has not submitted the related unit standard enrolment records for these qualifications (Section 4.11.2).