

**Exploring the practice of quality control in the onscreen marking of
Ordinary Level Biology in Zimbabwe**

By

Ebba Masiri

Submitted in accordance with the requirements

for the degree of

Doctor of Philosophy

In the subject

Education

At the

University of South Africa

Supervisor: Prof. M.T. Gumbo

February 2020

Table of contents

Chapter 1

Orientation into the study

1.1	Introduction.....	1
1.2	The context of the Study.....	2
	1.2.1 Public examinations in Zimbabwe.....	2
	1.2.2 Paper based marking in Zimbabwe.....	4
	1.2.3 Information communication technology in education.....	7
	1.2.4 Introduction of onscreen marking in Zimbabwe.....	9
1.3	Statement of the problem	13
1.4	Research question.....	14
1.5	Aim and objectives of the study.....	15
1.6	Significance of the study.....	16
1.7	An overview of the research design and methodology.....	16
1.8	Delimitations of the study.....	18
1.9	Definition of key terms.....	18
	1.9.1 Onscreen marking.....	18
	1.9.2 Quality control.....	19
	1.9.3 Seeding	20
	1.9.4 Percentage double marking.....	21
1.10	Chapter outline.....	22

Chapter 2

Quality control in the marking of public examinations

2.1	Introduction.....	24
2.2	The concept of educational assessment.....	24
2.3	The purpose of assessment.....	27
2.4	A brief history of public examinations.....	29
2.5	Controversy surrounding examinations.....	32
2.6	Marking examination scripts.....	36
2.7	Examiner training.....	43
2.8	Standardisation of marking.....	51
2.9	Monitoring quality of marking.....	55
2.10	Examination questions and mark schemes.....	61
2.11	The conceptual framework.....	65
2.12	Conclusion.....	66

Chapter 3

The onscreen marking technology

3.1	Introduction.....	68
3.2	Development of e-assessment.....	68
3.3	Challenges of e-assessment.....	70
3.4	Overview of the OSM technology.....	75
3.4.1	The administrator module.....	76
3.4.2	The marker module.....	82
3.5	Global use of OSM.....	83
3.5.1	OSM in the United Kingdom.....	83

3.5.2	OSM in Hong Kong.....	88
3.5.3	OSM in China and other countries.....	96
3.5.4	OSM in Zimbabwe.....	97
3.6	Advantages of OSM.....	102
3.7	Challenges of OSM.....	104
3.8	Conclusion.....	105

Chapter 4

Research Methodology

4.1	Introduction.....	107
4.2	Location of the study.....	107
4.3	Constructivist paradigm.....	108
4.3.1	Ontology of the study.....	109
4.3.2	Epistemology of the study.....	110
4.3.3	Methodology of the study.....	111
4.3.4	Axiology of the study.....	112
4.4	Case study research.....	113
4.4.1	History of the case study research.....	113
4.4.2	Qualitative case study.....	116
4.4.3	Binding the case.....	117
4.4.4	Research sub-questions.....	121
4.5	The population.....	122
4.6	The sample.....	126
4.7	Data collection methods and instruments.....	128
4.7.1	Document Review.....	129
4.7.2	Interviews.....	132
4.8	Data presentation and analysis.....	144
4.9	Trustworthiness.....	147

4.10	Ethical considerations.....	151
4.11	Conclusion.....	156

Chapter 5

Data presentation and analysis

5.1	Introduction.....	158
5.2	Data processing and presentation.....	158
5.3	The Context of the OSM in Zimbabwe.....	159
	5.3.1 The assessment framework.....	159
	5.3.2 The technological infrastructure.....	164
	5.3.3 The human resource capacity.....	167
5.4	Capacity building of examiners	175
	5.4.1 Examiner recruitment and training.....	175
	5.4.2 SM training.....	181
	5.4.3 The standardisation process.....	184
5.5	Monitoring the quality of marking.....	190
	5.5.1 Qualification.....	190
	5.5.2 Seeds.....	192
	5.5.3 Dealing with stopped markers.....	199
	5.5.4 Permanently stopped markers.....	202
	5.5.5 Dealing with Suspect seeds.....	204
	5.5.6 Escalated problem scripts.....	208
	5.5.7 Examiner feedback.....	212
	5.5.8 Automatically generated reports.....	213
5.6	Test design issues.....	217
	5.6.1 Question paper structure.....	217
	5.6.2 Type of questions and mark schemes.....	224
5.7	Opportunities and challenges of quality control in OSM.....	233

5.7.1 Opportunities.....	233
5.7.2 Challenges.....	236
5.8 Summary of the findings.....	243
5.9 Conclusion.....	247

Chapter 6

Discussion of findings and proposed framework

6.1 Introduction.....	249
6.2 The context of OSM in Zimbabwe.....	249
6.2.1 The assessment framework.....	249
6.2.2 The technological infrastructure.....	254
6.2.3 Human resource capacity.....	256
6.3 Sub-question 1: Capacity building of examiners.....	258
6.4 Sub-question 2: Monitoring quality of marking.....	265
6.5 Sub-question 3: Influence of question papers and mark schemes....	271
6.6 Sub-question 4: Opportunities and challenges of quality control.....	275
6.7 The framework for quality control in OSM.....	279
6.8 Conclusion.....	281

Chapter 7

Conclusions and recommendations

7.1 Introduction.....	283
7.2 Influence of training and standardisation.....	283
7.3 Mechanisms of monitoring quality of marking.....	284
7.4 Test design issues.....	285

7.5	Opportunities and challenges of quality control in OSM.....	287
7.6	Limitations of the study.....	287
7.7	Conclusion.....	288
7.8	Recommendations.....	289
7.9	Autobiographical reflection.....	291
	List of references.....	293

List of Tables

Table 1.1	Factors influencing ICT in education in Zimbabwe.....	7
Table 2.1	General taxonomy of mark schemes.....	62
Table 3.1	Reasons why UK organisations adopted OSM.....	83
Table 4.1	Subjects components marked onscreen	118
Table 4.2	Interview participants.....	127
Table 4.3	List of documents reviewed	128
Table 4.4	Research sub-questions answered by data collection instruments.....	142
Table 4.5	Six-stage thematic data analysis.....	145
Table 5.1	Assessment objectives and skills for Biology (5008).....	160
Table 5.2	Weighting of assessment objectives.....	161
Table 5.3	Assessment scheme for Biology (5008).....	162
Table 5.4	Quality control activities in the OSM environment.....	167
Table 5.5	OSM training material provided to ZIMSEC.....	183
Table 5.6	Automatically generated reports.....	214
Table 5.7	Challenges and solutions: constrained and unconstrained papers.....	218
Table 5.8	Number of scripts marked by examiners.....	239

List of Figures

Figure 2.1	Bias in standardized tests.....	33
Figure 2.2	Hierarchy of examiners.....	50

Figure 2.3 The conceptual framework for studying quality control in OSM.....	66
Figure 3.1 The administrator module.....	77
Figure 3.2 Component settings in the administrator module.....	80
Figure 3.3 The marker screen.....	82
Figure 4.1 Target population of the study.....	125
Figure 6.1 A framework for quality control in the ZIMSEC context.....	280

Appendices

Appendix A Letter to the Director ZIMSEC.....	312
Appendix B Participant information sheet.....	314
Appendix C Consent form.....	317
Appendix D Document analysis form.....	318
Appendix E Interview schedule for subject managers.....	320
Appendix F Interview schedule for examiners.....	325
Appendix G Face-to-face interview transcription.....	326
Appendix H WhatsApp chats transcription.....	331
Appendix I Question paper review.....	343
Appendix J Mark scheme review.....	348
Appendix K Findings from documents: Question 1.....	353
Appendix L Parameter calculator guide review.....	361
Appendix M Editing certificate.....	365
Appendix N Ethical clearance.....	367

Declaration

Student number: 58526110

I declare that **Exploring the Practice of Quality Control in the Onscreen Marking of Ordinary Level Biology in Zimbabwe** 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

(Mrs. E. Masiri)

23 January 2020

DATE

Abstract

The purpose of this study was to explore the practice of quality control in the onscreen marking (OSM) environment of Biology (5008) examinations between 2013 and 2017. Examination marking is gradually being migrating from paper-based marking (PBM) to OSM in a bid to improve the efficiency and quality of marking. The Zimbabwe School Examinations Council (ZIMSEC) introduced OSM for some O Level subjects in June 2012, in a context characterised by a persistent economic crisis, patchy internet coverage, erratic power supplies and low digital literacy, among other challenges. The Council encountered some difficulties related to quality control, which triggered this qualitative instrumental single case study that was informed by the ontology, epistemology, methods and axiology of the constructivist philosophy. Data were collected through face-to-face and focus group interviews on the WhatsApp platform with 4 subject managers, 11 senior markers and 18 normal markers, and through document review. The findings of the study suggest that the quality of marking was influenced by the context in which the examinations were marked. The socio-political climate that prevailed in Zimbabwe impacted on the technological infrastructure for the OSM and the digital literacy of the examination personnel. The capacity of the examiners to work in the OSM environment was influenced by knowledge and skills transfer from training and standardisation to the live marking. The quality of marking was monitored by the seeds approach to script moderation, automatically generated reports and audit trails, and escalation of problem scripts. It was also influenced by the structure of the question papers, cognitive demands of the questions and mark schemes on the examiners, spaces provided for candidates' responses and mark scheme features such as language and marks to marking points ratio. The assessment framework provided by the syllabus guided the design and marking of Biology examinations. From these findings, a framework that could guide the practice of quality control in the OSM environment was formulated. OSM technology could enhance the quality of marking Biology examinations, thereby eliminating challenges associated with PBM. Some of the opportunities were, however, reduced by the challenges encountered during the OSM of the examinations. It is recommended that ZIMSEC put in place policies and procedures that could guide specific quality control activities in the OSM environment and establish computer centres in the provincial capital towns. The Council could also consider benchmarking examiner recruitment, training and standardisation procedures with international examination authorities.

Keywords: Examinations; Onscreen marking; Quality control; Seeding; Constrained; Unconstrained; Questions; Mark schemes; Senior markers; Normal markers; Subject managers

MUSUMO WETSVAKURUDZO

Tsvakurudzo ino yanga yakananga kuvandudza nharaunda yemakwenyero ebvunzo kubudikidza nemichina pachidzidzo cheBhayaraji, 5008 pakati pemakore a2013-2017. Vandudzo iyi iri kuuya zvishoma nezvishoma kubva pakukwenya pamapepa zvichienda mukukwenya nemichina (on screen marking :OSM) nechinangwa chekuda kukwenenzvera mhando yebasa rezvekukwenya . Bazi rebvunzo reZimbabwe School Examinations Council (ZIMSEC) rakavarura kukwenya kubudikidza nemichina (OSM) kubvunzo dzedanho reOdhinari revhuru muna Chikumi 2012 , mumamiriro anozivikanwa ematambudziko ezveupfumi, masaisai eindaneti asingavimbiki anouya zvigamba zvigamba , magetsi asingawanikwe nguva dzose, nezivo yezvemichina muvakwenyi isina kupararira pakati pezvimwe zvimhingamupini.

Kanzuru yezvebvunzo yakasangana nemamwe matambudziko ane chokuita nounaku hwezvemakwenyerwe ebvunzo hwakakonzero kuti paitwe tsvakurudzo ino yezveudzamu (*qualitative*) muchinzvimbo chiduku chakasarudzwa (*case study*) yaitungamirirwa nemaziviro evacho vanoona nezvekukwenya bvunzo (*interpretivist epistemology*) nemaziviro okuti chokwadi chinosiyana nekusiya kwenharaunda nokuti chigadzirwa chevarimukati mekukwenya (*constructivist ontology*), nekuumba mufungo kubudikidza neumbo huchabuda mutsvakurudzo (*inductive theory*) nenzira nezvinokosheswa nenharaunda mukuumba ruzivo (*constructivist philosophy*).

Umboo hwetsvakurudzo hwakawanikwa kubudikidza nebvunzurudzo ine udzamu padungamunhu nemumapoka nekupindurana padare reWatsiApu nevanotungamira zvidzidzo (*Subject managers*) vana , zvidza mune zvokukwenywa bvunzo gumi neumwe, nevamwewo vakwenyi gumi nevasere uye kuongorora magwaro.

Mamiriro ezvemagariro nematongerwo enyika muZimbabwe akava nechokuita nezvezvivakwa, midziyo yezvemichina yekukwenya (OSM) kubudikidza nemichina neunyanzvi hwezvemichina muvashandi vezvekukwenya bvunzo. Magonero evakwenyi pane zvenharaunda yekukwenya nemichina hunodyidzana neruzivo, unyanzvi kubudikidza nokudzidzira neyananiso inobva mukukwenya chaiko. Unaku hwemakwenyero hwaicherechedzwa kubudikidza nemwero (*approach*) wekungonyukura nekukandakanda muvakwenyi zvipenga zvemimhinduro zvinenge zvambokwenywa zvikatenderanwa zvibobzwa neunyanzvi (*seeds*). Kuti tinzwisise

maonero akapamhamha akasiyana-siyana pakufambiswa kwekukwenya kwakashandiswawo kukwenya kwedzokororo (*moderation*), magwaro engororo aibuda ipapo ipapo (*automatic reports*), nekuronda matsimba (*audit trail*) nokusimudza mhinduro dzevadzidzi dzairatidza kana kusumudza zvigozhero .Zvimwe zvakabatsira zvaive nechokuita nemamiriro emibvunzo, kunjereka kwaitarisirwa paudzamu, nemidonzvo yemakwenyero yebvunzo . Kwakatariswawo zvakare nzvimbo yekupindurira yaive yakapiwa vadzidzi nezvimwe zvine chokuita nemidonzvo zvakaita semutauro , urongwa nemagoverwo ezvibodzwa pamwe nenhungamiri (*assessment framework*) yemakwenyero inopiwa nebumbiro rezvidzidzwa yaitungamira urongwa hwemakwenyero ebvunzo dzeBhayoraji.

Kubva mune zvakabuda mutsvakurudzo, panogona kuwanikwa nhungamiri (*framework*) inogona kutungamira mwero wenharaunda younaku hwemakwenyero ane umhizha kubudikidza nezvemichina , hwakaronga kukwenya kubudikidza nemichina kunosimudzira unaku hwemakwenyero ebvunzo dzeBhayaraji nekupedza zvimhingamupini zvaive nechokuita nekukwenya kwepamapepa. Zvimwe zvakanakira kukwenya kubudikidza nemichina (OSM) zvisinei zvakaderedzwa nezvimhingamupini zvakasanganikwa nazvo pakukwenya kubudikidza nemichina.

Zvinokurudzirwa kuti bazi rinoona nezvebvunzo reZIMSEC riise pachena mitemo nematanho anobatsira kujekesa kuti pasimudzirwe zvine chokuita namanakiro enharaunda yezvekukwenya kubudikidza nokudzika mizinda yezvemakombiyuta mumadhorobha ematunhu. Bazi rebvunzo iri ngaritarise mwero wemapinziro evakwenyi, madzidzisirwo uye kuyanana matanho emakwenyero nemamwe mapazi ezvebvunzo epasi rose.

Mazwi akakosha: bvunzo, kukwenya kubudikidza nemichina, chipimo chemanakiro, kunyukura, kusununguka nekusungukana, mibvunzo, midonzvo yokukwenya.

Ngokufitjhazana/Ngokurhunyenziweko

Ihloso yaleli rhubhululo bekukuphenya indlela ikhwalithi elawulwa ngayo ngehlelo lokutshwaya ngekhomphyutha kwe enhlahlubo zeBhayiloji (5008) phakathi komnyaka ka-2013 no-2017. Ukutshwaywa kwe enhlahlubo kancanikancani kuyasuka ehlelweni lokutshwaya iphepha ngesandla (PBM) kuya ehlelweni lokutshwaya ngekhomphyutha (OSM) ngomzamo wokuthuthukisa umsebenzi omuhle kanye nokuletha iqophelo eliphezulu lokutshwaya. Hlangana nezinye iintjhihilo, UMkhandlu wezokuTshwaywa kwe eNhlahlubo eZimbabwe (*Zimbabwe School Examinations Council*) (ZIMSEC) sewungenise ihlelo le-OSM kwezinye iimfundo zesigaba sika-O Level ngenyangaka Mgwengweni 2012, ngaphasi kobujamo obumbibe zomnotho, kobujamo obumaratha be-inthanedi, obuqokeme kobokuphakelwa ngegezi kanye na ngaphasi kwezinga eliphasi lefundo ye dijithali. UMkhandlu uhlangabezane nobunye ubudisi obumalunga nanokulawulwa kwekhwalithi, okubujamo obukhwezelele isizathu sokobana kube nerhubhululo linye elisebenzako elisebenzisa indlela yerhubhululo yekulumo, kanti lokhu kwabangelwayi-ontholoji, i-ephistemoloji, iindlela zerhubhululo kanye ne-akziyoloji yefilosofi i-*constructivist philosophy*. Idatha ibuthelelwe ngendlela yehlolombono yokubuza umuntu ngamunye ubuso nobuso kanye nokubuza iinqhema zabantu ezinqotjihiwe kokukundla yezokucocisana, i-*WhatsApp platform* kanye nabaphathi beemfundo aba-4 subject managers, abatshwayi abakhulu abali-11, kanye nabatshwayi abajayelekileko abali-18, kanti lokhu kwenziwa ngokubuyekeza umtlobo. Ilwazi elifumane keerhubhululweni liphakamisa kobana izinga lekhwalithi lokutshwaya laba nomthintela wobujamo/wendawo lapho iinhlahlubo zatshwaywa khona. Ubujamo bezehlalakuhle yabantu kezepolotik iebebusezweni leZimbabwe laba nomthelela phezu komthanga lasisekelo wethekinoloji, kanti kwathinta abasebenzi behlelo le-OSM kanye nezinga lefundo yedijithali. Amandla wekghono labatshwayi lokusebenza ebhodulukweni le-OSM lalilawulwa kudluliselwa kwelwazi kanye namakghonofundwa ukusukela ekubandulweni kanye nokwenza izinto ngendlela efanako ehlelweni elibonakala ngamehlo lokutshwaya. Izinga lokutshwaya lalitjhejwe yindlela yokulinganiswa kwamaphepha atshwayiwako, ihlelo le-*seeds approach to script moderation*, kanti ihlelwe lingokwalo lihlanganisa imibiko begodu lilandelela ukuhlolwa, kanti goduli yakwazi nokuveza amaphepha ane miraro. Leli hlelo begodula lilawulwa sisakhiwo sephepha lemibuzo, lilwazi elifunekako ephepheni lemibuzo kanye namaskimu wamaksi phezu kwabatshwayi bamaphepha, iinkhala ezenzelwako bana abafundi baphendulele kizo kanye namaskimu wokutshwaya okunje ngelimi

kanye namamaksi asesilinganiswe nisamamaksi, phecelezi-*marking points ratio*. Isakhiwo sokuhlola sinikelwa yisilabhasi, okungiyo eyikombandlela yedizayini kanye nokutshwaywa kwe enhlahlubo zeBhayiloji. Ngalelilwazi elitholakeleko, kukghonakele ukuthi kutlanywe isakhiwo ebesingabayi kombandlela yendlela engalandelwa ukulawula ikhwalithi ebhodulukweni ye-OSM. Ithe kinoloji ye-OSM beyinga siza izinga lokumakha iinhlahlubo zeBhayiloji, ngalokho lokhu bekungaphungula iintjhijilo ezihlobene nehlelo le-PBM. Nanyana kunjalo, amanye amathuba, aphungulwazi intjhijilo ekuhlangabezenwe nazo nakutshwaywa iinhlahlubo zehlelo le-OSM. Kuye kwa tjhukunyiswa ukobana i-ZIMSEC izene mithethomgomo kanye ne enkambiso ezingabayikombandlela elayela imisebenzi ethile koyokulawulwa kwekhwalithi ebhodulukweni le-OSM kanye nokuhloma iinkhungo zekhomphyutha kumadorobhahloko we emfunda. UMkhandlu begodu ungatjheja yokubeka izinga lokuqatjhwa kwabatshwayi, lokubandulwa kanye nehlelolokwenza izinto ngendlela efanako neyamaziko we entjhabatjhaba alawula iinhlahlubo.

Amagama aqakathekileko: Iinhlahlubo, ihlelo lokuMakha ngeKhomphyutha, Ukulawulwa kwekhwalithi, isidingi, ukukhinyabezeka, ukungakhinyabezeki, Imibuzo, Ama Skimu wokuTshwaya.

Dedication

I dedicate this thesis to God Almighty, Jehovah, whose name is a strong tower, my light and my salvation. I thank you Lord for giving me the health and strength to go through this research. I also dedicate this work to my husband, Godfrey Masiri who has encouraged me all the way and whose encouragement has made sure that I give it all it takes to finish that which I have started, waking me in the wee hours of the night, and to my children, Kudzaishe and Tinotenda, who spurred me on. To my mother, Selina and my late father, Richard Mapokotera for sending their ‘girl-child’ to school. My love for you all can never be measured. God bless you all!

Acknowledgement

I am indebted and grateful to the following people who have contributed immensely to making this study possible:

I am particularly grateful to my supervisor, Professor Mishack T. Gumbo. I was inspired by the interest you showed in my work; the timely feedback; encouraging comments; and the follow-ups when I seemed to take too long to accomplish tasks. Thank you for your patient guidance and encouragement. I pray to God Almighty that He may grant you the desires of your heart and establish your plans.

I am thankful to Dr. Wilfred Mazani who kept nagging me to register for a PhD until I finally did. Thank you, Pastor! I am very grateful to Prof. T. Javangwe for taking his time to edit the language of this thesis – ndinokutendai, Musunda!

I am grateful to the Director of ZIMSEC, Dr. L. Nembaware for granting me the permission to conduct this study at the Council and to all the participants of this study for taking your time to share your experiences with me.

I say a big thank you to ‘Mai Titos’ family for the support, with special mention of Ebba Junior and Pride for helping me when I needed extra hands and computer assistance. Frank, I thank you for the support and encouragement.

To my colleagues and friends, Gamuchirai Charumbira, Elenia Javangwe, Grammar Chadya, Simbai Kapfunde and Charles Nyandoro, thank you for the support.

I pray that the Lord God of Heaven bless you all!

Chapter One

Orientation into the study

1.1 Introduction

The purpose of this study was to explore the practice of quality control in the marking of Ordinary Level Biology in the onscreen marking environment in Zimbabwe in order to propose a framework which can help to improve the practice. Examination marking is increasingly migrating from the paper-based marking (PBM) to onscreen marking (OSM), in a bid to improve the efficiency and quality of examinations. In OSM, candidates' scripts are scanned into the digital format and sent to the examiners for marking on computer screens via a secure system (Pinot de Moira, 2013; Coniam, 2011a; Roan, 2009; Fowles, 2011; Hudson, 2009). The Office of Qualifications and Examinations Regulation [Ofqual] (2013:3) posits that the examination boards should put in place the robust systems that promote the reliability of marking (quality marking), and to prevent and remedy poor marking when it occurs. Such systems include examiner training and standardisation, and monitoring of marking through script moderation (Ofqual, 2013; Hudson, 2009).

Quality control systems have also migrated from PBM to OSM, with literature showing that OSM enhances quality control by offering online training and frequent, flexible real-time script moderation mechanisms and monitoring of marking that automatically stops deviating examiners from marking until they receive further training (Data Research Services [DRS], 2013; Ofqual, 2013; Ramakrishna, Navya, Sri Harish, Swarna & Vasundhara, 2012; Hudson, 2009). Research shows that the reliability of marking is influenced by the type of training and standardisation examiners receive, the efficiency of the script moderation and marking monitoring system and the type of questions and marking schemes in the examination (Ofqual, 2014a; Tisi, Whitehouse,

Maughan & Burdett, 2013; Ofqual, 2013; Ahmed & Pollit, 2011). The Zimbabwe School Examinations Council marked twenty Ordinary Level subject components on screen, including Biology (5008) from 2012 to 2017, hence the need to study the quality control system in the onscreen marking of the subject.

Raikes, Greateorex and Shaw (2004:13-14) raised several research questions about quality control in the OSM environment. Some of the questions included the criteria to be used to select the answers for ‘seeds’ (pre-marked standard answers that are regularly presented to examiners to monitor marking accuracy) or double marking; the roles of senior examiners in the new quality control system; the feedback that should be provided to the examiners, when and how it should be communicated, and many others. Most research studies in OSM have focused on the comparison of scores and marker behaviours in PBM and OSM, and the attitudes of examiners towards OSM (Coniam & Yan, 2016; Yan & Coniam, 2014; Conium, 2013; Johnson, Hopkin, Shiell & Bell, 2012a; Johnson et al, 2012b; Coniam, 2011a; Johnson, Nádas & Bell, 2010; Coniam & Yeung, 2010; Conium, 2009). A few researches have focused on the practice of quality control in a live onscreen marking environment (Ofqual, 2014a; Pinot de Moira, 2013; Johnson & Black, 2012; Fowles, 2011), with Pinot de Moira (2013:4) lamenting the lack of evidence that supports the benefit of OSM quality control in live examinations. This study seeks to fill this gap by exploring the practice of quality control in the marking of Ordinary Level Biology in the onscreen marking environment in Zimbabwe in order to propose a framework which can help to improve the practice.

1.2 The context of the study

1.2.1 Public examinations in Zimbabwe

The Zimbabwe School Examinations Council (ZIMSEC) was established by an act of the Zimbabwe Parliament, Number 17 of 1994, and started operating in 1996. The act enumerates several functions for the council, which include to:

- organise and conduct examinations in subjects that form part of a course of primary or secondary education as the minister may in writing direct;
- consider and approve subjects for examinations;
- appoint panels or boards of examiners;

- approve and register examination centres;
- review rules and regulations relating to examinations;
- confer or approve the conferment of certificates, diplomas and other awards to persons who have passed examinations;
- enter into arrangements, whether reciprocal or otherwise, with persons or organisations inside or outside Zimbabwe for the recognition of certificates, diplomas and other awards granted in respect of examinations organized or conducted by the council;
- do all things necessary to maintain the integrity of the examination system in respect of primary and secondary education in Zimbabwe; and
- do any other thing that the council may be required to do by or under this act or any other enactment.

(ZIMSEC, 2013; Act Number 17 of 1994:68-69).

Before 1996, the school examinations were set and marked by United Kingdom-based examinations boards, with the Examinations Branch of the Ministry of Education managing the administration of the examinations in Zimbabwe. When localisation started, the University of Cambridge Local Examinations Syndicate (UCLES) assisted ZIMSEC to take over (ZIMSEC, 2013; Musarurwa & Chimhenga, 2011). When the localisation of examinations was completed ZIMSEC took over the responsibility to assess the primary and secondary education in Zimbabwe. The council sets, administers, marks and grades all examinations in Zimbabwe, although there are a few private schools that still prefer United Kingdom to ZIMSEC examinations (Mashoko, Mateveke, Kufakunesu & Mashoko, 2013: 463).

The Zimbabwe education system comprises nine years (from early childhood development [ECD] through Grades 1 – 7) of primary and six years (Forms 1 – 6) of secondary school. The six years of secondary school are divided into four years of Ordinary (O) Level and two years of Advanced (A) Level courses, examined after four and two years respectively (Ministry of Primary and Secondary Education [MOPSE], 2015; Southern Africa Association for Educational Assessment [SAAEA], 2014). As mandated by Act Number 17 of 1994, the Zimbabwe School Examinations Council sets and administers school examinations in Zimbabwe. The primary school examinations are at Grade 7, which is the end of the primary course (MOPSE, 2015;

SAAEA, 2014). Grade 7 learners currently sit for five subjects, Mathematics, English Language, General paper, Agriculture and an Indigenous Language. The Indigenous Languages examined at Grade 7 in the old curriculum that will be last examined in October 2020 are Shona, Ndebele, Tonga, Nambya, Tshivenda, Xichangana, Kalanga and Sesotho (MOPSE, 2015; personal experience). The Ordinary Level examinations are written at the end of Form 4, with ZIMSEC examining 37 subjects that translated to 93 examination papers, while advanced Level examinations are written at the end of Form 6, and the council examined 23 subjects made up of 74 examination papers in the old curriculum that was last examined in June 2018 (SAAEA, 2014; ZIMSEC Entry Procedure Booklet, 2008-2012; Question Paper Development Record, 2017).

The primary and secondary school assessments have been characterised by the public examinations only. The Ministry of Primary and Secondary Education reviewed the Zimbabwe curriculum in 2015 and came up with a curriculum framework that guides teaching, learning and assessment until 2022. The new curriculum was first implemented in January 2017 and was examined for the first time in November 2018 at O and A Level as mentioned earlier. The first Grade 7 examination will be sat by the 2017 Grade 3 cohort, in 2021. In addition to public examinations; the new curriculum introduced continuous assessment that contributes 30% of the final mark at Grade 7, Ordinary and Advanced levels (ZIMSEC, 2017; MOPSE, 2015). The continuous assessment has however faced numerous challenges that are outside the focus of this study. The Ordinary Level subjects components that were marked onscreen were in the old curriculum and were replaced by new syllabi; hence the need to study the practice of quality control in the OSM environment to propose a framework that could guide the practice in future examinations. Before 2012, all examination scripts were marked on paper.

1.2.2 Paper-based Marking in Zimbabwe

Before 2010, ZIMSEC was using the traditional paper-based marking system where a single examiner would mark candidates' scripts from an examination centre and award marks, using the whole script marking approach (Ngara & Ngara, 2013; Bukenya, 2006; ZIMSEC, 2003). The marking was supervised by the team leaders who moderated selected scripts to check adherence to the mark scheme and asked deviating examiners to remark their scripts (Risiro, 2014;

ZIMSEC, 2003). Research has indicated that whole script marking lowers the marking reliability due to the examiner bias emanating from the halo effect, examiner competences, interruptions during marking, transition from very bad to very good scripts and vice versa (Ofqual, 2013; Chinamasa & Munetsi, 2012; Pinot de Moira, 2011). In a research study, Mashoko et al (2013:468) established the existence of concerns about examiner bias in the marking of public examinations in Zimbabwe. Examination candidates who participated in the research expressed concern that day secondary and boarding secondary schools were treated differently during the marking of examinations. Examiners who participated in the research acknowledged that there were good centres and bad centres in examinations but mark schemes would solve the issue (Mashoko et al, 2013:468).

In a literature review on item level marking, Ofqual (2014a:3-5) confirmed the examiner bias that exist in whole script marking and that item level marking indeed eliminates the bias. The concerns of Zimbabwean candidates about examiner bias may, therefore, be genuine. The formal examinations are considered as high stakes assessments because they determine the fate of candidates and should therefore be conducted in a manner that allows a measure of transparency and scrutiny to ensure reliability, validity and fairness so as to gain public confidence (Isaacs, Zara, Hebert, Coombs & Smith, 2013; Hill, 2013). An examination board can, therefore, not ignore concerns about examiner bias in whole script paper based marking, lest they lose credibility.

In addition to the concerns about examiner bias in whole script marking, there is the need for the senior examiners to monitor the quality of marking. A team leader is required to sample and moderate at least ten percent of scripts in an envelope to ensure adherence to the mark scheme and hence, consistency and accuracy of marking (Mashoko et al, 2013; Bukenya, 2006). Several disadvantages of monitoring marking in the PBM environment have been enumerated. It is the examiner who selects scripts for the team leader, at fixed intervals, resulting in possibilities of examiners being thorough when marking scripts for moderation; the examiners continue to mark in the intervening periods, without feedback from team leaders; and there are logistical challenges of moving scripts between examiners and team leaders, limiting the frequency of script moderation (Johnson & Black, 2012; Hudson, 2009). These shortcomings of whole script

PBM could have prompted ZIMSEC to introduce conveyor belt marking, which is paper-based as well, in 2010.

The conveyor belt marking (CBM) system is where examiners are organised into groups and each marker assigned questions to mark, resulting in a candidate's script being marked by several examiners (Ngara & Ngara, 2013; Chinamasa & Munetsi, 2012; Mwanyumba & Mutwiri, 2009). Research has shown that CBM is more reliable than the whole script marking (Chinamasa & Munetsi, 2012; Bukenya, 2006). Ngara and Ngara (2013:33) gathered the opinions of ZIMSEC markers on the CBM system, who enumerated the following advantages: it is candidate friendly, as the script is marked by more than one marker; enhanced marker efficiency; it is easier to internalise the mark scheme; there are possibilities of obtaining reliable marks; it is easier to supervise marking and moderate scripts. The examiners, however, highlighted the challenges that emanated from CBM, which include painful delays by slow examiners, resulting in conflicts among team members; pressure is created on slow markers as they fail to cope with speed of team members; lower remuneration than in whole script marking; more work and pressure that is created by re-allocation of questions (Ngara & Ngara, 2013:38).

Mashoko et al (2013:462) also gathered the views of ZIMSEC examiners about CBM and established that participants believed CBM is more reliable than whole script marking. The participants, however, expressed concerns that CBM was time consuming; reduces competition among examiners; leaves no room for examiners to relax and unnecessarily confines them. The advantages and disadvantages of CBM were confirmed by Chinamasa and Munetsi (2012:188) when they studied the reliability of the specialised marking of examination questions at Chinhoyi University of Technology (CUT) and emphasised that challenges had more to do with organisation than with the actual marking. The organisation of the script moderation in CBM also requires sampling, which frequency might be limited by the pressure to meet deadlines and group dynamics mentioned by Ngara and Ngara (2013:38). These challenges of CBM and the quest to improve marking efficiency and quality could have prompted ZIMSEC to introduce some O level subject components to onscreen marking. The technology was adopted in a context with limited use of information communication technology (ICT) in the education sector.

1.2.3 Information Communication Technology in education

A survey on of the ICT and education in Africa identified enabling and constraining factors that influenced adoption of ICT in education in Zimbabwe. Shafika (2007:7) summarises the factors as presented in Table 1.1. The survey was conducted at a time when Zimbabwe was experiencing an economic crisis characterised by hyper-inflation, which stood at 100% in 2006, with over 50% of the population leaving on less than a dollar a day (Shafika 2007:2). The economic crisis could have further constrained the adoption and use of ICT in general and in education in particular, by depleting the resources needed to implement ICT policies. The national ICT policy cited in the survey was crafted in 2005 and was in place until 2015 when the government, through the Ministry of Information Communication Technology, Postal and Courier Services (MICTPCS) crafted a new national ICT policy in 2015.

Table 1.1: Factors influencing adoption of ICT in education in Zimbabwe (adapted from Shafika 2007:7)

Enabling factors	Constraining factors
An ICT policy which includes references to ICT in education	No specific national policy on ICT in education
Dedicated champions for ICT within the civil society sector	Limited human resource capacity
National policy promotes the idea of developing an ICT infrastructure that includes a local industry	Limited fiscal resources committed by government to support ICT access and use
Government leaders and civil society demonstrated an enthusiasm and positive attitudes towards ICT development in general and in education in particular	Little digital education content based on the local curriculum frameworks available in educational institutions.

Although the country had a national ICT policy in place since 2005, its pronouncements remained on paper, hampered by persistent economic challenges that have bedevilled the country for more than ten years. The same challenges cited by Shafika (2007), and many more, still needed to be addressed by the next national ICT policy. The following challenges to ICT access and use were enumerated in the National ICT Policy (2015:13-15):

- Inadequate communication infrastructure: High speed broadband coverage is still patchy, with most rural and remote areas remaining uncovered;
- Inadequate electric power: The national power grid does not cover the whole country, with a significant population resorting to alternative power sources that are expensive. Those who are on the national grid experience erratic supply;
- Inadequate investment capital: The high perceived country risk has resulted in higher landing rates for foreign borrowing. The current liquidity crunch has made it impossible to secure funding for ICT projects, adversely affecting ICT infrastructure development and growth in Zimbabwe;
- Low digital literacy: The education curriculum does not include ICT, resulting in low ICT uptake and usage;
- Absence of an internet governance framework to deal with national and international internet traffic. This has resulted in high internet tariffs.

The National ICT Policy (2015:13) made pronouncements to address these challenges, a task that might not be accomplished in the near future, given the state of the economy. The challenges evidently persisted through the period that was studied, 2013 to 2017, up to the time of writing this report. The Zimbabwe Council of Churches (ZCC) reports that more than 80% of households in Zimbabwe are living in dire poverty and have now adopted various coping strategies such as borrowing, cutting on food expenses, assistance from relatives and friends, cutting on health and education expenses. The ZCC called on members of parliament to fight for better living conditions and salary adjustments (Langa, 2019:4). Literature shows that the economic woes in Zimbabwe persisted from 2006 to 2019 (Shafika, 2007; Langa, 2019). The persistent economic challenges could have impacted negatively on the implementation of the ICT policy and the OSM technology in Zimbabwe. In the United Kingdom, examiners mark from their homes (Raikes et al, 2004:20), an arrangement that would not be possible in Zimbabwe where internet coverage is patchy. The high internet charges would impact on time frames for training, standardisation and the actual marking, hence quality control. The erratic power supply would possibly disrupt marking and quality control activities. The low digital

literacy could compromise examiner competencies in the OSM environment. There was, therefore, need to study quality control in the OSM environment in the Zimbabwean context.

The low digital literacy could be addressed by the new curriculum which introduced ICT education from ECD to A Level (MOPSE 2015; ICT Syllabi 2015-2022). The National ICT Policy (2015:14) made commitments to promote the ICT skills development and to increase ICT use in primary and secondary schools through enhanced teaching and learning. One of the goals about education was to promote e-learning and use of e-materials throughout Zimbabwe (National ICT Policy, 2015:13-15). However, Zimbabwe does not have a policy that guides ICT in education. Calls have been made for the country to craft a policy to guide ICT in teaching and learning (Chindaro, 2013; Shafika, 2007). Ridgeway, McCunsker and Pead (2004:5) posit that there is an intimate relationship between teaching, learning and assessment. An ICT policy in education would, therefore, guide the use of technology in assessment as well. As mentioned earlier, the OSM technology was adopted at a time when there was limited use of ICT in general and in education in particular, hence the purpose of this study, to explore the practice of quality control on the marking of Ordinary Level Biology in the onscreen marking environment in Zimbabwe in order to propose a framework which can help to improve the practice.

1.2.4 Introduction of onscreen marking in Zimbabwe

The ZIMSEC introduced onscreen marking of candidates' scripts at Ordinary Level in the June 2012 examination session, beginning with two subject components, Mathematics Paper 1 and Integrated Science Paper 3 (DRS, 2013; Kachere, 2012; Karombo, 2012; personal experience). The questions for the two components demanded short answers and spaces were provided on the question papers, where candidates wrote their responses, even before onscreen marking. Such examinations are constrained, and seeding is the approach to quality control (DRS, 2015; 2013; Hudson, 2009). Seeding was therefore used for both Mathematics and Integrated Science. More O Level subject components were gradually added to OSM and a total of 20 were marked onscreen in the November 2016 examination session (Examination Circular Number 10 of 2015; Examination Circular Number 8 of 2015; Examination Circular Number 42 of 2013; Examination Circular Number 41 of 2013). It was reported that ZIMSEC was the first

examination authority to use OSM in Africa (Coniam & Yan, 2016; Kachere, 2012; Karombo, 2012) and challenges specific to the context were encountered.

Commerce Paper 2 was marked onscreen for the first time in June 2013 (Examination Circular Number, 42 of 2013; personal experience). The paper was a free response examination, where candidates were provided with separate answer booklets that were scanned for marking on screen. Hudson (2009:5) calls them unconstrained papers and posits that percentage double marking is the appropriate approach to quality control for unconstrained examinations. Some of the questions in the paper required short objective answers worth two or three marks, while others demanded longer answers worth up to ten marks. After marking a specified number of scripts, fast examiners were stopped from marking so that their scripts could be presented to another marker (DRS, 2013:94). This retarded the marking pace and examiners got frustrated by frequent stoppages. A decision was taken to use seeding for short answer questions, combined with double percentage marking for longer responses, in the November 2013 Commerce paper 2, hoping to increase the marking pace (personal experience).

The November 2013 Commerce Paper 2 scripts were marked in January 2014 and the OSM technology presented challenges that threatened to delay the release of the results (Dube, 2014; NewsdzeZimbabwe, 2014; personal experience). One of the challenges was associated with the seeding approach to quality control. Some examiners were presented with a series of seeds and a few actual scripts. Such examiners were effectively marking seeded scripts and a few live scripts, retarding the marking pace that the council intended to quicken. DRS (2015; 2013) and Hudson (2009:5) concur that seeding works well with constrained papers, where answers are short and candidates write their responses on spaces provided on the question paper. According to DRS (2013:51), a single paper can use both seeding and percentage double marking approaches to quality control. Why then did the seeding approach fail in the Commerce paper, when it was used to control marking of short answers? Could it be possible that the system was confused by using seeding for short answers written on separate answer sheets? How does the question paper structure influence quality control in OSM? What type of questions and mark schemes enhance quality control in the OSM environment? The OSM software provider offered remote support and marking was completed (personal experience). A decision was made to provide answer

spaces on question papers for all subjects that would be subsequently marked on screen (Examination Circulars Number 10 of 2015; Examination Circulars Number 8 of 2015).

Ordinary Level Biology (5008) components, Papers 3 and 4, were marked onscreen for the first time in November 2013 (Examination Circular Number 42 of 2013). The two papers were constrained even before OSM. Paper 3 was a practical component, whereas Paper 4 was an alternative to practical and candidates sat for either of the papers (O Level Biology 5008 Syllabus, 2011-2020:5). Paper 3 was returned to PBM in the next examination session, after examiners had argued that the nature of the mark scheme required them to see the name of the examination centre and to mark whole scripts, so as to award appropriate marks to the candidates (personal experience).

Literature suggests that examinations can be marked onscreen either as whole papers, or they can be segmented into individual questions and marked at item level (Ofqual, 2014a; 2013). Whole script marking is accused of increasing examiner bias emanating from the halo effect, examiner competence, interruption during marking, transition from very bad to very good scripts and vice versa (Ofqual, 2014a; Tisi et al, 2013; Pinot de Moira, 2013; Chinamasa & Munetsi, 2012), compromising the credibility of examinations. Ofqual (2014a:4) cites a United Kingdom examination authority that marked an examination on screen using the whole script approach, later marked at item level and reverted to whole script marking. Examination boards probably ignore the effects of examiner bias in order to preserve the validity of the examination. Onscreen marking could ease logistical challenges involved in PBM and improve efficiency of the marking process (DRS, 2015; Ofqual, 2013; Fowles, 2011). The new curriculum eliminated the alternative to practical paper (Paper 4) and maintained the practical component that was returned to PBM (O Level Biology 4025 Syllabus, 2015-2022:42). Marking the practical paper on screen could ease logistical challenges associated with paper-based marking. There was, therefore, need to study the practice of quality control and propose a framework that could inform the marking of the practical component in the new curriculum.

Biology Paper 2 of the same syllabus (5008) was marked onscreen for the first time in the November 2015 examination series. Before OSM, candidates wrote responses to section B of the

paper on separate answer sheets that were provided by examination centres. Beginning 2015, candidates wrote their answers on spaces provided on the question paper (Examination Circular Number 8 of 2015). Eight examiners abandoned the marking exercise, arguing that the paper was difficult to mark on screen (personal experience). The examiners were probably not adequately trained to mark in the OSM environment. Bhawani (2004:591) defines transfer of training as the extent of retention and application of the knowledge, skills and attitudes from the training environment to the workplace environment. In other words, transfer of training is the degree to which trainees effectively apply the training context to the job. Wajdi, Khalil and Maria (2014:16) posit that training can only increase effectiveness if it is effectively designed, delivered and transferred to the job. They further argue that transferring of training is the only way in which training influences organisational-level outcomes. Grossman and Salas (2011:103) posit that transfer of training is related to trainee characteristics such as cognitive ability, self-efficacy, motivation and perceived usefulness of training; training design and the work environment. It could be possible that the training and standardisation had not been transferred effectively to the live marking due to examiner incompetency, poorly designed training or a marking environment that was not conducive to the task.

Literature shows that the OSM quality control system identifies examiners who are not marking within the agreed standards and stops them from marking until they receive further training (DRS, 2015; 2013; Pinot de Moira, 2013; Hudson, 2009). Examinations are marked to meet specific timelines (Ofqual, 2013; personal experience), exerting pressure on both examiners and their supervisors, to mark and meet the deadline. High internet tariffs may force the examination council to set short training, standardisation and marking periods. Standardisation meetings may be hurried, leading to inconsistent application of mark schemes by examiners, hence the frequent stoppages. Those who are supposed to train the stopped examiners may not do so and instead allow them to continue marking so as to meet deadlines. Erratic power supplies may also disrupt marking and quality control activities, leading to examiner frustrations. What mechanisms have been put in place to ensure adherence to mark schemes and to enforce the re-training of stopped examiners?

In a research on monitoring of marking quality in the OSM environment, Johnson and Black (2012:400) established that discrepancies between examiners' marks and definitive marks on seeds emanated from (i) the examiner awarding a wrong mark, (ii) grey areas in the mark schemes, where the answers were open to examiner interpretation and (iii) wrong marks on seeded scripts. As team leaders in the research monitored examiners, some gave feedback that guided examiners throughout the marking period, while others gave meaningless feedback that left examiners confused. Other team leaders did not give any feedback to examiners throughout the marking process, resulting in examiners feeling isolated. It is therefore possible that the eight examiners who abandoned the marking of Biology Paper 2 did not receive guidance from their team leaders during marking, or they received meaningless feedback. There was therefore need to study the monitoring of marking by senior examiners in the OSM environment and how mark schemes influenced quality control in Zimbabwe.

Research shows that questions that elicit short responses are marked more accurately than questions that elicit long answers (Ahmed & Pollit, 2011; Johnson & Nádas, 2008). Biology Paper 4, made of short answer questions, had been marked on screen since November 2013 and there were no reported challenges emanating from the OSM quality control system. Biology Paper 2 was made of two sections, A and B. Section A was made of short answer questions worth one or two marks, and section B had long response questions worth more than five marks. The challenges encountered with Paper 2 in 2015 were not reported in the June and November 2016 examination sessions. Could it be that questions on Paper 2, Section B, had been shortened so that they elicit short responses that are easier to score? Ramakrishna et al (2012:15) and Roan (2009:7) concur that the advantages of an onscreen marking technology would be valuable if it supports assessment practices and principles of the examination body. There is a risk that examinations could be designed to suit the demands of technology, compromising validity of the results. It is against this background that this study was conducted to explore the practice of quality control in the OSM environment in the case of O level Biology (5008) examinations.

1.3 Statement of the problem

Examinations' marking is gradually migrating from PBM to OSM in a bid to improve the efficiency and quality of marking (Ofqual, 2013; Ramakrishna et al, 2012; Roan 2009). The

ZIMSEC introduced some O Level subjects to OSM in June 2012, in a context characterised by a persistent economic crisis, patchy internet coverage, erratic power supplies and low digital literacy among other challenges (Kachere, 2012; Karombo, 2012; personal experience). The technology allows real-time monitoring of marking, where deviating examiners are detected and stopped from marking specific questions until they are re-trained (DRS, 2015; 2013; Pinot de Moira, 2013; Hudson, 2009). Quality of marking is however determined by the quality of training and standardisation before marking, the efficiency of the mechanism of monitoring quality of marking and the nature of questions and mark schemes in the examination (Ofqual, 2014b; 2013; Johnson & Black, 2012; Ahmed & Pollit, 2011; Johnson & Nádas, 2009).

Quality control challenges were encountered in the onscreen marking of Commerce for the June and November 2013 examination session, where examiners were frequently stopped from marking specific questions, delaying the progress of marking (Personal experience; O Level Biology 5008 Papers 3 and 4 were marked on screen in the November 2013 examination session (Examination Circular Number 42 of 2013). Paper 3 was returned to PBM, with examiners arguing that they needed to mark whole scripts and not items (personal experience). Biology Paper 2 was marked onscreen in November 2015 and eight examiners abandoned the marking exercise, citing challenges in the marking of the paper (Examination Circular Number 8 of 2015; personal experience). There was therefore a need to explore the practice of quality control as influenced by examiner training and standardisation, mechanisms of monitoring the quality of marking, questions and mark schemes in the OSM environment.

The statement of the problem led to the research question and sub-questions stated in 1.4 followed by the aims and objectives in 1.5.

1.4 Research question

Main research question

How does the practice of quality control influence the marking of O Level Biology in the OSM environment in Zimbabwe?

Sub-questions

1. How does training of examiners and standardisation activities influence the quality of marking O Level Biology in the onscreen marking environment?
2. How is the quality of marking O Level Biology monitored in the OSM environment?
3. How do O Level Biology examination questions and mark schemes inform quality control in the OSM environment?
4. What are the opportunities and challenges of quality control in onscreen marking of O Level Biology?
5. How can quality control in the OSM of O Level Biology examinations be framed to provide guidelines for its practice?

1.5 Aim and objectives of the study

This study aimed to explore the influence of the practice of quality control on the marking of Ordinary Level Biology in the OSM environment in Zimbabwe in order to propose a framework which can help to improve the practice. This aim was accomplished by studying the training and standardisation procedures, mechanisms of monitoring quality of marking and the nature of examination questions and mark schemes used to mark O Level Biology in the OSM environment in Zimbabwe.

The objectives of the research were to:

1. Examine the influence of examiner training and standardisation activities on the quality of marking O Level Biology in the onscreen marking environment.
2. Investigate the mechanisms of monitoring quality in the OSM of O Level Biology.
3. Examine the influence of examination questions and mark schemes on quality control in the OSM of O Level Biology.
4. Identify the opportunities and challenges of quality control in the OSM environment.
5. Develop and propose a framework that guides the practice of quality control in marking O Level Biology in the OSM environment.

1.6 Significance of the study

In light of the problems identified in this study, a crucial contribution was eminent. This study contributed a framework that could guide quality control in the OSM of O Level Biology specifically and of similarly structured subjects in general, as well as literature on quality control on the OSM platform in Zimbabwe. The study highlighted operational opportunities and challenges of quality control in the OSM of O Level Biology in Zimbabwe. Since Zimbabwe is the initiator of onscreen marking of examinations in Africa, a study of this nature can be helpful in terms of exposing the quality control issues that this move has attracted. The study thus conscientises the Council and its parent ministry in the country of these issues, which can trigger actions to address them guided by the findings of the study. The framework that the study contributed could help direct such actions. It is also envisaged that other contexts that will implement the onscreen marking of examinations will benefit from the contribution of the study. The study will also attract other studies that will explore this phenomenon further.

1.7 An overview of the research design and methodology

The process of marking O Level Biology (5008) examinations, from 2013 to 2017, was selected as an instrumental case (Starman, 2013; Creswell, 2009; Baxter & Jack, 2008; Stake, 2005) that was studied to understand the phenomenon of quality control in the OSM environment. The study adopted the single case study design, guided by the constructivist worldview which assumes that reality is a subjective construct and researchers seek to understand the people's idea of reality (Yazan, 2015; Tracey, 2013; Starman, 2013; Creswell, 2009). This study, therefore, adopted the relativist ontology of the constructivist paradigm (Charmaz 2017; Lal, Suto & Ungar, 2012; Creswell, 2014; 2007; Guba & Lincoln, 1994) and assumed that examiners and subject managers held and shared multiple realities about quality control in the OSM environment, created during the marking of O Level Biology (5008) examinations. I sought to understand how these officials interpret the OSM quality control, the factors that influence their interpretations and how their interpretations of quality control vary with experiences, time and context. The constructivist epistemology guided the qualitative research design.

Purposive sampling strategies were used to select the case and participants. Typical case sampling was used to select O level Biology (5008) out of the 20 subjects marked on screen from

2013 to 2017. Expert and stratified purposeful sampling strategies were used to select the participants who made the accessible population of the study, who were subject managers, senior markers and normal markers (World Health Organisation [WHO], 2017; Creswell, 2009; Ponelis, 2015; Palyis, 2008). Four subject managers out of the 6, 11 out of 13 senior markers and 29 out of 45 normal markers who participated in the marking of O Level Biology examinations (Paper 2 and Paper 4) in 2017, were selected and interviewed face-to-face and by WhatsApp focus groups (Shahid, 2018; United Nations Development Programme ([UNDP], 2018; Oltman, 2016; Bolderston, 2012). A total of 54 documents that were generated for OSM were selected, described in detail in Table 4.2, and reviewed (Triad, 2016, Ahmed, 2010; Bowen, 2009). The results were reported according to themes that emerged (Nowell, Morris, White et al, 2017; Creswell, 2014; 2009), and summarised into a framework that explains the practice of quality control in the OSM environment.

In accordance with the constructivist axiology (Charmaz, 2017; Creswell, 2014; Starman, 2013; Creswell, 2007), I declared my personal background and experiences that might have influenced the findings of this study. I brought into the study my experiences and values acquired from teaching Biology in Zimbabwean schools and assessment at tertiary institutions. I also bring my experiences as an examiner and subject manager for O and A Level Biology (5008) and Biology (9190) respectively with ZIMSEC, experiences that have given me considerable knowledge about quality control in both paper-based and onscreen marking. These experiences influenced selection of the case and my interpretation of the data and the framework developed from this study as explained in the relevant sections. The credibility and trustworthiness of the study was assured by a clear definition of the population using the systematic and organised specification (SOS) put forward by Asiama, Mensah and Oteng-Abayie (2017:1607-1621); triangulation of sources and member checking (Korstjens & Moser, 2018; Nowell et al, 2017; Yazan, 2015:150).

Ethical clearance was first applied for and granted by the College of Education at the University of South Africa in which the study was housed (see Appendix N), through the assistance of my supervisor. Official access to the ZIMSEC and informed consent (Sabar and Ben-Yehoshua, 2017; Farooq & de Villiers, 2017; Maramwidze-Merrison, 2016; Dawson, 2007) were sought from the Director and participants respectively (see Appendix A, B and C). A letter was written

to the director of ZIMSEC (see Appendix A), who granted access in writing. I personally visited the subject managers in their offices to invite them to participate in the study as guided by Farooq and de Villiers (2017:14), who posit that pre-interview conversations allow the researcher to address participants' concerns, create interest, build rapport and explain the interview style. I accessed the examiners on two WhatsApp groups (see Appendix H) through the subject manager for O Level Biology (4025).

1.8 Delimitations of the study

Single case studies can be analytically generalised to the context when the studied case has typical characteristics and a theoretical or conceptual framework has been used to collect and interpret data (Starman 2013; Yin 2003). The study was, therefore, delimited to the quality control process in the marking of O Level Biology that were marked from 2013 to 2017 and other examinations that were marked on screen by the ZIMSEC. I did not intend to generalise the results of this study to contexts other than Zimbabwe.

1.9 Definition of key terms

1.9.1 Onscreen marking

Onscreen marking is a marking technology first introduced to mark examinations for general qualifications by Pearson Edexcel in 2003. In onscreen marking, candidates' scripts are scanned into digital format and sent to examiners for marking on computer screens, via a secure system (Ofqual, 2013:11). Fowles (2011:6) describes some onscreen marking technologies where scripts are distributed to examiners over the internet to remote locations, where the examiners access and mark the scripts on a computer screen and Geranpayeh (2011:16) calls it online marking or onscreen marking. Fowles (2011:7) describes another onscreen marking technology that does not depend on the internet, but allows complete scanned scripts to be downloaded to compact discs (CDs), for distribution to examiners to mark at home. Some scholars refer to onscreen marking as e-marking (Pinot de Moira, 2013; Fowles, 2011).

The terms onscreen marking, e-marking and online marking have been used by researchers (Pinot de Moira, 2013; Geranpayeh, 2011; Fowles, 2011; Coniam, 2011; Coniam, 2010; Coniam, 2009; Raikes et al, 2004) to describe a marking system where candidates scripts are scanned and

their images are distributed to examiners for marking on computer screens. This study, therefore, adopted the term onscreen marking as used by these researchers. The terms e-marking and online marking have the same meaning as onscreen marking, whenever they are used in this study. Although the mode of script distribution, internet or disks, does not seem to change the concept of onscreen marking, this research studied quality control in the marking of scripts distributed over the internet because that is the mode of distribution that was used in Zimbabwe.

1.9.2 Quality control

Reliability is defined as the extent to which test scores are accurate, consistent, stable and reproducible (Thompson, 2016; Race, 2009). Ofqual (2013:3) defines quality of marking as the accuracy and reliability of marking. Rust (2002:2) describes a reliable assessment as one in which several examiners working independently using the same criteria and mark scheme would come to exactly the same judgment about a given piece of work. Pinot de Moira (2013:7) defines reliability as the probability of a candidate to be awarded a true mark, and argues that in the practice of onscreen marking, the ‘true’ mark is the one that is awarded by senior examiners, and errant markers are the examiners who deviate from that mark. Rust (2002:2) and Ofqual (2014b:5) concur that a mark is a human judgment that can never be exact across markers, and therefore remains an estimate of the true mark, with Ofqual (2013:3) arguing that examination authorities should put in place robust systems and controls to monitor and promote accurate marking, prevent poor marking and correct poor marking when it happens.

Other scholars bring in the issue of validity, which they define as the extent to which a test measures what it purports to measure (Thompson, 2016; Isaacs et al, 2013; Race, 2009). Thompson (2016:2), concur with Odendal (2011:4) that test scores can be manipulated to make inappropriate inferences and decisions, compromising its validity. Sweiry (2013:2) argues that validity is the interaction between a test question and the candidate’s mind, and examiners should credit the evidence that shows that the candidate has answered the question.

Some scholars state that validity and reliability are inversely related, increasing one would diminish the other, further arguing that if test scores are not reliable, they cannot be valid (Thompson, 2016; Sweiry, 2013; Meadows & Billington, 2005). From these arguments, it can be

deduced that quality of marking is a function of accuracy and consistency of marking that enhance the reliability and validity of the scores.

Haggie (2008:3) describes quality of marking as referring to standardization, quality control and improved support for examiners. Data Research Services (DRS, 2015; 2013) outline methodologies that are used to monitor quality of marking in the OSM environment. Examiners are trained to familiarise with the scripts before they begin live marking. The standardisation process ensures that examiners understand the mark scheme and demonstrate their ability to apply it consistently before live marking begins. The script moderation procedures such as seeding and percentage double-marking identify deviating examiners and automatically stop them from marking.

There is no explicit definition of quality control in the literature reviewed so far but an implied definition can be deduced. In the context of this research, quality control refers to procedures or mechanisms put in place by an examination authority to promote and sustain the reliability and validity of test scores.

1.9.3 Seeding

The seeds are the pre-marked scripts that are periodically presented to examiners to monitor marking consistency and adherence to the mark scheme (Pinot de Moira, 2013; Ofqual, 2013; DRS, 2013). DRS (2013:40) explain the term seeding as ‘...to “seed” a marker’s marking queue with clips of answers to which the mark is already known, as they have been marked in advance by a senior examiner....’ This implies that pre-marked scripts, called seeds, are presented to examiners at set intervals for marking. The examiners should not be able to distinguish between a normal answer and a seed. The seeds are presented in pairs and subject managers determine the number of seeds (seed window) an examiner fails before the system stops them from marking (Pinot de Moira, 2013; DRS, 2013; Hudson, 2009). The subject managers control the seed bank and set the frequency at which seeds are presented to the marker, size of the seed window, and the maximum and minimum numbers of seeds per question (DRS, 2013:41). Pinot de Moira (2013:7) calls it the hierarchical seeded system, arguing that true marks are defined by senior markers and reliability in that context is the level of agreement with the senior marker. Hudson

(2009:5) states that mark tolerance can be set that reflects the degree of agreement required between a marker's mark and the standard mark set for the seed, and this is usually zero for small value items. Deviating examiners are temporarily stopped from marking that particular question, and given additional support until supervisors are satisfied that they can mark within the standardised approach (DRS, 2015; Pinot de Moira, 2013; Ofqual, 2013; Hudson, 2009). Ofqual (2013:18) posit that senior examiners can spot-check samples of examiners' work to supplement the use of seeds.

The authors cited above concur in their definition of seeding and the mechanics of using seeds to monitor quality of marking. This research, therefore adopted a summary of the concept as put forward by these authors. In the context of this research, seeding refers to pre-marked scripts that are presented to examiners at set intervals for marking, from a monitored bank. Examiners who deviate from the pre-determined marks will be temporarily stopped from marking that question until they receive further training.

1.9.4 Percentage double marking

Hudson (2009:6) defines double percentage marking as the OSM quality control mechanism where one examiner's marks are compared with another examiner's marks according to a set sampling percentage, to keep marking within acceptable tolerance, and DRS (2015; 2013) explains that the term derives from the fact that there is a configurable percentage of scripts that can be marked by two examiners. Roan (2009:4) defines percentage double marking as the process where two marking opinions are compared in real-time, supported by automated business rules and adjudication by a senior marker. Elaborating on business rules, Roan makes reference to agreed mark tolerance ranges within which examiners must mark. Unlike seeding, there are no pre-marked scripts in percentage double marking. Pinot de Moira (2013:7) and Hudson (2009:7) concur that when two examiners award marks beyond the tolerance range to a script, the system escalates the script to a senior marker who will adjudicate the two by marking the same script. When the senior marker agrees with one marker, the system automatically penalises the other marker by stopping them from marking that particular item. If the senior marker awards a different mark from the two, the senior marker's mark stands and the two examiners are penalised (Pinot de Moira, 2013; DRS, 2013; Hudson, 2009).

Roan's (2009:4) definition summarises all the concepts involved in percentage double marking. In this study, percentage double marking was used as defined by Roan (2009:4), who defines it as the process when two marking opinions are compared in real-time, supported by automated agreed mark tolerances within which examiners must mark, and adjudication by a senior marker.

1.10 Chapter outline

Chapter 1: Introduction and orientation into the study

This chapter introduced the research problem, discussed the context of the research and articulated the research problem, statement of the problem, purpose, aim, objectives and research question. The significance and delimitations of the study were also discussed. An overview of the research methodology was given and key terms were defined as used in the study. The context of the study articulated the problem under the following headings: -public examinations in Zimbabwe; paper-based marking in Zimbabwe; introduction of onscreen marking in Zimbabwe.

Chapter 2: Quality control in the marking of public examinations

This chapter reviewed scholarly literature related to quality control in the marking of public examinations so as to explore and understand how the quality control changes with marking mode, paper based and onscreen. A conceptual framework that guided the study was developed from the literature reviewed in this section.

Chapter 3: The onscreen marking technology

This chapter reviewed scholarly literature relating to onscreen marking in order to explore the nature of the OSM and how it works, its use in the marking of public examinations by examination boards in other countries and in Zimbabwe specifically, its advantages and challenges. The chapter is arranged as follows: an overview of OSM; OSM in other countries; OSM in Zimbabwe; advantages of OSM; challenges of OSM.

Chapter 4: Research design and methodology

This chapter described the case study methodology, how it shaped this research and justified why it was adopted. The chapter is arranged as follows: The constructivist paradigm; case study methodology; the research population; sample and sampling procedures; data collection instruments and procedures; data analysis and interpretation; trustworthiness; and ethical considerations.

Chapter 5: Data presentation and analysis

This chapter analysed data presented as guided by the case study methodology. The data were presented about the context in which Biology (5008) examinations were marked and for each research question.

Chapter 6: Discussion, proposed framework, conclusion and recommendations

This chapter discussed the findings in relation to the literature and the conceptual framework. The framework that can guide quality control in the OSM of Biology examinations was proposed (Figure 6.1) This chapter also presented the major conclusions that were drawn from the findings, discussed the limitations of the study and made recommendations for action and further research.

Chapter Two

Quality control in the marking of public examinations

2.1 Introduction

The impact of technology on education has resulted in the emergence of e-assessment, which involves assessment tasks or processes designed, accessed and stored through the medium of ICT (Isaacs et al, 2013:41). Onscreen marking was designed to automate existing marking procedures without changing them (Ramakrishna et al, 2012; Roan, 2009) and is used to improve the efficiency and quality of marking examination scripts. This chapter reviewed the scholarly literature that contributed to the development of the conceptual framework that guided the study of the practice of quality control in the OSM of O level Biology examinations in Zimbabwe. The central variable, i.e. quality control of the OSM for assessment becomes the pivot on which this scholarly literature review happened. The concept of assessment in the educational context was articulated. The purpose of educational assessment was discussed, and examinations were introduced as guided by their purpose. A brief history of examinations was discussed so as to contextualise the controversy that surround them, and the need for examination authorities to put in place quality control strategies that promote the credibility of public examinations. Literature was reviewed in the context of these established quality control strategies in order to build a conceptual framework for studying quality control in the OSM environment. The conceptual framework is presented towards the end of the chapter.

2.2 The concept of educational assessment

Educational assessment is founded on the principle of scientific measurement and psychology that can be traced back to the 19th century when Johann Friedrich Herbart proposed that Mathematics could be used in psychology (Odendahl, 2011:1). Herbart argued that mental phenomenon such as perception and emotion could exist in greater or less degrees, so it could be quantified the same way as physical phenomenon (Odendahl, 2011:1). A series of milestone researches were conducted in the field of psychology, leading to the evolution of educational measurement, where scientific methods were used to measure cognitive abilities.

In the 20th century, Thorndike supported Herbart's proposal, arguing that whatever exists at all exists in some amount; therefore, the effect of educational instruction could be quantified. In 1910 Thorndike designed a fourteen-point scale to score handwriting in terms of legibility, beauty and general merit (Odendahl, 2011:2). In 1905, A French psychologist Alfred Binet designed an intelligence test which was used to predict an individual's cognitive potential. This laid the foundation for standardised tests (Muir 2017; Odendahl, 2011), which comprise public examinations which will be discussed later in this chapter.

Thorndike acknowledged that teachers who used his scale did not always agree on a score, raising issues of subjectivity in educational measurement. Teachers raised concerns about subjectivity in grading practices leading to a search for more scientific methods of testing learners (Muir, 2017; Odendahl, 2011). The continued search for scientific methods led to the adoption of written tests and numerical scoring, setting the standards for quantitative, standardised approaches to measuring learners' cognitive abilities. Researchers who believed in the quantification of cognitive abilities introduced statistical methods that mimic the ones used in natural and physical sciences (Odendahl, 2011:15), leading to the development of educational assessment.

Educational assessment was borrowed from the concepts of evaluation described by Michael Scriven way back in 1967, referring to any procedure or activity that is designed to obtain information about knowledge, attitudes or skills of a learner or group of learners (Taras, 2005; Kellaghan, 2004). Scriven argued that the process of evaluation is a methodological activity which is similar, regardless of what is being evaluated; coffee machines, teaching machines, plans for a house or plans for a curriculum (Taras, 2010: 125). Taras (2009:58) concurs with Scriven that assessment or evaluation is a process, and posits that the assessment process is the mechanism which carries out a judgment about learners' work. Taras (2010; 2009; 2005) argues that the assessment process is the judgment which can be justified according to specific weighted goals, yielding either comparative or numerical ratings. The process of assessment, therefore, requires criteria against which it can be conducted, as argued by Taras (2009:58), who wrote, that "...a judgment cannot be made in a vacuum, and therefore points of comparison (i.e. criteria

and/or standards) are necessary and in constant interplay....” The assessment process, according to Taras (2010; 2009; 2005), needs to justify the:

- data gathering instrument/criteria;
- weightings;
- selection of goals; and
- judgement against the stated goals and criteria.

In support of this notion, Kapuyaka (2013:85) argues that unless each of the aspects of assessment is meticulously scrutinised and justified, assessment might mislead and does not yield the desired results. Khan (2012:579) concurs with Kapuyaka (2013:85) when he posits that there is a need to develop clear criteria when analysing assessment information, and emphasises the need for the comprehensive criteria for setting and marking learners’ work. Ghaicha (2016:201) summarised the concept of educational assessment as a part of education where the learner achievement is appraised by collecting, measuring, analysing, synthesising and interpreting relevant information about a particular object of interest under controlled conditions. Assessment should be related to curricular objectives set for the level of the learners, and should be procedural and systematic. Ghaicha (2016:201) emphasised that assessment requires assignment of numerical values (measurement) to describe the extent to which the learners possess the skills, attitudes or behaviours that are being assessed. There is consensus that educational assessment is procedural and systematic, governed by purposes and involves educational measurement and evaluation (Ghaicha, 2016; Taras, 2012; 2010).

In the context of this study, educational assessment was used, as summarised by Ghaicha (2016:201), as a procedural and systematic process of observing, collecting and analysing numerical or descriptive information about learners’ performance so as to make informed judgements. This view considers educational assessment as a combination of procedures for educational measurement and evaluation.

Educational assessment, however, is not influenced by the scientific methods only, but, as Pellegrino (2004:7) posits, by the curriculum and the socio-political context of education. This

implies that the assessment process can either be supported or undermined by the socio-political factors of a country, hence the need to study the practice of quality control in the OSM of O Level Biology in the Zimbabwean context. Since assessment procedures are determined by the intended purpose, it is important to discuss the purpose of assessment.

2.3 Purpose of assessment

Educational assessment provides information that is used to make informed decisions. Ghaicha (2016:215) posits that educational assessment informs decisions about the policy, curriculum and learners. Some of the purposes of educational assessment are listed below thus:

- **Diagnosis:** assessment is a well established criteria for establishing learners' learning difficulties that need interventions, and can be used as a basis for reforms in the curriculum and teaching methods;
- **Setting standards:** specify clear goals for teachers and learners and ensure that standards are maintained;
- **Accountability:** holds schools and teachers to account for their practice, thereby improving classroom practice;
- **Comparison:** provides criteria for entry into secondary schools and universities, and a basis for the comparison of learner performance across schools, districts and provinces;
- **Certification:** can be used to access employment;
- **Educational management:** can help improve educational time management by focusing teachers and learners on specific outcomes of the curriculum;
- **Social evaluation:** provides legitimate membership to the global community and mobility in the international community;
- **Motivation:** motivates learners, teachers, and administrators to work ever harder to boost achievement; and
- **Qualification:** high school graduates will have the academic skills requisite for success in the workplace.

(Kaukab & Mehrunnisa, 2016; Isaac et al, 2013; Taras, 2012; Johnson & Johnson, 2009; Race, 2009; Kellaghan, 2004)

Educational assessments are therefore designed to suite their purpose, resulting in different types of assessments. Four main types of educational assessments can be identified, i.e. school-based assessments, public examinations, national assessments and international assessments. Scholars describe the school-based assessments as the ones that are designed and administered by teachers or other instructional staff, are subjective, informal, immediate, ongoing and intuitive (Braun et al, 2006; Kellaghan & Greaney, 2004). The assessments provide information and feedback that can be used to improve teaching and learning, and are rarely used for high-stakes decisions such as promotion to the next grade and are called formative assessments (William, 2014; Taras, 2012;2010; 2005; Race, 2009) that lead to classroom interventions.

National assessments, international assessments and public examinations are standardised tests that are administered to large groups of learners. National assessments are activities designed to describe the level of achievement of the whole education system of a country, and provide information to policy makers so that they can evaluate various aspects of the education system. The results can be used for accountability purposes, allocation of resources or to alert the public of issues in the education system (Ghiacha, 2016:216). International assessments provide information about the achievement of learners in a country relative to achievements of learners in other countries (Kellaghan & Greaney, 2004; Braun et al, 2006). This study focused on quality control in the OSM of examinations, a technology mainly used in public examinations. Literature was therefore reviewed in relation to script marking in public examinations.

Ghaicha (2016:213) defines a test or an examination as an instrument that can be used to elicit and measure skills, attitudes or behaviours, from which inferences can be made about a learner's performance. Public examinations have been used to make critical, often life-changing decisions, resulting in them being referred to as high-stakes tests. Such decisions may include the denial of a high school certificate, selection of learners into higher learning institutions and secondary schools, the labeling of learners and schools, employment, withholding of funding, and even the closing of a school. Learners who may do well in school all year but fail a high-stakes test may be required to repeat the course and retake the examination (Isaacs et al, 2013; Johnson & Johnson, 2009; Nichols & Berliner, 2008; Kellaghan & Greaney, 2004, Braun et al., 2006).

Because of their high-stakes nature and purpose, public examinations are taken by an increasingly large number of candidates.

The need to examine large numbers of candidates resulted in standardised tests, which make up modern-day examinations. Scholars in educational assessment consider standardised tests as the type of tests that are consistent in their scoring across all candidates, who are made to take the tests with the same questions and given the same amount of time and graded according to a predetermined set of rules so as to ensure fairness and comparability of the results (Kaukab & Mehrunnisa, 2016; Fletcher, 2009; Braun et al, 2006). Kaukab and Mehrunnisa (2016:126) point out that standardised tests are predominantly multiple-choice questions but can be true or false, short answer or essay questions. Isaacs et al (2013:43) argue that all formal tests and high-stakes examinations are standardised to give them some degree of transparency, and that the results can be scrutinised to establish their validity, reliability, fairness and equity. Examination authorities have, therefore, put in place procedures to implement quality assurance strategies for standardised tests. In the context of this research, standardised tests and public examinations mean the same. Public examinations have been in use for a long time, with procedures becoming more and more elaborate to make the examination results credible.

The next section traces the history of public examinations.

2.4 Brief history of public examinations

Written examinations can be traced back to the Chinese Sui Dynasty (581-618), when they were used to select candidates for the imperial Civil Service. The candidates were judged on the quality of their work. The examinations were taken by men only, with the examination preparation starting as early as five, when young boys were taught to recite lines from selected texts, and teenagers were attached to the masters who taught them poetry, essay writing and the Confucian philosophy (Odendahl, 2011; Pollit, 2011). Those who performed well in the examinations got government positions that were determined by the examination scores, personal influence and available openings. The Chinese civil service examinations ran for almost 2000 years up to 1905 when they were reformed.

Even in those ancient times, rudimentary procedures were put in place to reduce bias and promote fairness in the assessment of large numbers of candidates. Pollit (2011:158) describes some of the procedures for ancient Chinese examinations thus:

At every level candidates' papers were anonymised using seat numbers to avoid any risk of favouritism. From provincial level up the papers were re-written by clerks and checked by proof readers to eliminate the influence of calligraphic skills. Candidates and officials were body-searched on entry and isolated for up to three days at a time to minimise cheating: Punishments as extreme as execution were applied to candidates and officials if any cheating was discovered.

The risk of bias remained, coupled with the practical challenges of dealing with the large numbers of candidates. The Chinese government of the day established a large bureaucratic system involving thousands of officials, soldiers, clerks, governors, ministers and the emperor himself (Pollit, 2011:158). These challenges can still be discerned in modern day examinations, where the number of candidates continues to grow, resulting in an ever increasing need to eliminate bias, hence the use of technologies such as the OSM. The notion of examinations had also spread to the other parts of the world, with written tests gradually replacing oral tests in universities, and later on, in schools.

In the western world, the Industrial Revolution ushered in a movement to return school-age farmhands and factory workers to the classroom. Standardised examinations enabled the newly expanded learner body to be tested efficiently (Hays, 2013; Fletcher, 2009). In France, the psychologist Alfred Binet began the development of a standardised test of intelligence; work that would eventually be incorporated into a version of the modern intelligence test, dubbed the Stanford-Binet Intelligence Test. In Belgium, written tests were introduced at the Louvain University in the 1400s. In Italy, oral examinations were introduced at the University of Bologna in the 1200s, and written tests came to St. Ignatius of Loyola in the 1540 (Muir, 2017; Huddleston & Rockwell, 2015).

In England, oral examinations were administered at Oxford University in the 1600s and were replaced by written examinations in 1702. The University of Cambridge and the Harvard adopted written tests in the 19th century (Muir, 2017; Odendahl, 2011). The first school examinations

were written in 1858 when schools approached universities to provide examinations for boys, with girls sitting their first examination in 1867. The 1858 examinations were sat in English Language and Literature, History, Geography, Geology, Greek, Latin, French, German, Physical Sciences, Political Economy and English Law, Zoology, Mathematics, Chemistry, Arithmetic, Drawing, Music and Religious Knowledge (University of Cambridge Local Examinations Syndicate [UCLES], 2008: www.cambridgeassessment.org.uk/news/how-have-school-exams-changed-over-the-past-150-years/). Today, several examinations authorities set and administer examinations in England and Wales, under the supervision of the Ofqual (Ofqual, 2014c; 2013).

In America, examinations were used to recruit soldiers for the First World War. In the 1800s, examinations had reached American schools and were used to assess aptitudes and achievement, when Horace Mann introduced the concept of using examinations in Boston schools. The objective was to obtain objective information about the quality of instruction and compare schools and teachers within each school. The most common American tests are the TerraNova, the Stanford Achievement Tests (SAT) and the Metropolitan Achievement Test (MAT), to name a few (Fletcher, 2009; Nichols & Berliner, 2008).

Examinations are also an important aspect of education in Africa, with many countries adopting examination systems from their former colonial masters. Kellaghan (2004:5) posits that there are three major examinations in most African countries, one at the end of primary school and two at secondary school, with examinations managed by the education ministries in Francophone countries and examination agencies in Anglophone countries. Examinations agencies, awarding bodies or examination boards are institutions that develop and award qualifications that are used in schools, colleges and workplaces, and are governed by the codes of practice that ensure quality, consistency, accuracy and fairness in assessment (Isaacs et al 2013:25). Zimbabwe, being a former British colony, adopted the English examination system, with the Zimbabwe School Examinations Council managing the examinations (Mashanyare & Chinamasa, 2014; SAAEA, 2014). Examinations have always been controversial, with some groups advocating for, and others against, them, as discussed in the next section.

2.5 Controversy surrounding examinations

Examinations are not without controversy, and in America, movements have been formed for and against standardised tests and examinations. The proponents of standardised tests have put forward several advantages that are purportedly derived from them. Munoz (2013: www.education.com) encouraged American parents and learners not to despair towards the countdown to high-stakes tests, and outlines the pros of the tests that should be kept in mind, thus:

High-stakes test results can be used to help teachers create a learning plan based on your kid's needs—helping her in the long run. Look at your child's test results as a tool for progress, not as a judgment on ability or intelligence. Data from state-wide testing is almost always publicly available. As a parent, you can look at these results to see how well, or poorly, your child's school is performing. Access to this information will help you make more informed decisions about where and how your child will get the best education.

The proponents of high-stakes examinations justify the decisions that derive from their use. Johnson and Johnson (2009: <http://www.education.com>) posit that high-stakes tests proponents believe that for them to be effective, the consequences of low achievement should be severe, hence, the use of sanctions such as repeating a grade, withholding a high school certificate, or school closure. Critics of high-stakes examinations, however, continue to raise concerns that cannot be ignored by their proponents, resulting in elaborate procedures that are strictly followed in the design, scoring and grading of the examinations (Kandur, 2017; Kaukab & Mehrunnisa, 2016). The issues of fairness, validity, security, reliability and many others have been raised by the critics of high-stakes examinations and these are summarised according to Kandur (2017: <https://www.dailysabah.com/feature/2017/09/23/testing>) (illustrated in Figure 2.1), and according to the author, standardised tests and examinations:

- are unavoidably biased by social-class, ethnic, regional and other cultural differences;
- unfairly advantage those who can afford test preparation;
- radically limit teacher ability to adapt to learner differences;
- provide minimal to no useful feedback to classroom teachers;

- hide problems created by margin-of-error computations in scoring;
- penalise test-takers who think in nonstandard ways (which the young frequently do);
- give control of the curriculum to test manufacturers;
- encourage the use of threats, bribes and other extrinsic motivators to raise scores;
- assume that what the young will need to know in the future is already known;
- emphasise minimum achievement to the neglect of maximum performance;
- produce scores which can be manipulated for political purposes;
- create unreasonable pressures to cheat;
- use arbitrary, subjectively-set pass-fail cut scores;
- reduce teacher creativity and the appeal of teaching as a profession; and
- lessen the concern for and use of continuous assessment.

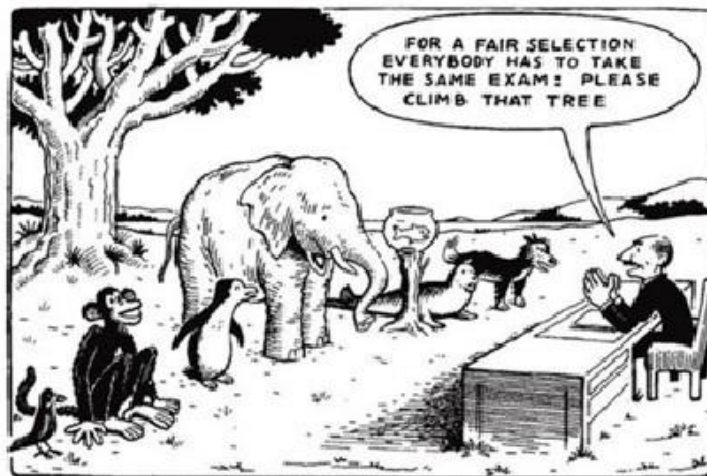


Figure 2.1: Bias in standardised tests (adopted from Kandur, 2017)

A close analysis of this cartoon shows that the bird and the monkey are adapted to accomplish the task, if climbing means getting to the top of the tree, even if the bird is likely to be more advantaged than the monkey. There is no way a penguin or a fish can climb the tree, implying that they will not be selected for the purpose the test was meant to serve. It is for reasons such as these that Kandur (2017: <https://www.dailysabah.com/feature/2017/09/23/testing>) raises a lot of questions about examinations and wrote thus:

There has always been controversy about exams. Are we really testing our children on what they learn, on their abilities, or on what they can memorize and regurgitate? Are exams the best way to evaluate children? Einstein is reported to have said: "...if you judge a fish by its ability to climb a tree, it will live its whole life believing that it is stupid..."

Even the ancient Chinese examinations were allegedly biased, favouring learners from certain social classes. The examinations theoretically seemed to be open to all men but Hayes (2013: www.factsanddetails.com/china/cat2/4sub9/entry-5385.html) thought that children of merchants, landowners and families with money had an advantage in that their parents could hire tutors to teach them. The performance of the Chinese men in the public examinations was therefore dependent on their social classes.

Some African countries are still to recover from the impact of colonisation in spite of educational reforms that were meant to undo the racial disparities in the colonial education systems (Shizha & Kariwo, 2011; Bayat, Louw & Rena, 2014). The colonial government in Zimbabwe provided more educational funding for whites than for blacks, resulting in two education systems pursuing different curricula. At independence the new government crafted educational policies that were meant to address racial imbalances, schools were built in marginalized areas, more teachers were trained and more learners were enrolled in schools, with primary education being free (Shizha & Kariwo, 2011; Kanyongo, 2002). However, the policies encountered several challenges that included unrealistic time frames for goals and shortage of financial resources to sustain free primary education, forcing the government to introduce levies in both primary and secondary schools (Kanyongo, 2002:71). This move created an access barrier to learners who could not afford the fees. Some learners dropped out of primary and secondary schools because they could not afford to pay the levies, with more girls dropping out than boys (Marist International Solidarity Foundation [FMSI], 2011:2). These access barriers can still reflect in the examination system.

South Africa is still trying to deal with the legacy of apartheid that created poor quality education for poor learners compared to their wealthier counterparts (Bayat et al, 2014:183). Research shows that black learners in townships are often hungry and ill; do not have proper clothing and

study materials; they lack parental support, motivation, self-esteem and language proficiency; and frequently change schools (Bayat et al, 2014:184). The authors posit that in 2013 the World Bank ranked South Africa second from last in the world for Mathematics and Science education, ahead of Yemen. The international assessment used to rank the countries did not take cognisance of the variations in the circumstances of the learners, creating the situation illustrated by the cartoon in Figure 2.1. Critics of examinations might really have a point when they say examinations are inherently biased.

Despite the intensified criticism, examinations continue to dominate the education systems of the world probably for two reasons. First, elaborate examination procedures have been put in place to ensure quality examinations, and to justify the inferences made from them. Secondly, there is no better way of selecting the candidates for placement in universities, colleges, jobs and for implementing interventions in education system. As long as there is no alternative method of providing information about learners' performance, then there is always a need to intensify research so as to inform procedures that guide test design, scoring or marking and grading, hence the purpose of this study, to explore the practice of quality control in the OSM of O Level Biology examinations in Zimbabwe.

Assessment procedures need to provide clear criteria for judging the quality of assessment systems in the 21st century, based on the changing demands of today's workforce, advances in other nations, and original analysis (Darling-Hammond, Herman, Pellegrino et al, 2013:4). This is supported by Hill (2013:19), who posits that because examination results are the main, if not the only basis for making high-stakes decisions, the results should always be accurate and error free. Hill (2013:20) enumerates quality assurance strategies that modern examination authorities should adopt– paying attention to recruitment and training of examination personnel; creating a culture of assuming responsibility for improving quality; establishing effective system of internal controls; automating examination processes to eliminate human error; and implementing fair and transparent results and appeals processes.

Automation of assessment processes should not compromise the quality of the assessments. This study, therefore, interrogated quality control to inform the practice in the OSM environment in

Zimbabwe and similar contexts. The next section reviews scholarly literature on the procedures put in place by the examination boards to control the quality of script marking.

2.6 Marking of examination scripts and issues of validity and reliability

Script marking is an important aspect in examinations and has three major purposes, which are, to:

- ensure the consistent application and interpretation of assessment performance standards in the subject;
- ensure that scores awarded to candidates across examination centres are fair and comparable; and
- ensure that results are valid and reliable.

(Government of South Australia [GOSA], 2017:3).

These purposes of script marking should be accomplished regardless of the marking mode or approach, on paper or on computer screens.

Validity and reliability are important principles of assessment that have determined the procedures that govern large scale assessments such as public examinations. Validity is defined as the extent to which a test measures what it purports to measure, as discussed in Chapter 1, Section 9.2. Thompson (2016:2) posits that validity refers to the meaningfulness, usefulness and appropriateness of inferences made from a test, farther arguing that validity is not an inherent characteristic of a test or measurement procedure, but in the ‘...reasonableness of using the test scores for a particular purpose or inference....’ When test scores are used for purposes that they were not intended to serve, then they become invalid.

To illustrate the issue of validity, Odendahl (2011:4) traced the milestone researches that led to the development of educational assessment, and described how craniometry (measurement of human skulls) was used to make unwarranted inferences. Craniometrists believed that the larger the skull the keener the mind that once occupied it. Two craniometrists, Samuel Morton and Pierre Paul Broca, selectively chose and manipulated data of the skull measurements to claim

that white males were more intelligent than women and men of other races. Odendahl (2011:5) commented thus:

But even where the skull measurements were precise and accurate, the inferences the craniometrists drew were without warrant. Offering no evidence about their subjects' ability to perform intellectual tasks and ignoring the fact that modern humans have smaller brains than those of whales, elephants, dolphins ..., craniometrists nonetheless asserted a strict correspondence between intelligence and size of the physical brain.

This example illustrates that test scores can be manipulated to make inappropriate inferences and decisions and justifies the concerns raised by the critics of standardised tests. There is, therefore, a need to provide evidence to justify the validity of inferences and decisions made from the test scores. Evidence should be provided to justify the scoring or marking criteria. However, some critics argue that validity should not be the responsibility of score interpreters but of test designers.

Sweiry (2013:2) argues that the notion that validity is a property of test interpretation ignores the role of test developers in making sure that question papers and mark schemes contribute towards valid assessments. This has resulted in an alternative notion of validity. Some scholars argue that validity is the interaction between a test question and the candidate's mind, and examiners should credit the evidence that shows that the candidate has answered the question (Sweiry, 2013:2). This argument implies that validity is actually inherent in the test instrument. In the context of this research, validity is viewed as defined by Sweiry (2013:2) who believes that question papers and mark schemes contribute towards the validity of assessments. This view of validity explains its relationship with reliability that will be discussed later in this section.

Reliability is defined as the extent to which test scores are accurate, consistent, stable and reproducible (Thompson, 2016; Race, 2009). Meadows and Billington (2013:9) define marking accuracy as the absolute difference between the mark given by the examiner and the estimated true mark, and marking consistency as the correlation between the mark given by the marker and the estimated true mark. Rust (2002:3) and Ofqual (2014b:5) concur that a mark is a human

judgment that can never be exact across markers, and therefore remains an estimate of the true mark. To this effect, Tisi et al (2013:1) argue thus:

Many people accept the fact that a test result could have been different if the candidate had taken the exam on a different day or if different questions had come up. This uncertainty in the system appears to be accepted as the “luck of the draw”. However, when it comes to human error in the process of assessment, including marking variability, the general public are understandably less tolerant.

This implies that the unreliability that emanates from the marking process erodes the public confidence of any examination system. Given the controversy that surrounds examinations, any marking process should, therefore, promote accuracy and consistency of marking, so that candidates are awarded actual marks.

There is a general agreement in assessment circles that the candidates’ actual mark at any particular time is made up of their ‘true’ score and a certain amount of measurement error (Tisi et al. 2013; Pinot de Moira, 2013; Ofqual, 2013). It can therefore be argued that a true mark is almost impossible to obtain but interventions can be made to bring the mark awarded to the candidate as close as possible to the true mark. There are statistical means of estimating the true score as guided by theories such as the Classical Test Theory and the Item Response Theory (Pinot de Moira, 2013; Baird et al, 2013; Meadows & Billington, 2005). Such statistical estimates were not discussed in this study because they were outside its focus.

In the context of this study, the true score is viewed as defined in the practice of examination scripts marking. Pinot de Moira (2013:7) argues that in the practice of script marking, the ‘true’ mark is the one that is awarded by senior examiners, and errant markers are the examiners who deviate from that mark. The quality of marking in the OSM environment is monitored mainly by seeding and percentage double marking, where the final mark for a standard script is fixed by senior markers (Pinot de Moira, 2013; Ofqual, 2013). All quality control measures are meant to ensure that examiners do not deviate from the true mark, or they deviate within acceptable ranges. Quality control measures should, however, preserve the validity of the examination results.

Even though they view validity differently, Thompson (2016:1) and Sweiry (2013:2) concur that reliability is a component of validity, arguing that if test scores are not reliable, they cannot be valid. Meadows and Billington (2005:13) confirm that reliability is a prerequisite for validity, further noting that validity and reliability are inversely related. To illustrate their point, Meadows and Billington (2005:13) write:

Reliability and validity are in tension. Attempts to increase reliability, for example, making the marking scheme stricter often have negative effects on validity because candidates with good answers not foreseen in the mark scheme will not be given high marks. Another way of increasing reliability is to test a smaller sample of the curriculum. However, this would be a less valid test of candidates' knowledge and skills in the subject area and would provide an incentive for schools to improve their test results by teaching only those parts of the curriculum actually tested.

This confirms Sweiry's (2013:2) notion that the validity of assessments is dependent on how the candidates interact with questions and what is given credit by examiners. The relationship between validity and reliability calls for caution by the examination authorities to adopt the marking strategies that do not compromise the validity of the test scores. Concerns have been raised that, in the wake of the OSM technology there is a risk of tests being designed to suit the demands of the technology at the expense of test validity (Fowles, 2011; Roan, 2009). Meadows and Billington (2005:7) identify the three sources of unreliable scores and these are factors in the test itself, factors in the candidates taking the test and factors related to scoring. Sweiry (2013:2) emphasises that marking reliability poses the greatest threat to the overall reliability of assessments, hence the need to focus on quality control in the OSM environment.

Several researches were conducted on examinations marking and identified several factors that influence the reliability of marking. Meadows and Billington (2005: 20) reviewed literature on the reliability of marking in which they identified several factors that influence reliability of scores. These are enumerated below as follows:

- Marking bias that emanates from the contrast effect that arises from the quality of the immediately preceding script; the candidate's gender, ethnicity and culture; candidate's handwriting;
 - Changes in the consistency and severity of marking over time;
 - The examiner's background that results from training and experience, fatigue, mood, poor concentration and lack of vigilance;
 - Question format with highly structured questions being marked more reliably and open-ended questions being marked less reliably; and
 - Mark schemes that are unsatisfactory as sources of unreliable marking.
- (Meadows & Billington, 2005: 20)

Meadows and Billington (2005:68) concluded that unreliability is inherent in assessment in general and in marking in particular. The examination authorities can improve marking reliability through marker training, use of experienced examiners and paying close attention to mark schemes.

Tisi et al (2013:1) reviewed literature on marking reliability in which they detailed the statistical methods of quantifying the marking reliability. As stated earlier, such methods are outside the focus of this study, so they were not discussed. The review, however, enumerated advances in improving the marking reliability, and these include:

- Increasing item constraint, highly specified mark schemes, lower maximum marks and questions targeted at lower cognitive skills;
- Marker education and experience affect accuracy, but the relationship is not simple and depends on item type;
- Item level marking is more reliable than whole script marking because it reduces the effects of examiner bias;
- OSM marking appears to be as reliable as PBM, even for long answer and essay questions;
- OSM facilitates item level marking and all its associated benefits; and

- OSM allows continuous monitoring of marking, enabling inaccurate marking to be detected, and eliminates errors resulting from addition and transcription of marks.
(Tisi et al, 2013:21).

The OSM technology, therefore, has the capacity to improve the quality of marking by automation of tasks such as addition and capturing of marks and many others. Odendahl (2011:141) enumerates some more strategies for improving fairness and reducing bias at scoring of scripts, which are:

- Removing names or the types of identity information from responses;
- Distributing responses randomly to examiners;
- Specifying the scoring criteria in written guidelines;
- Illustrating the scoring criteria with sample responses;
- Training examiners; and
- Monitoring examiners.

Ofqual (2014c; 2013) posits that examination boards should make sure that marking is carried out to the standard. The Ofqual (2014c:17) claims that in this regard:

It is crucial that there are rigorous checks to ensure that the mark schemes are interpreted correctly and consistently by each examiner, and examiners are consistent throughout the marking session. It is also important to make sure there are no clerical errors and marks are added correctly and assigned to the correct student.

The examination authorities across the world have, therefore, put in place procedures that ensure quality marking. The Ofqual, a quality assurance body for UK examinations, reviewed literature on marking internationally and accessed research data from New South Wales, Canada, China, Hong Kong, New Zealand, Korea and the United States. The literature review was meant to answer the following four questions:

- Who marks assessments in other countries?

- Do any jurisdictions use double or multiple marking? Is there any evidence as to the rationale behind and the impact of this? How is it managed? Is it targeted at certain subjects and/or types of questions?
 - Is there onscreen marking of assessments in other countries? If so, do they use item level marking? Is there any evidence of the impact of this?
 - How do other jurisdictions' quality assurance of examination marking compare with our own? This includes how the quality of the interpretation of the mark scheme is monitored, how clerical errors are eliminated and inter-rater (the degree of agreement among examiners) is monitored and reported.
- (Ofqual, 2014c:5)

The review established that examiners are recruited from practicing or retired teachers, and are required to have specified levels of education and good knowledge of the subjects they mark. In a few countries, however, practicing teachers are not eligible to mark examination scripts. The examiners respond to the adverts sent out by examination authorities (Ofqual, 2014c:6). Double marking and multiple marking of essay-type, open-ended, constructed response and short answer questions are done in most countries. In double marking, two independent examiners mark each candidate's response and the final mark is the average of the two marks. In multiple marking, more than two examiners are used. The Ofqual noted that the use of double marking and multiple marking to produce final scores is an acknowledgement that legitimate differences in opinion can exist between the markers. However, in the UK multiple and double marking are not commonly used (Ofqual, 2014c:6).

The quality of marking is monitored using a sampling approach where senior examiners review or re-mark a sample of scripts marked by each examiner (Ofqual, 2014c:9). The review also established that just like in the UK, examination authorities in other countries train examiners in the application of mark schemes through a standardisation process. Multiple choice answers are automatically marked and open-ended questions are marked by examiners across the world. Live marking is monitored by senior examiners through double, multiple or sample marking. Multiple choice responses are automatically marked by optical mark-reading machines, while some open-

ended questions are marked on the computer screens by examiners. Quality control in the OSM environment is enhanced by real-time monitoring of examiners (Ofqual, 2014c:17).

The results of this review show that the quality of marking is assured by recruiting and training the examiners with the requisite subject expertise, standardisation and the monitoring of live marking by double, multiple and sample marking. It is also apparent that these quality assurance measures are implemented by the majority of the countries that provided data for the review. However, the review did not access the research data from any African country, hence the need to study quality control in the OSM environment in Zimbabwe.

The next section discusses the literature on examiner training.

2.7 Examiner training

Examiners are trained at two occasions, i.e. well ahead of marking to build a pool of examiners, and just before marking to ensure correct and consistent application of the mark schemes. The training conducted just before marking is called standardisation (Ofqual, 2014e; 2013) and is discussed later on. It is standard practice for the examination boards to recruit part-time staff to mark the examination scripts (Hill, 2013; Ofqual, 2014c), and examination boards need to provide comprehensive training programmes for the examiners. The Ofqual (2014e:2) posits that the UK examination boards maintain a pool of trained examiners for each subject component where they choose examiners who mark scripts for a particular examination session. New examiners are trained before they are added to the pool of examiners and additional training is offered to all examiners at standardisation. Hill (2013:20) argues that errors tend to occur when part-time personnel are not sure of what they should do in some situations, further explaining thus: “...training needs to anticipate such situations and provide clear guidance on appropriate responses. Training manuals and videos are essential tools to codify practice....”

Examination boards recruit and train examiners who mark responses to different types of questions. Tze Ho and Chong Sze (2013:1) reviewed marking and grading procedures for Liberal Studies examinations in Hong Kong. The review intended to inform the public about marking procedures and to gain their confidence in the qualifications conferred by the Hong Kong

Examinations and Assessment Authority (HKEAA). To prepare for the marking of Liberal Studies on screen, the HKEAA conducted three examiner training sessions between 2010 and 2011, training a total of 1319 examiners for the subject. According to Tze Ho and Chong Sze (2013:3), the participants were invited to a marking centre, and the training was meant to:

- familiarise examiners with the marking process;
- induct examiners on the criteria and standards of marking Liberal Studies;
- familiarise examiners with OSM; and
- collect the marking statistics of the participants to facilitate the selection of examiners for the live examinations.

(Tze Ho & Chong Sze, 2013:3)

The participants first attended a three-hour meeting where they were briefed on marking guidelines, criteria and standards. The participants then marked sample scripts that they discussed in group meetings led by assistant examiners. The group discussions helped the trainee examiners to align their marking with the set standards. The participants then marked 15 scripts each for Papers 1 and 2, whose marks had been standardised by the experienced examiners. Their marking statistics were collected and analysed to guide the appointment of examiners who mark live examinations (Tze Ho & Chong Sze, 2013:4). The review, however, did not specify the type of the examiners who were recruited and by HKEAA, as done by UK examination boards. The examination authorities in the UK recruit and train three major categories of examiners; clerical, graduate and expert examiners (Ofqual, 2014e:3).

The Ofqual (2014e) conducted a research on the quality control activities of UK examination boards. Data were collected by interviews and visits to the examination boards. The study sought to describe the main quality control activities carried out by the examination boards between March and October 2013. On recruitment, the research established that examination boards recruited examiners through external advertising. The prospective examiners are required to complete an application form which was reviewed by the examination board. One board administered aptitude tests to the prospective examiners (Ofqual, 2014e:3). The UK examination boards set minimum requirements for examiners, although they vary by examination board and

subject. Minimum requirements for clerical markers and graduate markers are lower than for expert markers, although they should hold an undergraduate degree. Expert markers must have an undergraduate degree in the subject they wish to examine or a related subject, and must be qualified teachers. Most examination boards specify at least one year's teaching experience, often in the subject and at the level they wish to mark. Although a few examination boards do not require their examiners to have experience of teaching the subject they wish to examine, the majority do (Ofqual, 2014e:3). The same examiner categories were used by UCLES as early as 2004.

Raikes et al (2004:9), writing about OSM at UCLES, posit that the technology offers the flexibility of splitting the candidates' scripts by question and distributing the responses appropriately to the three types of examiners described as follows:

- Clerical markers: these are trained and standardised markers who have little or no knowledge of the subject and are proficient in the language in which the scripts are written and should possess the adequate IT skills
- Graduate markers: these are trained and standardised markers who are recent graduates or post graduate learners in the subject; and
- Assistant Examiners/expert markers: these are trained and standardised markers who have experience in the performance of candidates at the level of the examination in addition to language proficiency, ICT skills and subject expertise.

(Raikes et al 2004:9).

The different types of examiners mark the responses to different types of questions. These have a bearing on the accuracy and hence, the quality of marking. This was demonstrated by Suto and Nádas (2008:9), who conducted a research to investigate the key factors that contribute towards the personal expertise of the examiners, hence the accuracy of marking the International General Certificate of Secondary Education (IGCSE) Biology. The researchers conceptualised that marking accuracy is influenced by the demands of the task and marker expertise. The research involved 42 markers, comprising five groups: (i) eight experienced examiners, (ii) nine biology teachers with no marking experience, (iii) eight graduates in biology, who had no teaching or

marking experience, (iv) nine graduates in other subjects, who had no teaching or marking experience, and (v) seven non-graduates, who had no university education, teaching experience or marking experience. This design enabled the relative effects on the accuracy of the following factors to be elicited: marking experience, teaching experience, highest education in a relevant subject, highest education in any subject, and gender. Twenty-three examination questions, from the November 2005 examination, were explored, varying in format, number of marks, and difficulty for candidates, and cognitive marking strategy complexity. The researchers prepared the following four samples of the candidates' responses:

- Practice sample made of five different responses to each question;
- First standardisation – made of 10 different responses to each question;
- Second standardisation – made of 10 different responses to each question (this sample was marked by participants who could not meet the level of accuracy required in the 1st standardisation); and
- The main sample – 50 different responses to each question.

All markers marked identical response samples for each question. Logistic regression and ANOVA were used to model the accuracy of the data yielded. The results showed that marking accuracy was influenced by the factors listed below (in order of importance):

- Highest level of education in any subject;
- Highest level of education in a relevant subject;
- Teaching experience;
- Marking experience; and
- gender (women marked more accurately than men did).

The researchers, however, noted that the contributions of these factors to marker expertise are not independent of each other (Suto & Nádas, 2008:9). The results also showed that marking accuracy was generally high, with expert examiners topping the list and non-graduates at the bottom of the list. The questions that demanded simple cognitive marking strategies were marked with high accuracy by all examiner categories. The questions demanding complex cognitive

marking strategies were marked with less accuracy by all examiner categories, with experts being the most accurate and non-graduates being the least accurate (Suto & Nádas, 2008:10).

These results confirm the importance of setting minimum qualifications for examiners and standardisation (Ofqual, 2014e; Tisi et al, 2013). The finding that non-graduates, who had no university education, teaching experience or marking experience were the least accurate shows that examiner recruitment has a bearing on the quality of marking. The results are important to this study because the Biology examiners were investigated. The results could guide the recruitment and training of examiners for the O Level Biology examinations in Zimbabwe. However, the researchers did not mention the mode of marking. It is not clear whether the questions were marked on paper or on screen. All the same, the factors that influence the marking accuracy could apply in any mode of marking, with OSM having the advantage of enhancing the efficiency of quality control mechanisms (Ofqual, 2014c; 2014e; Tisi et al, 2013). It was therefore important to investigate the influence of examiner training and standardisation on the quality of marking O Level Biology examinations in the OSM environment in Zimbabwe.

A similar study was conducted by Meadows and Billington (2013:9), who studied the effect of marker background and training on the quality of marking in the General Certificate of Secondary Education (GCSE) English. The research aimed to establish the effects of marking experience, subject knowledge and teaching experience on the marking reliability of GCSE English. The study was a quasi-experimental design involving (i) 97 GCSE English examiners, (ii) 81 trainee English teachers, (iii) 99 English undergraduates and (iv) 82 graduates from other disciplines. All groups of markers marked 199 part-scripts composed of five questions of varying cognitive demand. Two questions required relatively short answers while three required relatively long answers. The markers initially marked 100 scripts using the marking scheme and later received standardisation training. After the training, the markers then used the standardised mark scheme to mark 99 scripts. The scripts were randomly selected from over 22 000 scripts that had been marked in the summer of 2005. The data were analysed by two-way Anova and simple effect analysis. The marking quality was assessed by relative marking severity, indicated by the mean mark awarded by the marker; marking accuracy, measured by the absolute difference between the mark awarded by the marker and the estimated true mark; and marking

consistency, indicated by the correlation between the mark awarded by the marker and the estimated true mark.

The results showed that all marker groups generally marked accurately. There were some undergraduates who marked as well as the best GCSE examiners. The background had no effect on marking accuracy, but had an effect on the marking consistency, with GCSE examiners and trainee teachers marking more consistently than the undergraduate groups. Training improved marking accuracy across all marker groups but to a small extent. However, training had no effect on the marking consistency of the undergraduate groups but seemed to influence the consistency of trainee teachers. The overall results indicated that GCSE examiners were more consistent than trainee teachers and the two undergraduate groups. The item level analysis showed that the marker groups were equally accurate in marking short responses. The GCSE English examiners marked the longer responses more accurately than trainee teachers and undergraduate groups (Meadows & Billington, 2013:15).

Some conclusions made by Meadows and Billington (2013:15) were contrary to the conclusions made by Suto and Nádas (2008:10). The former concluded that the examiners' background does not affect the marking accuracy, with the undergraduate markers marking as accurately as the best expert markers, while the latter concluded that the examiners' background contributes towards personal expertise, hence accuracy of marking, with expert examiners being the most accurate. The two studies, however, concur that expert examiners mark long responses more accurately than other marker groups. Despite the differences on the marker background, the two researches provide an insight on the recruitment of examiners and justify the categorisation of markers by U.K examination boards.

However, a concern has been raised about the risk of designing examinations with more short response items that are easy to mark, compromising the validity of the examinations, especially in the OSM environment, given the inverse relationship between validity and reliability (Fowles, 2011; Meadows & Billington, 2005). This implies that the flexibility of distributing scripts to different types of examiners, offered by the OSM technology, may improve the accuracy and efficiency of marking, but may also compromise the validity of examinations. The examination

boards should therefore put in place mechanisms that preserve assessment validity, in the wake of the OSM technology, when they categorise examiners according to the type of items.

All examiner categories have hierarchies among them and well-defined reporting lines. The UK examination boards have senior examiners who supervise marking at prescribed levels, led by the principal examiner (PE) (Ofqual, 2014e; 2013; Baird et al, 2013; Tze Ho & Chong Sze, 2013). In Ontario, Canada, the top hierarchy is made of scoring leaders and scoring supervisors. At the bottom rank are the scorers. The scoring leaders and scoring supervisors are trained by education officers of the examination authority. The scoring leaders and supervisors have the responsibility to train all scorers; oversee the scoring of items; resolve issues that arise during scoring; and ensuring that scoring materials are applied consistently. The scoring leaders and supervisors are selected from a pool of experienced examiners (Ofqual, 2014b:63).

In Hong Kong the PE is called the chief examiner while other marking supervisors are referred to as assistant examiners. The examiners are simply called markers (Tze Ho & Chong Sze, 2013:7). Team leaders (TL) supervise the smaller teams of examiners (five to six) during marking and in OSM they select the scripts that will be used for seeding. The TLs are, in turn, supervised by a PE and assistant PEs (Ofqual, 2014e; 2013). The Ofqual (2014e:3) research established that TLs were recruited from the high performing examiners. The PE vacancies are externally advertised and the high performing TLs are encouraged to apply. The hierarchy of examiners is summarized in Figure 2.3.

Such rigorous recruitment procedures could guarantee the quality of standardisation and monitoring practices, hence quality of marking. This research sought to study recruitment and training practices for O Level Biology examiners in Zimbabwe so as to design a framework for quality control in the OSM environment.

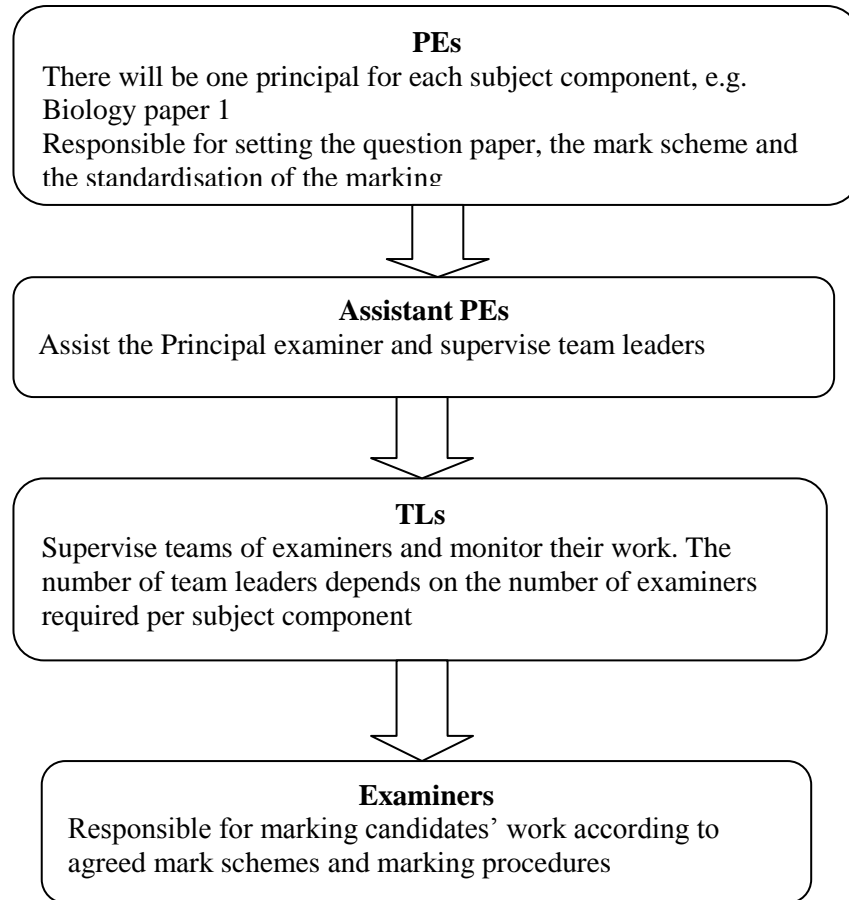


Figure 2.2: Hierarchy of examiners (Adopted from Ofqual, 2013:10)

Once they are recruited, the examiners are trained. The Ofqual (2014e:4) established that all UK examination boards provide online training for some subjects. The training and guidance documents are provided on technical and administrative tasks of examining, with some boards monitoring the extent to which examiners access the training materials. However, a few boards offer face-to-face training for selected subjects, according to the Ofqual research. The HKEAA of Hong Kong offered face-to-face training for Liberal Studies examiners, although they marked the training scripts on screen (Tze Ho & Chong Sze, 2013:3).

After the initial training, the UK examination boards monitor examiner performance until the end of each examination session to identify and rate them on a five-point scale. The examiners with training needs may either be retrained before they are invited to the next session, or removed from the pool of examiners when there is no shortage of examiners (Ofqual, 2014e:4). Removing

examiners with training needs might not be a good practice for the examination board. The Ofqual (2014e:4) posits that some examiners can choose to stop marking a subject. In such cases it might be cheaper to re-train than to recruit and train replacements. Examiners with training needs in one subject might be competent in another subject, as indicated by the Ofqual (2014e:5) who argue that performance on one subject is not necessarily a good indicator of future performance on another. Suto and Nádas (2008:9) established that accuracy of marking Biology was mainly influenced by the examiners' highest qualification in any subject. The examination authority could establish a system that allows examiners with training needs in one subject to transfer to another subject of their choice, instead of dropping them from marking teams.

Some examination boards train new TLs on how to use the OSM system to manage their teams. The training of the PEs varied with the examination boards. Some PEs go through accredited training programmes while others are guided and trained by subject managers who are in overall charge of specific subjects (Ofqual, 2014e:5). The online training may not be feasible in the Zimbabwean context where there are challenges of ICT access and use (National ICT Policy, 2015:13-15). The ZIMSEC may have to settle for face-to-face training only. It was therefore important to investigate how the recruitment and training of examiners influenced quality control practice in the marking of O Level Biology in Zimbabwe. The trained examiners still need to attend standardisation meetings before they can be allowed to mark candidates' scripts.

2.8 Standardisation of marking

Standardisation is a type of examiner training that is conducted just before the script marking begins, and takes place for each examination paper every examination session because mark schemes are specific to question papers, and like training, it can be conducted through face-to-face meetings or online (DRS, 2015; Ofqual, 2014b; Ofqual, 2013). Ofqual (2014b:59) posits that standardisation ensures that examiners are fully competent in applying the marking scheme consistently before they begin marking and has several purposes, which include providing a context within which the marking process takes place; defining the task to be performed by the examiners; and minimising the effects of variables such as item difficulty from the marking process. Ofqual (2014:59) argues that the good examiner training is evidenced by stable marks

throughout the marking process; marks that reflect the relative difficulty of items; and marks that reflect realistic expectations of the candidates' performance.

In the UK, there are two meetings that comprise the standardisation process, which are the pre-standardisation meeting and the standardisation meeting (Ofqual, 2014e; Baired, et al, 2013). The pre-standardisation meeting is attended by the PEs, assistant PEs and some or all TLs (Ofqual, 2014e; Baired et al, 2013), who supervise the quality of marking. The Ofqual (2014e:5) insists that the meeting is always conducted face-to-face, probably to maximise the interaction among the examiners. The meeting is intended to discuss the candidates' responses to questions in the examination so as to set the marking standards and criteria for that particular examination. The meeting includes:

- briefing on procedures, timelines, documents, contracts, nature and significance of standardisation;
- briefing on issues from current and previous examinations;
- discussion of the mark scheme;
- marking of selected candidates' responses;
- discussion on handling of unexpected, acceptable responses;
- confirmation of true scores awarded to selected scripts that will be used for practice at the standardisation meeting;
- review of the mark scheme in relation to candidates' responses; and
- confirmation of the final mark scheme.

(Ofqual, 2014e; Baired et al, 2013).

In PBM, the PE selects the practice scripts from their allocation of scripts before the pre-standardisation meeting. In OSM, where more scripts are required, the TLs select the scripts. The pre-standardisation meeting is therefore meant to prepare for the standardisation meeting.

The pre-standardisation meetings are also conducted by the HKEAA of Hong Kong. Tze Hong and Chong Sze (2013:7) did not describe the pre-standardisation meeting in detail but mentioned it in passing thus:

Before the markers' meetings, a representative sample of candidates' scripts was selected and marked by the Chief Examiner and a group of experienced senior Assistant Examiners, whereby a consensus was arrived at through professional discussion. Some of these standardised scripts were used for marking standardisation, training and qualifying purposes.

The standardisation meeting is attended by all the examiners appointed to mark the candidates' scripts for that particular session. Ofqual (2014e:5) concur with Baired et al (2013:10), that the meeting includes:

- briefing on procedures, timelines, documents, contracts, nature and significance of standardisation;
- briefing on issues from current and previous examinations;
- discussion of issues that emerged from the marking of practice scripts;
- discussion of the marking scheme;
- marking and discussion of sample scripts that illustrate the range of performance and possible response types; and
- handling of unexpected but acceptable responses.

Baired et al (2013:11) point out the responsibility of the PEs and write:

For each subject paper, the principal examiners were responsible for establishing and setting the standards for marking using professional judgement about how to interpret and apply the marking scheme. The principal examiner's judgement on these issues is always final.

This statement shows that the standardisation procedures in the UK place too much trust on the PEs by giving them the sole responsibility of setting marking standards. The interpretation and application of the mark schemes will depend on the competence of the PEs. As mentioned earlier, Zimbabwe, a former British colony, adopted the examination procedures from the UK, and is likely to place the same responsibility on PEs. The emphasis on timelines implies that the standardisation meetings have to be completed within stipulated deadlines so that the examiners

proceed to mark the scripts within the stipulated deadlines. This has implications on the mastery of marking standards and the monitoring of examiners during marking, hence the need to study standardisation meetings in the Zimbabwean context.

The UK examination regulatory body, Ofqual (2014b:4), conducted a literature review to determine the impact of different forms and stages of standardisation methods on marking reliability, and to establish international trends in standardisation of examinations. It searched for the literature from the educational research data bases, assessment specialist websites, known experts in examinations and research journals. The results of the study indicated that the examiners preferred face-to-face to online standardisation, but the standardisation approach had no influence on the marking reliability. The examiners believed that the face-to-face standardisation creates a shared understanding and a community of assessment practices that positively influence marking reliability. Empirical evidence shows that the community of practice does not have an impact on marking reliability (Ofqual, 2014b:4). The standardisation methods vary with countries, depending on convenience and cost. Ofqual (2014b:30), however, is cautious about the assertion that online standardisation does not have adverse impact on quality of marking, explaining thus:

Is it possible that the early positive results of online standardisation are ‘riding on the coat tails’ of pre-existing quality assessment practices, but that over time, the shared understanding and community of practice effects will be eroded, to the detriment of marking reliability? However, in contrast, may be initial negative reactions to new ways of working will be overcome with time, and online standardisation will lead to higher quality in the medium to long term.

Negative attitudes of examiners towards online standardisation could be a result of resistance to change and might not, in the early stages of transition from face-to-face to online standardisation, influence the quality of marking. An examination board may choose between face-to-face and online standardisation, or combine the two in one examination. There is, however, a need for further research on the influence of online standardisation using examination scripts that have been marked by examiners who never attended face-to-face standardisation.

Tisi et al (2013:2) reviewed the literature on reliability of marking in order to identify advances which have been made in improving and quantifying the marking reliability. They reviewed the literature published between 2004 and 2012. The literature search focused on documents published internationally, which were available in English and covered national examinations as well as teacher assessments. The search identified 240 sources, which were screened, resulting in 28 key sources that answered the research question. The review established that the accuracy of marking is affected by the nature of test questions and mark schemes, characteristics of markers and the process used for marking. There was consensus that the reliability of marking could be improved by constraining the mark schemes, selecting examiners with the relevant marking experiences and subject expertise and use of the OSM technology which allows item level marking and continuous monitoring of examiners (Tisi et al, 2013:3). The review, however, also established that there were conflicting results on the impact of standardisation on marking reliability. Some studies established that examiners who went through standardisation marked as accurately as those who did not, concluding that examiners marked accurately when they were provided with the mark scheme. Other studies established that examiners who went through standardisation marked more accurately than those who did not (Tisi et al, 2013:27). However, the quality of marking might be compromised if examiners feel that they have not mastered the marking standards as established by Falvey and Coniam (2010:14). There is therefore need for further research on the impact of standardisation on the quality of marking.

At the standardisation meeting, examiners mark samples of pre-marked scripts where their marking is assessed by team TLs. The examiners will only be allowed to begin the actual marking when the TLs are satisfied with the accuracy of marking the sample scripts (Ofqual, 2013; Tze Ho & Chong Sze, 2013), the stage referred to as approval in the marking procedures. As soon as live marking begins, senior markers start to monitor the quality of marking.

2.9 Monitoring quality of marking

As discussed earlier, literature reviewed by Ofqual (2014c:6) established that there are three methods of monitoring the marking of examinations, and these are double marking, multiple marking and sampling and regular sampling. The researchers noted that the use of double marking and multiple marking to produce final scores is an acknowledgement that legitimate

differences in opinion can exist between the markers. However, in the UK multiple and double marking are not commonly used (Ofqual, 2014c:6). This is supported by Pinot de Moira (2013:4), who posits that even though multiple marking is believed to be the most effective method of ensuring the final mark awarded is consistent with the mark scheme, costs and time constraints make it unsustainable in high stakes examinations in the UK. Hudson (2009:3) discusses double marking and regular sampling as the main methods of monitoring marking in PBM. It can therefore be argued that there are two functional methods of monitoring examinations at reasonable costs and within reasonable time frames, regular sampling and double marking.

2.9.1 Regular sampling

In regular sampling, a sample of an examiner's work is moderated (remarked) by the senior examiner to ensure the consistent and accurate application of the mark scheme (Ofqual, 2013:17). In Uganda, TLs were required to moderate ten per cent of the scripts in each envelop to ascertain consistency in marking. The acceptable deviation between the examiner and the TL was ± 2 (Bukonya, 2006:1). Research shows that regular sampling is useful when marking highly objective responses for subjects such as Mathematics, and long responses with highly specified mark schemes (Coniam, 2011b:2). Regular sampling in PBM has its own challenges.

Several drawbacks of regular sampling and remarking of scripts as a way of monitoring quality of marking have been described. When scripts are sampled at whole script level, individual examiner bias is retained. The sample scripts are chosen by the marker who might have paid special attention to their marking, and not to the marking of the non-sample scripts. True scores are determined by the senior markers who may also be biased. The frequency of sampling and the number of scripts are limited by the logistics of moving scripts between examiners. Poor marking that remains at the end of marking needs to be corrected by remarking or statistical adjustment (Hudson, 2009:3). One method of correcting poor marking at the end of marking is to engage script checkers who are not examiners, to go through the scripts and identify marking errors and unmarked portions. The examiners or team leaders are requested to correct the errors or mark skipped portions (Bukonya, 2006:1).

The challenges associated with regular sampling and moderation of scripts could compromise the quality of marking when senior markers decide not to moderate scripts due to time constraints or when they are not competent enough to set credible true scores. The examiners might decide not to revisit their scripts when instructed to do so by senior markers, due to time constraints. These challenges might have necessitated the use of the OSM technology to enhance the mechanisms of monitoring the quality of marking. Double marking is another method of monitoring quality of marking.

2.9.2 Double marking

Double marking is when all candidates' scripts are marked by two examiners and may need a third for adjudication, and a fourth marker when serious discrepancies arise. The closest of the pair of marks will be taken as the final mark for the response concerned (Tze Ho & Chong Sze, 2013:9). Double marking is generally used to monitor the marking of open-ended questions with the mark schemes that leave room for the examiner's interpretation of the candidate's responses (Tze Ho & Chong Sze, 2013; Coniam, 2011b). Literature shows that double marking is dying a natural death in PBM because it requires large numbers of examiners who cannot be paid by examination authorities, as explained by Hudson (2009:3) who stated:

Setting up double-marking processes in a paper-based environment is complex and costly in its own right. Those awarding bodies internationally that have achieved this have well-thought out systems, but these are surrounded by teams of administrative staff supporting the process. Double-marking almost always takes place in a marking centre, the sampling of markers' marking and the adjudication of difference between one marker and another tends to take place as marking takes place. This adds stress and the risk of error because of the logistical and time constraints that exist.

These logistical challenges of double marking were overcome by the use of the OSM technology, which enables immediate random distribution of responses to examiners for marking at item level.

2.9.3 Monitoring quality of marking in the OSM environment

Literature shows that there are three mechanisms of script moderation in monitoring the quality of marking in the OSM environment. Double marking is where all scripts are marked by two examiners (Tze Ho & Chong Sze, 2012; Coniam, 2011b) or a percentage of the scripts are marked by two examiners (DRS, 2015; DRS, 2013; Hudson, 2009). Pinot de Moira (2013:25) calls it peer-pair quality assurance system. When the two examiners award marks that differ beyond a set tolerance range, the first examiner is stopped from marking. In Hong Kong, all scripts for Liberal Studies were marked on screen by two examiners, where the two marks were automatically compared. A discrepancy of more than 20% prompted the system to send the response to a third marker for adjudication (Tze Ho & Chong Sze, 2013; Coniam, 2011b). The HKEAA conducted research to ascertain the reliability of double marking of Liberal Studies in the OSM environment in which it was established that it was high (0.8) (Tze Ho & Chong Sze, 2013:11). Seeding is another mechanism of monitoring the quality of marking.

Seeding is where senior examiners mark the sample standard scripts which are automatically presented to examiners at set intervals for blind marking. Examiners who deviate from the pre-determined mark beyond a set tolerance are automatically stopped from marking (DRS, 2015; Ofqual, 2013; Pinot de Moira, 2013; Hudson, 2009). According to Pinot de Moira (2013:6) the seed system is hierarchical in that the senior marker's mark would be defined as the true mark. If the marker failed to mark the seed accurately then the final mark would be that of the senior marker. The author argues that reliability, in this context, would be the level of agreement with the senior. There is, therefore, need to set a tolerance range for the difference between the marker and the senior marker.

Back-reading is where a senior examiner re-marks sample scripts that have already been marked by a marker and penalises the markers who deviate beyond tolerable ranges. It is used for small entry examinations where seeding or double marking cannot be used (DRS, 2015; Ofqual, 2013; Pinot de Moira, 2013; Hudson, 2009).

Pinot de Moira (2013:1) used a mathematical simulation to explore the seeding and percentage double marking models of quality control when used in item and whole script levels. The paper

focused on the effect of item marking on the probability of awarding some true mark and the effect of sampling on the probability of identifying an errant marker. The research used marks from two subjects, Religious Studies and English. The Religious Studies scripts were marked on screen by the Assessment Qualification Alliance (AQA) in 2008. The Religious Studies examination was a short answer paper with 21 items totalling 83 marks (Pinot de Moira, 2013:9). The English paper was sat in the summer of 2006. The examination was made of two essays each worth 27 marks. The data about the English papers were obtained under experimental conditions while the data for Religious Studies were obtained in a live examination (Pinot de Moira, 2013:10). Simulation data were produced for the two subjects and used in seeding and double marking.

On comparing the Religious Studies and the English papers in the hierarchical quality control system, the study established that the short answer paper would be marked more reliably than the essay paper. For Religious studies only 13% of the candidates would be awarded the true mark but nearly 80% would be within 4% of the true mark. It was concluded that almost all of the Religious Studies candidates would be awarded marks within 10% of the true mark. For English, 13% of the candidates would be awarded a true mark and only 40% would get a mark within 4% of the true mark (Pinot de Moira, 2013:12). The study provides evidence that the seeding approach to quality control is most appropriate for short answer question than for essays. The study also provides evidence that the nature of questions in an examination influence accuracy of marking, as will be discussed later in this chapter.

The scholar concluded that a quality control system that includes any element of sampling cannot directly influence marking reliability, but can only identify the deviating markers for retraining. Pinot de Moira (2013:24) emphasised thus:

The key, then, is to maximise the probability of identifying the errant marker and then to take appropriate remedial action. In so doing, marking reliability can be improved indirectly by a feedback mechanism. Indeed training has been shown to have a key role in reducing marking errors even though optimal feedback mechanisms remain ambiguous.....but, no matter which feedback mechanism is implemented, it remains the case that any quality assurance system must first identify the errant markers efficiently.

The study also concluded that item marking minimises the effect of systematic differences between markers but the extent to which this improves reliability at a paper level is dependent upon the number of items on a paper; and that peer-pair quality assurance (percentage double marking) has higher chances of picking the errant markers than seeding. To emphasise the advantage of peer-pair quality assurance Pinot de Moira (2013:25) argued thus:

The advantages of a peer-pair quality assurance system undoubtedly include the alleviation of time pressures at the beginning of the marking period and a move towards a more consensual view of the true mark. However, these advantages come at the expense of a single gold standard and an increased workload throughout the marking period.

This statement implies that scripts for peer-pair marking are not selected and discussed at pre-standardisation and standardisation meetings like seeds, reducing work for senior markers. Some studies reported on the examiner concerns about quality control in the OSM environment.

The examiners who participated in a pilot study conducted by the UCLES in January 2004 were concerned that the method of monitoring marking quality in the OSM environment smacked of 'Big Brother' and was deceitful. They felt the seeds were looking for deviations from the correct marks (Raikes et al, 2004:19). Coniam (2011a:1045) established that some examiners felt the quality control mechanism was useful, and one noted that they were not adequately trained to access and interpret their marking statistics. Some of the studies reporting on quality control in the OSM environment were pilot researches, with small samples, that were addressing research questions other than the OSM quality control system. These studies may not adequately inform the monitoring of marking examinations in Zimbabwe due to the differences in contexts and subjects. There was therefore a need to explore the mechanisms of monitoring the quality of marking O Level Biology (5008) in the Zimbabwean context. The quality of marking is also influenced by the nature of questions and mark schemes in the examination.

2.10 Examination questions and marks schemes

Questions or test items and mark schemes influence the quality of marking candidates' scripts. To emphasise the importance of questions and mark schemes, Ahmed and Pollit (2011:259) assert:

At the heart of every assessment lies a set of questions, and those who write them must achieve two things. Not only must they ensure that each question elicits the kind of performance that shows how good pupils are at the subject, but they must also ensure that each mark scheme gives more marks to those who are better at it.....It is futile to design excellent assessment tasks if an equal amount of care is not put into the design of the mark schemes that govern the marking process; good questions will be wasted if the evidence they elicit is not judged appropriately.

This statement implies that mark schemes provide the criteria for judging the candidates' responses to examination questions and to awarding marks, and are designed together with the question paper. The quality of test items and their mark schemes can threaten the validity of the test results and their interpretation. Ahmed and Pollit (2011:260) explain how the questions and mark schemes can pose as validity threats. According to these authors, the question may not elicit the desired responses from the candidates; different examiners may score the same response differently, causing marker unreliability; and examiners may credit features of the responses that do not reflect the intended achievement construct. The importance of examination questions and mark schemes in the quality of marking has resulted in intensified research on how they influence marking reliability and, hence quality.

Ahmed and Pollit (2011:265) developed taxonomy of mark schemes using sets of examination questions from Scotland, England and Northern Ireland. Three subjects, Geography, Business Studies, and Design & Technology were chosen for the variety of the question types they used and were believed to be difficult to assess. A total of 4843 items from Scotland and 1913 from England and Northern Ireland were analysed, resulting in a taxonomy of mark schemes that could help the examiners. The authors identified the types of questions that varied from highly objective to very subjective. On the basis of the mark schemes designed for the item types, they

categorised the mark schemes into a general taxonomy that is determined by the extent to which it helps examiners to assign marks to learners' responses as indicated in Table 2.1.

Table 2.1: General taxonomy of mark schemes (adapted from Ahmed & Pollit, 2011:266)

Level	Description of the taxonomy
Level 3	<ul style="list-style-type: none">• Define a principle for discriminating better from poor performances• Offers guidance on how to credit every possible answer
Level 2	<ul style="list-style-type: none">• A description of good and poor performance• May offer adequate guidance if the prediction is accurate
Level 1	<ul style="list-style-type: none">• A description of good performance• Offers no guidance on difficult border line responses
Level 0	<ul style="list-style-type: none">• Offers no guidance to assigning marks to responses

The mark schemes were further assigned to three categories, with each one made of levels zero to three. They were categorised by the degree of constraint in the question: very constrained, semi constrained and unconstrained as determined by the nature of the question. Ahmed and Pollit (2011:267) describe the mark scheme categories as follows:

In the simplest category– very constrained (VC) questions – the function of the mark scheme is to define the boundary between right and wrong or between scores of 1 or 0. For semi- constrained (SC) questions the range of responses is greater and the function of the mark scheme shifts to helping markers whether a particular response shows enough evidence of correctness or goodness to merit a mark, and perhaps to help them decide how many of the available marks it should be given. With unconstrained questions (UC), the concept of correctness may fade almost completely away, and the function of the mark scheme becomes mainly to help markers rate the quality of the responses they see.

Ahmed and Pollit (2011:267) emphasise that it is not necessary to design all mark schemes to the top level of the taxonomy because some mark schemes serve the purpose of the examination. They, however, posit that it is better to err on the side of too much than too little help for the markers. This taxonomy of the mark schemes, combined with other mark scheme features investigated in other studies, provided the criteria for assessing the type of questions and of mark

schemes used in O Level Biology examinations and how they could influence quality control in the OSM environment.

Another research on mark scheme features was conducted by Bramley (2008).

Bramley (2008:2) conducted a study to code salient features of examination questions and their mark schemes and to investigate the link between the coded features and the level of marker agreement. The marker agreement data came from the live marking of a wide range of subjects in the June and November 2006 examinations. Marking was monitored by team leaders re-marking sample scripts of their team's allocation, and the team leader's mark was considered as the true score. The scripts were marked at item level by different examiners. The coded features were the maximum possible mark for the question; the type of mark scheme, objective, point based or level-based; the amount of space where candidates could write their answers; the ratio of valid points to the marks available; whether the mark scheme specified qualifications, restrictions or allowed variations to the correct responses; and whether the mark scheme specified wrong answers. The level of agreement was measured by the simple P_0 statistics (the proportion of cases with exact agreement between the team leader and the examiner). The size and significance of the effect of the coded features was assessed by logistic regression modelling.

The study established that objective and point based mark schemes showed the same level of marker agreement for marks of magnitude ten and below. After allowing for the maximum mark, the amount of constraint in the correct responses was strongly related to the marker agreement, with objective items having three percentage points higher agreement than point-based items worth the same marks. The gap widened as the number of marks increased. The level of agreement decreased if there were more valid points than maximum marks in the mark scheme. The research concluded that most of variations in the marker agreement can be explained by the maximum possible mark, the mark scheme type and ratio of valid point to maximum possible mark (Bramley, 2008:2). The results on the ratio of the valid points to available marks refutes Ahmed's and Pollit's (2011:267) belief that the inclusion of more valid point would provide better guidance to the examiners.

The results of Bramley's study were significant to the study of the practice of quality control in the OSM environment because the data used came from live examinations and not from a pilot study. The results may also have an implication for the validity of examinations marked on screen. The highly objective mark schemes with the highest level of agreement between the markers could tempt the examination authorities to increase the proportion of such items in the examinations for the sake of increasing marker agreement, hence reliability, a move that compromises the validity of the assessments, given the inverse relationship between validity and reliability (Ahmed & Pollit, 2011; Meadows & Billington, 2005). It was therefore important to study the nature of questions and mark schemes for O Level Biology examinations marked on-screen in Zimbabwe to establish their possible influence on quality of marking.

Another study on the mark scheme features was conducted by Child, Munro and Benton (2015:5), who investigated how the mark scheme features influenced the reliability and general quality of marking as measured in terms of mark distribution, degree of agreement between the examiners and the PE, and examiners' perceptions of mark scheme usability. The PE for GCSE English Language – Information and Ideas, was recruited to design the two-mark schemes for a set of questions and to train examiners. The original mark scheme contained the same content and features as the mark scheme used in the live session in June 2014. The experimental mark scheme contained the same content as the original mark scheme, but had a number of features manipulated by changing the positioning of the guidance in relation to the questions, the salience of key terms and page formatting. Twenty examiners were recruited to mark 150 scripts comprising two questions (2 and 4) from the June 2014 English examination. Ten examiners were trained to use the 'original' mark scheme and ten to use the experimental one. The examiners attended a standardisation meeting where the PE provided information on how to apply the mark schemes correctly, using a sample of ten exemplar scripts.

The results indicated that the experimental mark scheme did improve the reliability of marking for question 2 significantly. Question 4 results indicated that the experimental mark scheme significantly improved the reliability in terms of the agreement between the markers and the PE, and between the markers themselves. The results for question 4 indicated that the experimental

mark scheme seemed to encourage the examiners to use a greater range of marks, and that this increase resulted in a greater proportion of variance attributed to the true score than to error. The change in the distribution of the scores was interpreted to imply that inconsistency in marking may have a smaller effect on grade outcomes when using the experimental mark scheme. The researchers, however, noted that there was not enough evidence to conclusively assert that the experimental mark scheme improved reliability of marking GCSE English Language. The examiners, however, related the features of the experimental mark scheme to their perception of usability and cognitive processing. These features included the bolding of key terms, the proximity of level descriptors to guidance and the one- page formatting of the experimental mark scheme. The study concluded that the changes to some mark scheme features are worthy of future consideration, with respect to improving the mark scheme usability and the consequent overall quality of marking.

The study was conducted using a level-based marking scheme, which was unconstrained, and the results seem to support Ahmed's and Pollit's (2011:267) belief that it is better to design the mark schemes that provide more guidance to examiners than highly constrained mark schemes that offer no guidance. The results of the study also provided some criteria for assessing features of mark schemes used in the marking of O Level Biology and their influence on quality control in the OSM environment.

2.11 The conceptual framework

The literature review indicated that the quality of marking is influenced by factors such as examiner training, pre-standardisation, standardisation, examiner monitoring and the nature of examination questions and mark schemes. The interaction of these factors is shown in the conceptual framework in Figure 2.4, which resulted from the deliberations on the pertinent concepts guiding this study – the conceptual understanding of OSM impacts quality control of the same. Hence, the practice of quality control in the marking of O Level Biology on computer screens was studied using this framework.

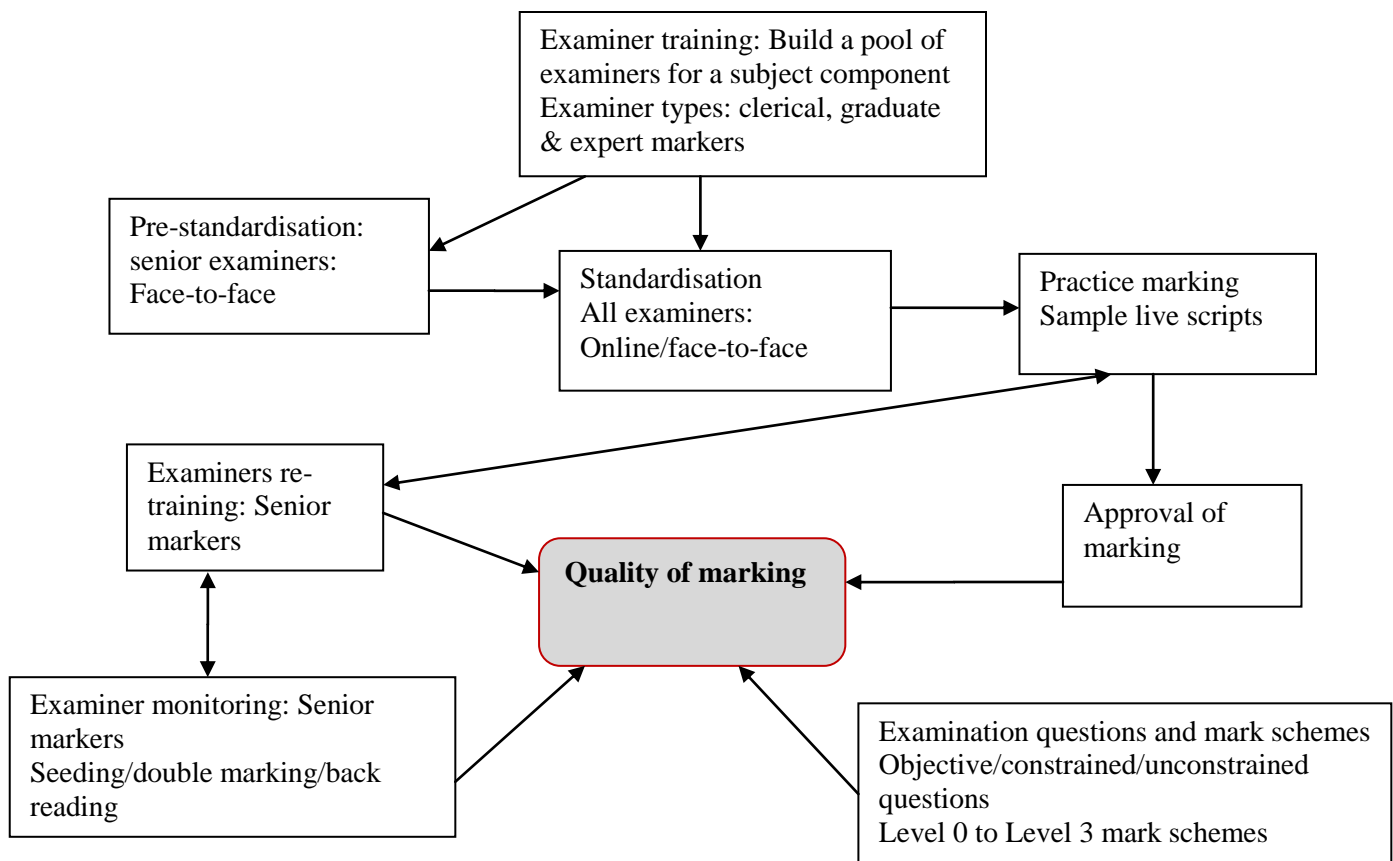


Figure 2.3: The conceptual framework for studying quality control in OSM environment

2.12 Conclusion

This chapter reviewed scholarly literature that guided the conceptual framework for the study of quality control in the OSM environment. The literature review established that the concept of educational assessment is premised on the principle of scientific measurement, guided by the assumption that anything that exists can be quantified. This assumption led to the development of procedures that guide educational assessment. There are four types of assessments, school based, national assessments, international assessments and public examinations. This research focused on public examinations which backdate to ancient China, where they were used to select candidates for appointment into public service posts. Since then, examinations have been surrounded by controversy that led to camps of proponents and critics. Despite the criticism examinations continue to dominate education systems because they are the only criteria for

making decisions about learners' performance. Procedures for the design, administration, marking and grading of public examinations are continually being refined to improve the quality of examinations, hence the purpose of this research.

Scholarly literature shows that quality of marking is determined by examiner recruitment and training; standardisation of marking; monitoring of marking; and the nature of examination questions and mark schemes. Examination authorities around the world recruit examiners who are practicing or retired teachers and train them to build examiner teams with hierarchies. Senior examiners set marking standards, pass them on to other examiners through standardisation meetings and monitor examiners during marking. Examiners are only allowed to mark live scripts after senior examiners are satisfied of their competences. In the OSM environment, quality of marking is monitored by double marking, seeding and back reading. The conceptual framework that guided the study of quality control in the OSM environment was designed pictorially to depict the interaction of the factors that determine the quality of marking. The next chapter reviews literature on the OSM technology.

Chapter 3

The onscreen marking technology

3.1 Introduction

Chapter 2 presented the discussion on the scholarly literature that guided the conceptual framework for the study of quality control in the OSM environment. As discussed in the introduction of Chapter 2, the impact of technology on education has resulted in the emergence of e-assessment, which involves the assessment tasks or processes designed, accessed and stored through the medium of ICT (Isaacs et al, 2013:41) and the OSM technology was designed to automate the existing marking procedures without changing them. The current chapter concerns scholarly literature review on the development and adoption of e-assessment in the public examinations so as to place the OSM technology into the context of this study. The chapter then explores the nature of the OSM, its use by the examination boards in other countries and in Zimbabwe specifically. The chapter also reviews research studies on several issues related to the OSM technology, its advantages and challenges. E-assessment emerged as a result of challenges associated with paper-based examinations and its use in public examinations is discussed in the next section.

3.2 Development of e-assessment

The advent of ICT has greatly changed the way of living, learning and working. ICT has improved access to information through the internet; communication through cheaper and faster ways such as text messaging, mobile phones, video conferences, Skype and emails (Boyle, 2010:1). Improved access to information and communication ushered in e-commerce, where goods and services can be obtained electronically, with sales personnel being trained online or through e-learning (Adewo, 2012; Bennett, 2002). Learners, therefore, need to acquire new skills that enable them to survive in the 21st century. In addition to the subject content that is taught in the classroom, 21st century skills include life and career skills, critical thinking, communication, collaboration, creativity and ICT skills (Kurshan, 2017; Partnership for 21st Century learning [P21], 2007). Technology enhances the acquisition of these higher cognitive skills (Pellegrino &

Quellmalz, 2010; Winkley, 2010; Erstad, 2008; Ridgeway, McCusker & Pead, 2007), hence the need to use technology in the assessment of those skills.

Literature suggests that learners are natural users of digital technologies in their everyday life and for learning, but their skills are assessed the traditional way, on paper. Tucker (2009:1) lamented the assessment of digital learners on paper and wrote thus:

Students are growing up in a world overflowing with a variety of high-tech tools, from computers and video games to increasingly sophisticated mobile devices. And unlike adults, these students don't have to adjust to the information age – it will be all they've ever known. Their schools are gradually following suit, integrating a range of technologies both in and outside of the classroom for instructional use. But there's one day a year when laptops power down and students' mobile computing devices fall silent, a day when most schools across the country revert to an era when whiteboards were blackboards, and iPhones were just a twinkle in some techie's eye – testing day.

Tucker's statement suggests that assessment is out of sync with everyday life and learning and should be harmonised. Boyle (2010:4) concurs with Tucker (2009:1), arguing that paper examinations deny candidates tools that they have always used when composing text, solving mathematical problems and finding information. A call to use technology in assessment has, therefore, grown louder, leading to the evolution of e-assessment. E-assessment dates back to the 1950s when optical scanners were used to score multiple choice responses, and these were replaced by computers in the 1960s (Weiss, 2011:2). The author posits that the optical scanners could not analyse items, so other machines had to be used to obtain the basic frequency counts and other statistics. Computers allowed more reliable and faster scanning as well as item analysis. According to Weiss (2011:2), the introduction of the personal computer (PC) in the 1980s brought major changes to the way tests were designed, stored, analysed and delivered.

As technologies advanced with time, item banks were conceptualised and designed to maintain testing programs (Weiss, 2011:4). The development of item banks led to the electronic test delivery (computer-based testing [CBT]) in the 1970s, eliminating both printed tests and answer sheets (Weiss, 2011:10). The electronic test delivery led to the development of on-demand tests

such as the City and Guilds online test that can be booked by candidates whenever they are ready to take it; and adaptive tests, where tasks are changed according to the progress made by the task taker (Winkley, 2010; Ridgeway, McCusker & Pead, 2007). The PC therefore improved the efficiency of test design, storage and delivery.

The use of ICT in public examinations brought a variety of e-assessments that were summarised by the Qualifications and Curriculum Authority [QCA] (2007:6) as follows:

- Assessments that are distributed, completed, marked automatically and administered electronically using local intranet/networks and individual workstations;
- Assessments that are distributed, completed, marked automatically and administered electronically using the internet;
- Assessments comprising a combination of automatic marking and manual marking that are delivered in either of the two ways described above;
- Electronic test delivery, with all marking completed manually on screen or on paper;
- A range of multimedia formats for submitting assessment;
- Electronic scanning of completed assessments for marking;
- Tests downloaded from the internet by the centre;
- Delivery of assessments and submission of completed assessments by secure email;
- E-portfolios to store and manage candidates' evidence electronically; and
- Assessments that are automatically marked and react adaptively to student performance.

However, the adoption of e-assessment in public examinations has not been easy, given the controversy that surrounds high-stakes assessment, as discussed in Chapter 2 Section 5. E-assessment has, therefore, been used to some varying degree by examination authorities due to the challenges associated with high stakes examinations. The following section discusses the challenges that militate against the adoption of e-assessment.

3.3 Challenges of e-assessment

Bennett (1998:1) of Educational Testing Services (ETS) predicated a three-generation model for adopting e-assessment which would be distinguished by the purpose of testing, test format and

content and the extent to which the tests make use of technology. Bennett (2015; 1998) summarises the characteristics of each generation, which receive brief explanation subsequently.

3.3.1 First generation computer-based tests (infrastructure building)

The assessments primarily serve institutional needs; measure traditional skills; use test designs and formats closely resembling paper-based tests, except that they are given adaptively; administered in dedicated test centres as a ‘one-time’ measurement; take limited advantage of technology (Bennett, 1998:22). Bennett (2015:371) argues that although much about first-generation e-assessment is traditional, they can capitalise on technology when used in computer adaptive testing, where the next item is selected based on the student’s competence level. Bennett, therefore, envisages that first generation e-assessments would be delivered via internet to the computer screens and the candidates would respond on computers. The adaptive nature of the tests meant that they would be different from candidate to candidate, violating the uniformity requirement of standardised tests, and attracting the wrath of critics, as discussed in Chapter 2, Section 4. The next-generation e-assessments were expected to be more advanced than the first-generation.

3.3.2 Next-generation electronic tests (qualitative change)

The next-generation e-assessments also serve institutional needs; use new item formats; automatic item generation; automatic scoring; networks to make assessment an integral programme component; measures skills; administered in dedicated test centres as ‘one-time’ measurements; allows candidates to interact with assessment agents entirely electronically (Bennett, 1998:22). Bennett (2015:371) emphasises that the next-generation e-assessments are driven by qualitative change and improvement of efficiency. The first two generations are restricted to test centres and measure performance once. Bennett predicated a third generation that he called the generation ‘R’, which means reinvention.

3.3.3 Generation R tests (Reinvention)

These assessments serve both institutional and individual purposes; are integrated with instruction through electronic tools to allow repeated performance sampling over time; are designed according to cognitive principles; use complex simulations that model real

environments and allow natural interaction with computers; are administered at a distance; and assesses new skills (Bennett, 1998:22). According to Bennett (2015:372), the generation R e-assessments are a radical departure, with distance learning assessment completely embedded in electronic curricular.

Isaacs et al (2013:42), however, argues that the mainstream large scale assessments are mostly still stuck in the first generation, with tentative steps having been made in the next generation. This argument is supported by Pellegrino and Quellmalz (2010:120), who posit that the new technologies in public examinations have focused on logistical efficiency and cost reduction, advocating for Bennett's (1998:22) next and generation R e-assessments thus:

A new generation of innovative assessments is pushing the frontiers of measuring complex forms of learning. The computer's ability to capture student inputs permits collecting evidence of processes such as problem-solving sequences and strategy use as reflected by information selected, numbers of attempts, approximation to solutions, and time allocation.

Pellegrino and Quellmalz (2010:122), however, acknowledge that the use of innovative technologies in high-stakes assessments in the United States of America (USA) is faced with accountability, regulatory, economic and logistical challenges, leading to computer-based assessments based on simple, highly structured questions that test factual recall. These challenges were cited for CBT implementation by the QCA in the UK. In April 2004 the Chief Executive of the QCA, Dr. Ken Boston, published a five-year programme for the implementation of on-demand e-assessment enumerated by Kingdon (2014:3) thus:

- Seventy-five percent of the Basic and Key Skills tests for Levels 1 and 2 of the contemporary national qualifications framework (NQF) were to be delivered on-screen by 2005;
- Field trials of the first on-screen GCSE subjects were to begin in 2005;
- The three English unitary awarding bodies were expected to offer the first on-screen GCSE examinations by 2006;
- Codes of practice plus audit and regulatory criteria were to be in place for 2007;

- Ten percent of GCSE examinations were to be delivered on-screen by 2007;
- Introduction of the first on-demand GCSE examinations was timed for 2008; and
- On-screen, on-demand delivery of GCSE examinations was to be the norm by 2009.

Kingdon (2017:3) explains Dr. Boston's reasons for the five-year plan thus:

He reasoned that new ways were required to focus markers' work, reduce their clerical burden and reform the time scales to which they worked. E-assessment, especially the on-screen, on-demand delivery of qualification units, was seen as the means of reforming contemporary assessments, examinations and qualifications system into something that better met the needs of learners and other stakeholders. QCA accepted from the start that, eventually, use would be made of automatic rating of students' extended written answers, with on-screen marking by clerical or expert markers when automatic rating systems were not considered to be appropriate.

Despite the noble intentions, the plan faced a huge regulatory challenge when multiple choice examinations were banned for some subjects; there was no relevant expertise; and there were financial risks that could not be taken (Kingdon, 2014:7). The QCA approved Dr. Boston's plan fully aware that automatic marking of extended essays might not be acceptable and made a fallback plan of OSM. Dr. Boston's plan clearly followed the e-assessment model predicted by Randy Bennett (Bennett, 2015; 1998). However, the OSM which served as the fallback plan is a radical departure from Bennett's model, and may provide a better alternative to CBT in high stakes examinations. This confirms Isaacs' et al (2013:42) argument that the mainstream large scale assessments are mostly still stuck in the first generation, with tentative steps having been made in the next generation. Winkley (2010:4) also confirms that high-stakes CBT in the UK public sector education remains more the exception than the norm, but have been used to evaluate the IT engineers and aircraft pilots by professional organisations such as Microsoft, Cisco and the Federal Aviation Administration.

The critics of CBT have raised security concerns, especially the failure to authenticate test takers. Khlifi and El-Sabagh (2017:62) posit that the main challenge facing the security of e-assessment is how to authenticate students because unauthorized persons can access and manage information. The security concerns that surround high-stakes examinations seem to be

compounded when the examinations are taken online, intensifying the controversy of high-stakes examinations that were discussed in Chapter 2, Section 5. Khlifi and El-Sabagh (2017:62) went on to propose a security scheme that could be used to authenticate students taking online tests. The details of the proposed scheme were outside the focus of this study and were therefore not discussed in detail.

Way back in 1998, Bennett acknowledged that very few first-generation e-assessments were offered in America, but anticipated that the volumes would dramatically increase (Bennett 1998:9). However, the challenges discussed earlier, coupled with the controversy that surrounds high stakes examinations, have militated against the adoption of e-assessment in the manner predicted by Bennet (2015; 1998). The low acceptance of innovative assessment technologies in educational assessment could be a result of the need to follow assessment procedures put in place by examination authorities as discussed in Chapter 2 Section 5, resulting in the rejection of technologies that seem to be a radical departure from the procedures.

Zimbabwe has adopted a number of assessment technologies that automate existing processes without reconceptualising them. As mentioned earlier in Chapter 1 Section 2.4, the ZIMSEC adopted the OSM technology for the June 2012 examinations. In the same year, the Council designed and implemented an e-registration programme, where candidates' details are electronically captured at the examination centre and passed on to the council for processing (Personal communication, 15 March 2018; The Herald, 2015; Karombo, 2012). An electronic programme with authoring tools that support the practical operational process of the examination authoring was adopted in 2016, and will enable the ZIMSEC to digitise the existing item bank, implement richer qualitative and quantitative review methodologies, and start to build banks of items to strengthen the authoring process (www.grademaker.com/about/news). Later in the same year the Council designed and piloted a mark capturing system that was used in live examinations for the first time in June 2017. The mark capturing system allows examiners to capture the candidates' marks directly onto the system rather than on marksheets that would be scanned later (ZIMSEC, n.d; Internal communications, 23 June 2017; 10 November 2017; 14 May 2018). These technologies seem to be tentative steps into Bennett's (2015; 1998) model of e-assessments.

The ZIMSEC might be able to adopt some of Bennett's (2015; 1998) first generation e-assessments if the effort put by government to lure investors into the education system come to fruition. The Ministry of Primary and Secondary Education (MoPSE) held a one day conference dubbed the Presidential Indaba on 18 July 2018. The purpose of the conference was to lure investors in the educational infrastructure development that would support the implementation of the new curriculum (MoPSE, 2018; Invitation letter, 11 July 2018; personal experience). The then Permanent Secretary in the Ministry of Primary and Secondary Education, Dr. Sylvia Utete-Masango, enumerated in the invitation letter to the conference, the additional infrastructure that was required to support the newly introduced competence-based curriculum, i.e. science laboratories, technical and vocational workshops, ICT infrastructure, renewable energy sources and clean water, and sporting facilities.

Speaking at the conference, some investors pledged to digitalise 92% of rural schools. Others pledged to digitalise 20 schools per year for the next five years. Sadly, there was no reference to investment in assessment and assessment technologies at the conference. It appeared as if assessment is not part of education even at classroom level (Personal experience). This failure to link assessment to teaching and learning could result in the unfortunate situation where candidates have to put their digital tools away and brace for paper examination (Boyle 2010; Tucker 2009). However, digitalisation of schools could enable the ZIMSEC to ride on the ICT infrastructure to administer computer-based multiple choice tests. This study, however, focused on the OSM only. The next section gives an overview of the OSM technology subsequently.

3.4 Overview of the OSM technology

The OSM technology is a platform that is used to mark paper-based examinations on computer screens. After the candidates have sat the examination their scripts are taken to a scan centre where they are scanned and their images saved and distributed to the examiners for marking on computer screens (Fowles, 2011; Coniam, 2011a, Coniam & Yeung, 2010). Newgen (n.d.: www.newgen.net) refers to scanning as the digitisation of the answer scripts, emphasising that the answer scripts are scanned using the high speed scanners and uploaded onto the OSM system for

further processing. The Hong Kong Examinations and Assessment Authority [HKEAA] (2015: www.hkeaa.edu.hk) summarises the marking process in the OSM environment as follows:

- Examination scripts are completed by candidates;
- Examination scripts are collected from examination centres;
- Examination scripts scanned and images saved;
- Images of examination scripts are distributed to markers for marking via a secure intranet system; and
- Marks and annotations by markers captured by the OSM system.

Chapter 2 discussed the marking process used by the examination boards around the world. Ofqual (2014g: 17) posits that the marking process is generally the same. Any OSM software should therefore be designed to improve the efficiency of the marking process described by scholarly literature and summarised in the conceptual framework of this study in Chapter 2, Figure 2.3. The OSM system is made of two main modules, the administrator and the evaluator/marker. The next sub-sections discuss the administrator and the marker modules in relation to quality control in the marking process.

3.4.1 The administrator module

Ramakrishna et al (2012:16) and DRS (2013:5) concur that the administrator module has a log-on screen that requests a username and a password before displaying a home page. Ramakrishna et al (2012:16) lists the functionalities on the home page of the administrator module: add/delete/edit subject, add/delete/edit/evaluators, assign subjects to evaluators, scan and add scripts, allocate scripts to evaluators, add marks to schema, generate reports, upload sample scripts, and reports. Figure 3.1 illustrates the administrator module.

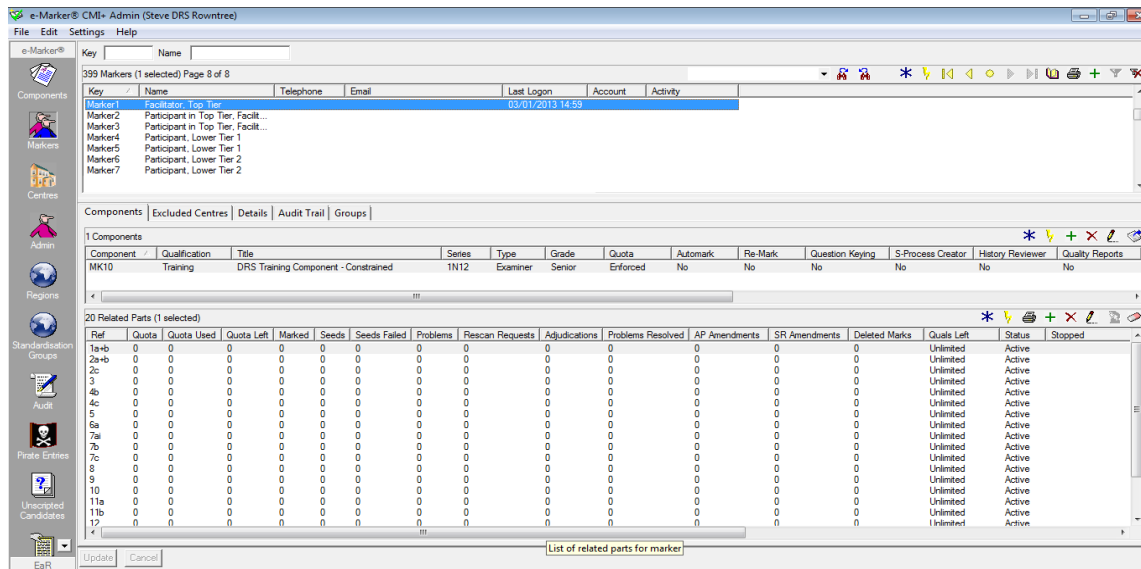


Figure 3.1: The administrator module of the OSM system (adopted from DRS 2013:6)

Newgen (n.d: www.newgen.net) describes the functions of the administrator module thus:

Administrator and Senior Examiners can configure complete examination related activities through this module. Subjects, question papers and marking schemes are managed in a paperless environment. Administrator can configure and initiate practice and live marking sessions. Practice Session/Standardization: Senior Examiner prepares a practice kit for training of examiners using actual sample of examination scripts. Examiners are required to mark the scripts in practice kit and submit it for review to senior examiners. After reviewing the marked scripts, senior examiners certify examiners for live marking.

This statement describes the pre-standardisation and standardisation meetings discussed in Chapter 2, Section 8. At the pre-standardisation meeting the senior examiners select standardisation scripts which are used for training purposes. At the standardisation meeting, the examiners are required to mark the standard scripts which are reviewed by the senior examiners before the examiners are allowed to mark live scripts (Ofqual, 2014e; Baired et al, 2013). The success of the standardisation meetings therefore depends on the competence of administrators and senior examiners in using the administrator module to set marking standards. It is therefore important to study the pre-standardisation and standardisation meetings in the OSM environment in the Zimbabwean context as it is an important quality control activity.

As discussed in Chapter 2 Section 8, the standardisation meetings can be conducted online or face-to-face. Figure 3.1 shows that standardisation groups can be set in the administrator module using the standardisation group icon. Ofqual (2014e:5) insists that the pre-standardisation meeting is always conducted face-to-face probably to maximise the interaction among the examiners. This implies that only the standardisation meeting can be set up in the OSM environment. As discussed in Chapter 2, research shows that the method of standardisation has no impact on the quality of marking; examinations boards can, therefore, choose the face-to-face, online standardisation or a combination of the methods in one examination. Ofqual (2014b:4) established that the examiners had negative attitudes towards online standardisation, preferring the face-to-face method which, they believed, created a shared understanding and encouraged the community of practice culture. Although empirical evidence suggests that neither the shared understanding nor community of practice impacts on the marking reliability (Ofqual, 2014b:30), their negative attitudes may reduce the credibility of the public examinations because they are important stakeholders in education, as established in literature (Dodd, 2014:1).

The Ofqual launched an online survey with teachers and head teachers to gather their views and experiences of marking with the aim of understanding the stakeholders' perceptions on the marking process in the UK. The survey was open to all those who wished to participate, and was conducted via a link on the Ofqual website from April 2013 to June 2013. Therefore, there was no sampling frame. A total of 981 responses were received. The survey was followed up with a call for evidence from 54 education, subject and teaching organisations through email. Six email responses were received. The majority (60%) of the teachers had never been examiners, and the 40% who were examiners acknowledged that the examiner experiences helped them improve their teaching, to an extent that head teachers encouraged their teachers to be examiners. The majority of teachers believed that although examiners were trained to use the marking scheme, they were not monitored during marking (Dodd, 2014:8). The survey also established that only 36% of the teachers and head teachers had confidence in the quality of marking compared to 54% who did not (Dodd, 2014:3). The results indicate that the majority of teachers had no confidence in the quality of marking. It does not augur well for the examination boards when examiners have negative attitudes towards a quality control activity because, as discussed in Chapter 2 Section 7, they are practicing or retired teachers who are important stakeholders in

education. It is therefore important for the examination boards to adopt the standardisation methods preferred by the examiners, lest they lose credibility.

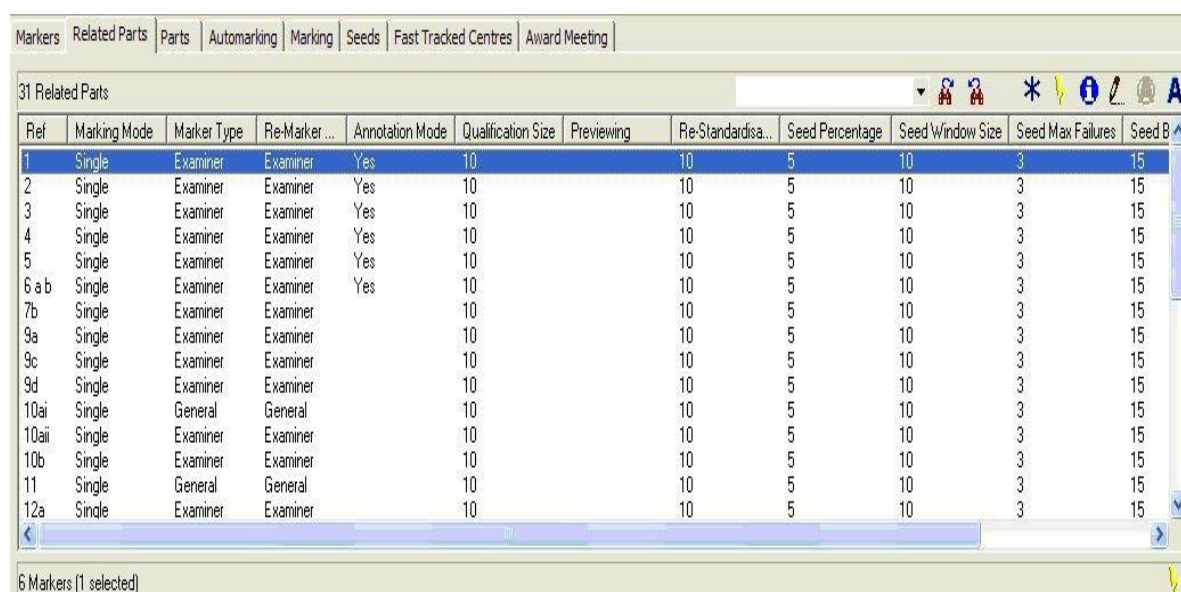
Falvey and Coniam (2010:1) conducted a qualitative study to gauge the responses of the English language raters to the OSM and PBM in Hong Kong. Semi-structured interviews were used to collect data from a total of 17 raters. Twelve of them had marked English on paper and on-screen (experienced raters) while the other five had marked on-screen only (new raters). The research questions were centred around the technical demands of the OSM technology, reading on screen, views on the reliability and efficiency of marking by OSM, training and standardisation, attitudes towards the OSM marking centres, and views on working from home.

In the study, all of the participants indicated that they possessed the right technical skills to work in the OSM environment. Some of the raters said reading on-screen was tiring their eyes and making their necks sore since they had to mark non-stop for three hours so as to utilise their booked space. Falvey and Coniam (2010:10) anticipated that the raters would get used to reading on-screen with time, and the complaints would be no more. The raters pointed out that some candidates used the correction fluid during the examination, resulting in their script images being blurred and illegible after scanning. The researchers state that illegible scripts were a challenge that could be overcome by powerful scanners. The raters had conflicting views on issues of reliability and efficiency when using the OSM. Some felt that they could mark faster and accurately on paper than on-screen while others felt they were faster and more accurate onscreen. Some new raters, who had marked the English language on-screen only said that they would be efficient marking onscreen while others said they would be efficient on-screen and on paper (Falvey & Coniam, 2010:12). Responses from these raters might not be valid given that they had no experience with PBM. The conclusion that more new raters than experienced raters felt that they were more efficient with the OSM might, therefore, not be valid as well. The new raters had no basis of comparison since they had marked in the digital mode only.

On training and standardisation, some examiners felt that they received satisfactory training while others felt that the training was rushed, with seniors dictating marks and comments to the examiners. They were concerned that a senior examiner, undoubtedly under pressure during the

training session, informed raters of the marks that had been awarded to scripts used in training instead of making them work through the scripts. There were also concerns that qualifying scripts were too few (only four), with some examiners arguing that they went through training without mastering the marking standards (Falvey & Coniam, 2010:14). Such practices can compromise the quality of marking and, therefore, need further interrogation in the context of Zimbabwe, hence the purpose of this research. It was therefore important to study the standardisation meetings for O level Biology examinations which marked on-screen. There are, however, conflicting reports about the impact of standardisation on the quality of marking.

The OSM technology allows the setting of quality control parameters in the administrator module as illustrated in Figure 3.2. The literature reviewed in Chapter 2 Section 9 indicated that there are three mechanisms of script moderation in the OSM environment, i.e. double marking, seeding and back reading. Figure 3.2 shows that the seeding mechanism was used to monitor the quality of marking at a threshold of 5% (15 seeds). The seed window had 10 scripts and the examiners were not allowed to fail more than 3 seeds (DRS, 2013:37). It was therefore important to study the setting of seeds in the marking of O Level Biology in Zimbabwe.



Ref	Marking Mode	Marker Type	Re-Marker ...	Annotation Mode	Qualification Size	Previewing	Re-Standardisa...	Seed Percentage	Seed Window Size	Seed Max Failures	Seed B
1	Single	Examiner	Examiner	Yes	10		10	5	10	3	15
2	Single	Examiner	Examiner	Yes	10		10	5	10	3	15
3	Single	Examiner	Examiner	Yes	10		10	5	10	3	15
4	Single	Examiner	Examiner	Yes	10		10	5	10	3	15
5	Single	Examiner	Examiner	Yes	10		10	5	10	3	15
6 a b	Single	Examiner	Examiner	Yes	10		10	5	10	3	15
7b	Single	Examiner	Examiner		10		10	5	10	3	15
9a	Single	Examiner	Examiner		10		10	5	10	3	15
9c	Single	Examiner	Examiner		10		10	5	10	3	15
9d	Single	Examiner	Examiner		10		10	5	10	3	15
10ai	Single	General	General		10		10	5	10	3	15
10aii	Single	Examiner	Examiner		10		10	5	10	3	15
10b	Single	Examiner	Examiner		10		10	5	10	3	15
11	Single	General	General		10		10	5	10	3	15
12a	Single	Examiner	Examiner		10		10	5	10	3	15

Figure 3.2: Component settings in the administrator module (adopted from DRS, 2013:37)

The OSM technology allows the examiners to mark the whole scripts or script portions as illustrated in Figure 3.2, where responses to individual questions were distributed to the markers who were identified as examiners and general. The literature reviewed in Chapter 2 Section 7 indicated that the OSM offers the flexibility of splitting the candidates' scripts by question (item level marking) and distributing them to the specific types of examiners depending on the type of the question (Ofqual, 2014a; Suto & Nádas, 2008). The advantages of item level marking were discussed in Chapter 1 Section 2.4. The Ofqual (2014a:2), however, posits that whole scripts can also be marked on-screen by single examiners, citing examples of UK examination boards which mark the whole scripts for all components marked on-screen. Another examination board allowed the examiners to choose between whole scripts or item level marking (Ofqual, 2014a:4). There is a need to explore the considerations for marking scripts at item or whole script level.

Ofqual (2014g: 20) posits that most of the examination boards which use the OSM technology mark scripts at item level, further asserting that the OSM is used for shorter and more constrained questions. Ofqual (2014g: 20) also states that more objective subjects such as science and mathematics are more likely to be marked on-screen than the subjective subjects such as English, drama and history. The more subjective questions which elicit longer answers seem to present challenges to the implementation of the OSM (Ofqual 2014g; 2011). The details of the challenges are discussed later in this chapter. The answers to the constrained and more objective questions are written in the spaces provided on the question paper, whereas the candidates are provided with the separate answer booklets to answer unconstrained and subjective questions (DRS, 2015; 2013; Hudson, 2009). The challenges associated with the unconstrained examinations might tempt examination boards to design test instruments to suit the OSM technology, compromising the validity of the examinations (Fowles, 2011:2). As discussed in Chapter 1 Section 2.4, the ZIMSEC made a decision to change the formats for some examinations, including Biology (5008), by providing the answer spaces on question papers for the examinations which were previously answered on separate answer sheets (Examination Circular Number 10 of 2015; Examination Circular Number 8 of 2015). It was therefore important to study the nature of the question papers and mark schemes in relation to the assessment schemes prescribed in the syllabi to determine any variations that were probably

caused by the need to mark scripts on-screen. The scripts are distributed to the examiners who access and mark them in the marker module.

3.4.2 The marker module

According to Ramakrishna et al (2012:17), the evaluator/marker module allows examiners to mark the candidates' scripts digitally and send their marks to the administrator databases. The module also has a log in function that requests a username and a password before the marking screen is displayed. The marker screen has the icons which display the scripts, the marking guide, specified fields where marks can be entered and maximum marks for the particular question (Ramakrishna et al, 2012:17). The marker module also has the icons which allow the examiners to enlarge the scripts, annotate the scripts as they mark, make comments and escalate the problem scripts to the senior examiners (Ramakrishna et al, 2012; Johnson et al, 2012b; Ofqual, 2011). The examiners can log onto the OSM system from home or in a designated marking centre (DRS, 2015: www.drs.co.uk). Figure 3.3 illustrates the marker module where the item is worth eight marks and the candidate has been awarded four.

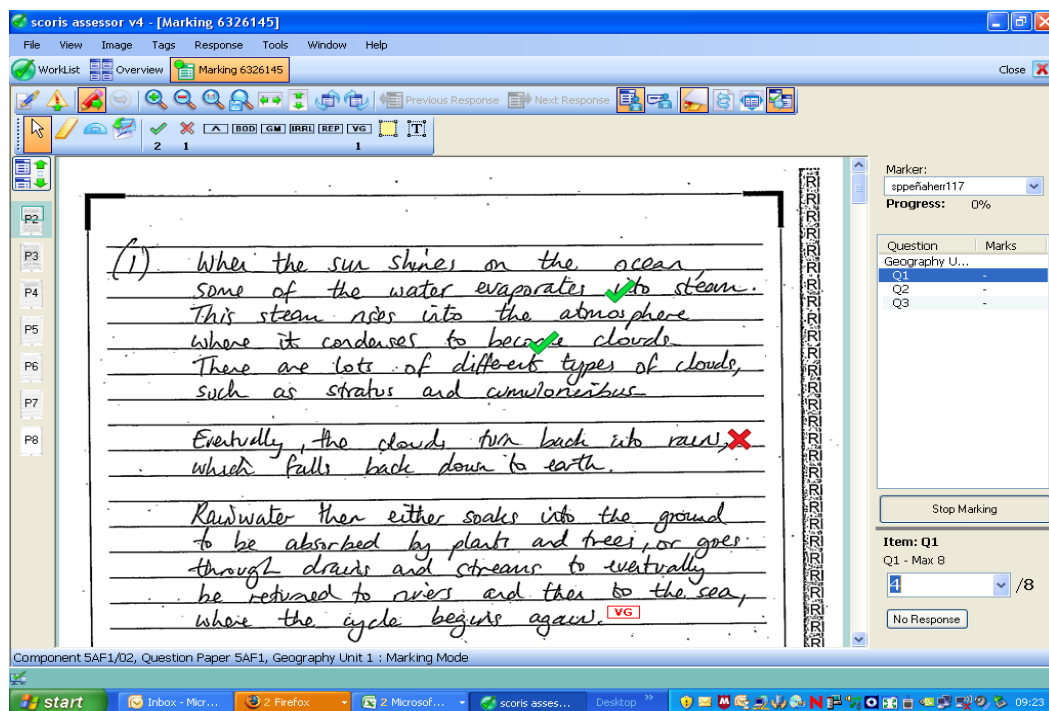


Figure 3.3: The marker screen in the OSM system (adopted from Adams 2011:4)

Many research studies in which examiners shared their experience working in the examiner module were conducted, and these receive attention in this chapter. As mentioned earlier in this chapter, the OSM technology has been widely adopted by the examination authorities around the world more than CBT. The next section discusses the use of the OSM technology by the examination authorities in other countries.

3.5 Global use of OSM

3.5.1 OSM in the UK

In the United UK the OSM technology was first used by Pearson Edexcel in 2003 and by 2013 two thirds of examinations were marked onscreen (DRS, 2014; Ofqual, 2013; Haggie, 2008). According to Haggie (2008:2), RM Education conducted a survey of 17 UK organisations that were planning to adopt the OSM technology. The organisations cited various reasons for adopting the technology and these are listed in Table 3.1.

Table 3.1: Reasons why UK organisations planned to adopt OSM (adapted from Haggie 2008:2)

Reason	Percentage
Quality	87
Cost efficiency	81
Risk reduction	69
More markers	25
Management control	43
Corporate policy	25
Speed of results	6
New markers	18

Quality is the most important reason why the examination boards adopted OSM. Some of the UK examination authorities that adopted the OSM technology are Oxford, Cambridge and RSA (OCR), Pearson Edexcel, Assessment and Qualifications Alliance (AQA), Welsh Joint Examinations Committee (WJEC), the International Baccalaureate Organisation (IBO),

Cambridge International Examinations (CIE) and the Council for the Curriculum, Examinations and Assessment (CCEA) (Ofqual, 2013; Fowles, 2011). Ofqual (2013:12), however, notes that the examination authorities in the UK use the OSM to some varying degree, with Pearson Edexcel marking 88% of their examinations on-screen, and WJEC marking only 13% on-screen by 2013. In the UK, examiners mark from their homes (Raikes et al 2004:20). Several studies have been conducted on the OSM in the context of the UK.

When the AQA planned to adopt the OSM for the 2005 summer examinations they considered a paper which focused on the initiatives in the area of the OSM, operational benefits, and an overview view of the costs involved with a comprehensive cost benefit analysis conducted by AQA's finance department and the software provider (Fowles, 2011:1). The paper also discussed the impact of the OSM on the characteristics of the assessments made, leading to the AQA conducting a literature review on the implications of the technology on the validity of the assessments, lest the technology undermined the quality of the assessments and hence, public confidence in the assessments conducted by AQA (Fowles , 2011:2).

As discussed in Chapter 2, Section 2.10, Fowles (2011:2) emphasises that a valid assessment must have reliability and a reliable assessment must be valid, warning of the danger of redesigning the assessments to suit the logistical efficiency and cost reductions. Fowles (2011:2) enumerated a number of questions related to validity and reliability which might be relevant to the migration of marking from paper to onscreen. The following questions were raised on validity:

- Is there any evidence of assessment schemes being developed and shaped by what technology can offer, at the possible expense of the validity of the assessment?
- Is the validity of the assessment retained when it is carried out, at question or part question rather than whole paper level?
- Is the validity of the assessment retained when it is e-marked by experts and generalists, or by the computer, depending on the complexity of the marking?

The following questions were raised about reliability:

- Do candidates' total marks differ when e-marking is of a whole component onscreen rather than on paper?
- Are candidates' marks affected by marking being carried out at question or part question level rather than at whole paper level (including such factors as examiner accuracy, halo effects, over-penalising of repeated errors). If so, which is the more reliable? Are any differences in reliability the same for components made up of short answer as opposed to longer, free response responses?
- How important to the reliability of marking is the facility to annotate scripts when marking conventionally? How satisfactory are computer-based annotations and comments?
- Can examiners revisit questions/papers they have already marked, whether to amend the mark given or simply to view?
- What are the implications for the reliability of assessments overall if examiner satisfaction is reduced when expert examiners mark electronically rather than conventionally?
- Are there any implications for reliability of marking if the pool of examiners for a particular component/subject loses examiners who are not technologically well equipped or willing to e-mark? Are they likely to be randomly distributed or unrepresentative in terms of their reliability of marking?
- Will there be problems in maintaining a pool of expert examiners for each component if the number of expertly marked responses varies significantly from series to series, and if so, will this affect reliability?

The AQA literature review established crucial findings that were important to this study on the practice of quality control in the OSM environment. The findings of the literature review are summarised as follows:

- Some research studies established that there were no differences in average scores awarded on paper and on screen for objective type of questions. Subjective answers that require examiner judgement are awarded higher marks on paper than on-screen.

- Examiners clearly expressed their desire to mark from home rather than from central venues. Most of the practical issues raised by examiners in pilot studies had been resolved by software designers.
 - A marking rate analysis indicated that OSM is 15% faster than PBM marking.
 - Senior examiners in a pilot study had pointed out their difficulty with blind marking of examiners' script, arguing that they needed to see the mark awarded by the examiner before they can make a judgement of the marker's competence.
- (Fowles, 2011: 4-7).

These findings are important to this study in several ways. Marks awarded to O Level Biology components marked on paper are comparable to components marked by the OSM; the difference in scores awarded to subjective questions in the OSM and on paper assisted in answering the research question on the influence of the question papers and mark schemes on the quality of marking. The results indicate that UK examiners are no longer concerned with design issues, but examiners in Zimbabwe might still be concerned with such issues, probably with an impact on the quality of marking. The OSM is faster than PBM, unless there are logistical challenges, which might impact on the quality of marking. Fowles (2011:5), however, noted that the review did not find the recent literature that answered the research questions raised by Raikes et al (2004:11-15) about the impact of the OSM on the quality of assessments marked on-screen.

The UCLES launched intensive research into the implementation of OSM in 2004. The senior markers participated in an exploratory OSM programme with the UCLES aiming to involve the examiners at the onset so that they could shape the modifications of the software (Raikes et al, 2004:3). The examiners raised a concern that the method of monitoring the quality of marking in the OSM environment smacked off 'Big Brother' and was deceitful. They felt that the seeds were looking for deviations from the correct marks. They preferred marking whole scripts rather than items. The researchers raised a lot of questions on the candidates and centres, examiners, marker training and standardisation, question papers and mark schemes, marking, quality assurance, and awarding (Raikes et al, 2004: 11-15). The AQA literature review, according to Fowles (2011:5), did not find literature that answered these questions. However, some research studies published

earlier and later than the AQA literature review answered some of the questions raised by Raikes and his colleagues, way back in 2004.

Johnson et al (2012a:814) conducted a study to explore whether the mode in which a set of extended essay texts were accessed and read systematically influenced the assessment judgements made about them. An essay question from a General Certificate of Secondary Education (GCSE) examination in English literature was used. Twelve experienced English literature examiners marked two matched samples of 90 essay examination scripts on-screen and on paper. The evidence from the statistical analyses suggested that mode presented no systematic influence on the marker reliability. The markers who tended to be more lenient on paper also tended to be more lenient on-screen and vice versa. The data suggest that within-marker variability levels were lower than the between-marker variability levels (Johnson et al, 2012a:824). The researchers noted that between-marker variability is not a surprising finding given the subjective nature of the subject, emphasising that the examination authorities continue to reinforce procedures such as standardisation exercises (Johnson et al, 2012a:825). The between-markers variability is, therefore, inherent in the marking of lengthy answers but might be compounded by the difficulty to read long texts on screen. There is need to interrogate the type of responses that can be marked on-screen without compromising the quality of marking.

Johnson, et al (2012b:55) conducted another research which aimed to study the marking processes that the examiners used to support their comprehension building whilst marking extended essays on-screen and on paper. Two research questions were formulated:

- How is examiner extended-essay navigation influenced by marking mode?
- How is examiner annotation practice when marking extended essays influenced by marking mode?

The researchers replicated the methodology of their earlier study (Johnson et al 2012a:814), replacing English Literature with Advanced Level American History essays with an average length of 900 words. Data were gathered from 12 experienced examiners who had marked the A Level General Certificate of Education (GCE) examinations in June 2009. In the study, each

examiner marked 90 scripts on paper and another 90 scripts on-screen. The examiners attended a two-day training meeting before marking. One day was spent on use of the OSM software and the other day was spent on standardisation of marking. Data on the examiners' essay navigation were gathered through direct observation of four randomly selected examiners, one from each marking group, as they marked for approximately one hour in each marking mode. Thirty matched essays from each essay sample were selected for annotation analysis, followed up by semi-structured interviews (Johnson et al, 2012:56).

The results indicated that examiners revisited previously marked essays more often on paper than on screen, arguing that it was more difficult to revisit the scripts on screen than on paper; the examiners navigated backwards more on paper than on screen. The researchers concluded that examiners read essays on screen in a linear fashion but in an iterative manner on paper (Johnson et al, 2012:58). The results also indicated that examiners used more varied annotations on paper than on screen with examiners using as many as 35 annotations per script on paper, compared to only six on screen. The researchers noted that annotation behaviours on screen were influenced by the restrictions on the software and the effort required in annotating the scripts. The researchers concluded that marking mode influenced navigation and annotation behaviours of examiners, although a few annotating behaviours were not influenced by the mode (Johnson et al, 2012:62). The navigation and annotation limitations on the OSM technology might contribute to the negative attitudes exhibited by examiners towards the technology when answers become extended as in Biology. It was therefore important to study the influence of examination questions and mark schemes on the practice of quality control in the OSM of O level Biology in Zimbabwe. The OSM technology was also adopted in Hong Kong.

3.5.2 OSM in Hong Kong

The HKEAA first marked all the English Language and Chinese Language scripts onscreen in 2007. A total of 14 subjects had their examinations marked onscreen in 2009 with the intention to totally phase out the PBM and replace it by the OSM for all examinations in 2012 (Coniam, 2011a; Coniam & Yeung, 2010). Literature does not clearly indicate if the HKEAA vision to replace PBM with OSM for all examinations had been realised. However, the statement by Coniam and Yan (2016:1151), that "...onscreen marking has been used for the majority of Hong

Kong examinations since 2012....” implies that there are a few exceptional subjects that are still marked on paper in Hong Kong. The Hong Kong government sponsored the development of IT infrastructure by giving the HKEAA US\$25 million in 2005, resulting in the establishment of three OSM centres that were ready for use in 2012 (Coniam, 2011a:1044). According to Coniam and Yan (2016:1154), the number of assessment centres increased later on, because these authors claim that there are ten marking centres where markers can access examination scripts via intranet and mark them on computer screens.

Coniam (2010:71) conducted a follow-up study involving 30 examiners who had rated the English language essays. Sixteen of them were negative about marking on-screen compared to marking on paper, whilst eight were generally positive about the OSM. The follow-up study was a direct response to the concerns that the attitudes of the two groups of raters (i.e., negative versus positive attitude) might be reflected in the scores awarded to the test-takers through the two marking mediums. It sought to investigate whether the attitudes of the 24 markers affected the OSM marks they awarded. Two hypotheses were formulated:

1. Raters who hold a negative attitude towards OSM will rate test-takers more harshly than will raters who have a positive attitude.
2. Markers who hold a negative attitude toward OSM will be more erratic than more positively oriented markers.

The results were analysed by correlations between the rater attitude and the different component of the HKCEE Writing paper, and multi-faceted Rasch measurement to examine rater fit and erratic behaviour in marking. The results indicated that a negative attitude toward the OSM does not appear to impact upon the reliability of the rating (Coniam, 2010:71).

The finding that negative attitudes do not seem to impact on the marking reliability comes as a relief to the examination authorities that would want to adopt the OSM. This builds on the confidence derived from the Ofqual research discussed in Chapter 2 Section 2.5, which established that the examiners preferred the face-to-face to the online standardisation, but the standardisation approach had no influence on the marking reliability (Ofqual, 2014b:4). The

negative attitudes exhibited by the examiners in these studies could be a result of resistance to change. However, as mentioned earlier in this chapter, the negative attitudes of the examiners do not augur well for the credibility of public examinations. Other researches explored the examiner perceptions on the different aspects of the OSM technology.

Coniam and Yeung (2010: 249) investigated the examiner perceptions on the OSM of Liberal Studies in Hong Kong. Data were collected from the 40 examiners who had marked Liberal Studies in 2009, who completed a pre and post marking questionnaire. Questions were posed on a 6-point Likert scale where 1 indicated a positive response and 6 a negative response. The markers were also asked to provide written comments on any aspect of the OSM process. They formulated two hypotheses:

- Markers will judge themselves to be sufficiently competent technologically to function effectively with in the new OSM medium.
- Markers will not be negative about the OSM medium and will show no preference for either marking medium.

Chi-square tests were used to analyse the results. The markers generally rated themselves as competent and therefore responded positively. The new markers were generally more positive than the experienced ones (Coniam & Yeung, 2010: 262). All markers, however, reported no problems with using computers either technologically or ergonomically. The first hypothesis was therefore accepted. The markers attitudes towards OSM were positive. The researchers noted that the markers were more positive in their post-marking questionnaires than they were in the pre-marking questionnaires. The second hypothesis was also accepted. The results, however, showed that examiners were not very positive about the type of support and feedback provided on their marking accuracy.

The results of this research were compared to the findings of a similar one conducted by Coniam in 2009 where data were collected from the English markers (Coniam and Yeung, 2010:262). It emerged that the Liberal Studies markers were more positive than the English Language markers. It was anticipated, therefore, that with time the OSM would be accepted as the marking

norm. This was especially important with regard to Liberal Studies, given that the marking panel for the subject would increase from around 50 markers to 500 in 2012. Technological efficiency and positive attitudes towards OSM are the important aspects to consider before adopting the technology. Although negative attitudes do not seem to impact on reliability (Coniam, 2010:71), examiners who hold such attitudes might deliberately retard the marking pace and increase the marking costs. It might be prudent for the examination authorities to manage negative attitudes towards the OSM technology. Support and feedback provided to the examiners about the quality of their marking are important aspects that need interrogation, hence the purpose of this study.

The attitudes of Liberal Studies examiners were further interrogated by Coniam (2011a:1042) when he conducted a qualitative examination of the attitudes of Liberal Studies markers towards the OSM in Hong Kong. Fourteen examiners composed of six new and eight experienced markers participated in the study. Data were collected through semi-structured interviews on the four key aspects of the OSM process, which are computer hardware and software; marking centres (environment, location and booking); marker training, support and standardisation; and marking-related issues.

The results indicated that markers generally had no problems with computer peripherals. Comments were made about the hardware and the special software designed for the OSM purposes in Hong Kong. Some markers commented positively on computer hardware stating that the screen resolution was acceptable, the monitor was at the right height, and the workstation chairs were comfortable. A few were unhappy with the ergonomics, arguing that the keyboard and mouse were a bit too high; mouse-click responses were too slow; and the user interface was not user friendly due to frequent pop-ups that were disruptive (Coniam, 2011a:1042). A problem was raised with annotating scripts with ticks or crosses. While the scripts can be annotated on paper, markers said it was much harder to do that in the OSM because there were not enough symbols to use, the same results established by Johnson et al (2012b:62) in the UK. The markers were also not happy about inaccessible internet services. For security purposes, the dedicated workstations did not have internet access. The markers argued that they needed to check that the content provided by the candidates was accurate, not the result of plagiarism and how this affected judgment in terms of the marks awarded (Coniam, 2011a:1045).

The issues that were raised about hardware and software could have implications on the quality of marking. The examiners would most likely be accurate where there are no software and hardware challenges. A slow system might frustrate the markers, leading to the sloppy marking. There might have been a genuine need to check the quality of answers on the internet, but some examiners could get distracted and retard their marking pace. The examination authorities might need to strike a balance between checking answers and marking progress.

The markers raised three major issues regarding marking centres, namely, centre layout, accessibility to the centres and the convenience of booking marking sessions. The markers generally favoured the design and layout of the three marking centres (one in each of Hong Kong's three major geographical areas). Some markers commented favourably on the general environment and the lounge. Others, however, mentioned the noise being an issue, particularly talking on mobiles. The location was a major issue, with some markers complaining that they had to travel a long way as there were only three marking centres. Some issues were raised about the booking of three-hour marking sessions. Some markers were happy with booking arrangements but others felt seriously inconvenienced as, in addition to travel, marking was in three-hour stretches and they could only book two three-hour sessions daily (Coniam, 2010:1046).

These results imply that the marking logistics need to be planned well. Coniam (2011a:1046) points out that the markers in Hong Kong are full-time teachers. It is not clear from the research if marking was done during school holidays or during the school term. If they had to mark during the school term then they would mark for a few hours per day due to the booking system that puts a constraint. Travelling to the marking centres might also tire them out, leading to a sloppy marking. It would be prudent to put up an OSM system that the markers can easily access at the convenience of all parties, the examiners and the examination authority.

The markers were generally satisfied with the amount of marker training they received with seven commenting positively on the use of standardising scripts (seeds), presented onscreen to the markers at certain intervals to check their marking consistency (Coniam, 2011a:1049). Some

markers said that the seeds kept them on track during marking and ensured fairness to the candidates. Others felt that the seeds were too few, with two presented with 40 scripts. This translates to 5% of the scripts being used as seeds. The same percentage is shown in Figure 3.2. This raises some questions: What happens if the seeds are set below or above 5%? How is the percentage arrived at? Who determines the percentage of scripts to be used as seeds? It was therefore important to interrogate the setting of quality control parameters in the marking of the O Level Biology in Zimbabwe.

Some markers enumerated several advantages of the OSM. They commented that it secured the candidates' scripts, preventing them from being lost; maintained candidate confidentiality; enabled item level marking which they perceived as more reliable than whole script marking; provided useful mark distribution statistics (Coniam, 2011a:1050). One examiner, however, pointed out that they were not adequately trained to use the mark distribution statistics during marking and, therefore, chose to ignore the statistics (Coniam, 2011a:1050). The Liberal Studies examiners in an earlier study had indicated a less positive attitude about the support and feedback provided to them (Coniam & Yeung, 2010:260). The English language examiners had also cited rushed training and too few qualifying scripts (Falvey & Coniam, 2010:14). These examiners' concerns imply that quality control in the OSM environment is an issue that needs to be interrogated. Issues for further investigation include the parameters for determining the frequency at which the seeds are presented to examiners; the consequences of using too few seeds; and support given to deviating examiners once they are identified by the seed mechanism. It was therefore important to explore the practice of quality control in the Zimbabwean context. More researches on the acceptance of the OSM were conducted in Hong Kong.

Yan and Coniam (2014:464-480) investigated the effects of the three key demographic factors, which include the language of marking, gender and age and the markers' reactions to the OSM. A total of 1743 markers completed a post-marking questionnaire consisting of two previously validated scales, i.e. ease of use and acceptance of the OSM scales. The results showed that the markers generally reported finding the system easy to use and positive acceptance of the OSM. The markers marking in both English and Chinese had higher perceived ease of use and acceptance than markers who marked only in English or in Chinese. Gender also had a

significant impact on the markers' responses to the two scales – favouring males. Age was not a significant factor influencing the markers' perceived ease of use, but the older markers revealed a significantly higher level of acceptance than younger markers.

The results of this research indicate that the OSM was becoming more acceptable to examiners in Hong Kong, as predicted by Coniam and Yeung (2010:262). The results also imply that the examination authorities do not need to worry about the age of examiners and the marking language when adopting the OSM technology. The impact of gender might need further interrogation in different contexts.

A similar study was conducted by Coniam and Yan (2016:1151), to compare the markers' reactions to the OSM, i.e. perceived ease of use and acceptance of the OSM against the backdrop of virtually all subject areas being marked on screen in Hong Kong. The study was meant to answer two research questions:

- What are the effects of subject area on markers' perceived ease of use in the OSM environment?
- What are the effects of subject area on markers' perceived acceptance of OSM?

The data were collected from a survey of 1743 markers who were all classroom teachers across 14 major subject areas. The markers' qualitative comments about the OSM system were collected and post-hoc interviews with a key informant from the Hong Kong Examinations and Assessment Authority (HKEAA) were conducted. The analysis of the survey data was triangulated by the markers' qualitative comments together with the HKEAA staff interview. The results showed that the markers generally revealed a high level of perceived ease of use and positive acceptance of the OSM. The effect of subject area for both scales was statistically significant. On the Ease of Use in the OSM Environment scale, the markers of ICT, Mathematics and Physics were the most positive, with the markers of History, Geography and Biology being the least positive (Coniam & Yan, 2016:1160). The markers of ICT, Mathematics and Physics listed more advantages of the OSM than disadvantages, while the markers of History, Geography and Biology listed more disadvantages than advantages of the OSM. The listed advantages

include improved efficiency of marking; no sorting/flipping of papers, time is therefore saved; marking is faster. The disadvantages that were listed include eye strain; and the hustle of travelling to marking centres which is time consuming and tiring (Coniam & Yan, 2016:1162). The concerns about travelling to the marking centres were also raised by Liberal Studies examiners in an earlier research study in Hong Kong (Coniam, 2010:1046). The HKEAA might have to listen to examiner concerns and revise their approach to examiner access to the system.

The markers for subject areas that were marked by section, such as ICT, Mathematics and Physics demonstrated higher levels of perceived ease of use and acceptance of the OSM than the markers for subject areas that were marked by single question, such as History, Geography and Biology. The ways in which the marker panels are formed are closely related to the question type predominating in the examination paper. The subject areas involving extended response questions were normally marked by question, while the subject areas involving limited response questions were marked by section (Coniam & Yan, 2016:1163). Coniam and Yan (2016:1163) elaborated thus:

In general, subject areas involving extended response (i.e., paragraph or essay) questions were marked by question, while subject areas involving structured response (writing one or two lines) and limited response (single word, multiple choice) questions were marked by section. Interestingly, this information regarding how marking panels were divided up reflected patterns of acceptance.

The researchers noted that the marking of lengthy answers on-screen might cause eye strain, whereas the marking of shorter answers might give a better sense of progress to the examiners. The researchers concluded that the subject area was found to have a significant impact on perceived ease of use and acceptance of the OSM; further explaining that the teachers of ICT have a higher computer self-efficacy and accept emerging educational technologies with greater ease than teachers of other subject areas (Yan & Coniam, 2016:1164). Coniam's and Yan's (2016:1151) statement that their research was conducted at the backdrop of OSM of virtually all subjects imply that a few subjects are still marked on paper. This could mean that there are some subjects that cannot be marked on screen.

The findings of this research relate to the studies on the impact of examination questions and marking schemes on accuracy of marking. As discussed in Chapter 2 Section 10, extended responses are marked less accurately than shorter and precise responses (Child et al, 2015; Ahmed & Pollit, 2011; Bramley, 2008). The marking of extended answers might even be less accurate given that reading on-screen is more difficult than reading on paper (Coniam & Yan, 2016:1163). The OSM technology was adopted for public examinations in China and other countries, as discussed in the next section.

3.5.3 OSM in China and other countries

The OSM technology is also widely used in China where the Guangxi Province first piloted the marking of the English university entrance test in 1999, with the senior high school entrance tests (called the gaokao) marked on-screen since 2010 (Coniam & Yan, 2016; Coniam, 2011a). Coniam (2011a:455) posits that it is difficult to gauge the actual numbers of the scripts marked on-screen in China because marking is decentralised to provinces and municipalities. However, he states that the Chinese Ministry of Education indicated that 28 out of 32 provinces and municipalities would be adopting the OSM. Coniam and Yan (2016:1152) state that over ten million scripts were marked on-screen for the gaokao in 2010, adding that the success of the OSM in the gaokao examinations spurred the adoption of the technology for other high-stakes examinations. The examiners in China also mark from dedicated centres as in Hong Kong (Coniam & Yan, 2016:1152).

The OSM was also adopted in the Caribbean Islands. Geisha (2012: www.guardianmedia.org) reported that the Caribbean Examinations Council (CXC) intended to introduce onscreen marking in three to four years' time. The examining authority used the centralised system where the examiners were accommodated at one place throughout the marking period and was reported to be paying between \$18 and \$19 million for the marking exercise (DRS, 2014; Geisha, 2012). Apart from paying a daily stipend, CXC also provided hotel accommodation and meals for markers. A CXC official was quoted as stating, "...while this may be good for teacher development and while teachers may like it, this is just not sustainable...." (Geisha, 2012: www.guardianmedia.org). Another CXC official who was also quoted said that onscreen marking would greatly reduce the marking budget of between \$18 and \$19 million dollars. Roan

(2009:6) concurs with the CXC official and asserts that reducing the cost of running public examinations with onscreen marking is an obvious factor considered by the examination boards. However, the CXC acknowledged that onscreen marking brings its own challenges, and singles out the problem with internet penetration. An official quoted by Geisha (2012: www.guardianmedia.org) elaborated the challenge thus:

We have to ensure that teachers who will be marking electronically will have access to the internet. We have to make sure they are in areas where they have access because some of our territories are not as developed as others in terms of internet accessibility so we have to be very careful.

The ZIMSEC is in the same predicament with the CXC and is likely to face challenges related to undeveloped ICT infrastructure (MoPSE, 2018; National ICT Policy, 2015). Undeveloped ICT infrastructure could cause serious disruption of the OSM, possibly compromising the quality of marking.

The OSM has been reported in Cyprus and Australia (Coniam, 2011a:455) and in a few African countries such as Ethiopia, Nigeria, Tanzania and Zimbabwe (Coniam & Yan, 2016:1152). The Namibian ministry of education used the OSM for Geography, Entrepreneurship, History, Design and Technology and Physical Science examinations at Grades 10 and 12 in 2014 (Namwandi, 2014: www.namibian.com). DRS (2014: www.drs.co.uk) posits that Nigeria piloted the OSM in 2010, using 1300 scripts for Agricultural Science, Biology and Economics, further asserting that the successful implementation of the OSM in Zimbabwe saw the increased adoption of the technology in Africa. The researcher did not come across published research studies from China and these other countries. The following section discusses the implementation of the OSM in Zimbabwe.

3.5.4 OSM in Zimbabwe

There are variations in literature regarding the actual time when the ZIMSEC first implemented the OSM. The Herald (2015: www.herald.co.zw/zimsec-sets-pace-in-e-marking) concurs with DRS (2014: www.drs.co.uk) that the ZIMSEC piloted the OSM in 2010 and marked the first two

subject components, i.e. Integrated Science and Mathematics, in June 2011. Some media houses, however, reported that the ZIMSEC launched onscreen marking of the candidates' scripts at Ordinary Level in the June 2012 examinations which were marked in July 2012, beginning with the two subject components, which are Mathematics Paper 1 and Integrated Science Paper 3. The examination authority held a one-day tour of the OSM venue (Chinhoyi University of Technology), as reported by several media houses, on the 14th and 15th of July 2012, where the director officially announced the commencement of the OSM. The chairman of the Information Communication Technology Committee, Dr. Brooking, was quoted saying that marking had started on Tuesday of the previous week, which was the 3rd of July 2012 (Kachere, 2012; Karombo, 2012). Document review, face-to-face interviews with subject managers for the pilot subjects and personal experience indicate that the OSM was used to mark live examinations in the June 2012 examination series. Although I did not participate in the June 2012 examination, I participated in the pilot in October 2011 at the Bindura University of Science Education and the November 2012 examination which was the second live marking session.

The ZIMSEC adopted the OSM with a view to improve the quality and logistical efficiency as well as to reduce the cost of marking. Statements relating to quality were said by officials who attended the tour of the marking venue. The then director of ZIMSEC, Mr. Esau Nhandara, said that the technology would 'watch over' the performance of markers, always ready to stop them should they deviate outside the parameters set in the programme. He further expected that the technology would ensure accurate marking and enhance the quality of results. Dr. Brooking said that the OSM would improve the quality of results as there was no chance for markers to cheat in the addition of marks at the end of marking, elaborating that the programme was designed to stop functioning if a marker awarded more or less marks than those expected (The Chronicle, 2012; Bulawayo24 News, 2012). The then OSM project manager, Mr. John Maramba, said that deviating examiners would be stopped from marking and would only be allowed to resume when their supervisors were satisfied that they had mastered the contentious sections of the mark scheme. He pointed out that the examiners who continued to deviate would be dropped from the marking teams (Kachere, 2012: www.sundaymail.co.zw). The technology was, however, implemented in a plethora of challenges that could militate against the quality control mechanisms that come with it.

As discussed in Chapter 1, Section 2.4, the OSM technology was introduced at a time when there was limited use of ICT in general and in education in particular; the internet coverage was patchy; power supply was erratic; and the economy was ailing. Mr. John Maramba, the project manager, even appealed to government to help the ZIMSEC acquire a minimum of 1000 computers to enable the expansion of the OSM project (Bulawayo24 News, 2012: www.bulawayo24.com/index-id-news-sc-education-byo). The project expanded and by November 2016, 20 components were marked on screen (Examination Circular Number 10 of 2015; Examination Circular Number 8 of 2015; Examination Circular Number 42 of 2013; Examination Circular Number 41 of 2013). The challenges related to ICT infrastructure may disrupt marking and quality control activities, leading to the examiner frustrations observed in Chapter 1, Section 2. It was, therefore, important to study the implication of these challenges for the practice of quality control in the OSM environment, hence the purpose of this research.

As mentioned earlier, the ZIMSEC adopted the OSM technology to improve logistical efficiency of marking. The Council's officials enumerated several logistical benefits that were expected from the OSM technology, which included:

- Easing the burden of adding the marks manually as it automatically adds and sends the final marks to the database within seconds;
- The OSM being much faster than the traditional belt-marking;
- The turnaround time for marking is significantly reduced because there is no bulk handling of scripts and there is no filling in of mark sheets after the marking session;
- There is no shuffling of papers during marking, eliminating the possibility of candidates' scripts being lost or torn once scanning has been completed; and
- The quick delivery of results – November results expected to be delivered in January.

(The Herald, 2015; DRS, 2014; Bulawayo24 News, 2012; the Chronicle, 2012; Kachere, 2012; Karombo, 2012).

Literature, however, warned that technologies should be used to support the existing assessment practices rather than designing assessments that suit the demands of technologies so as to achieve

logistical efficiency of the assessment process (Ramakrishna et al, 2012; Fowles, 2011; Roan, 2009). As observed in Chapter 1 Section 2, ZIMSEC made a decision to constrain all the examinations marked on-screen, after marking one unconstrained component for two examination sessions (Lessons learnt reports, June 2013; November 2013). Literature suggests that unconstrained examinations pose challenges in the OSM environment.

Ofqual (2011:4) reported that in the summer 2010 examination series, the AQA marked approximately 270,000 scripts across 54 unconstrained components. To facilitate the electronic segmentation of candidates' responses, a new question numbering system and answer book format was introduced for all components that used a separate answer book. The OSM of unconstrained components had been piloted in 2009 and in January 2010. The system failed to ensure that all creditworthy material presented by candidates was marked, resulting in 3353 candidates from 1335 examination centres receiving wrong marks for 48 out of the 54 unconstrained components marked on-screen. The system failure was only discovered after enquiries about the results by examination centres almost a month after the results had been published. The unmarked work could only be viewed on the hard copies of the scripts (Ofqual, 2011:4).

Ofqual (2011:5) enumerates the factors that contributed to the system failure as follows:

- The process for dealing with the variety of ways in which candidates recorded their answers;
- The process for fixing the segmented images of the candidate's responses before they are released to examiners for marking;
- The role and training of examiners in the onscreen marking process;
- The selection of components for onscreen marking of unconstrained answers in separate answer booklets;
- Limitations of the pilot exercises carried out in 2009 and January 2010;
- Inadequate user acceptance testing; and
- The absence of appropriate project and risk management arrangements.

Ofqual (2011:11) describes an elaborate procedure for dealing with unconstrained scripts in the OSM environment. It is apparent that the ZIMSEC decided to provide spaces on the question papers to avoid these procedures. The Examination Circular Number 19 of 2013 was dispatched to examination centres informing them that Commerce Paper 2 (7103/2) was going to be answered on booklets provided by the council, with an exemplar script attached to guide candidates and invigilators on how to fill in candidate details and to number the questions answered. The council faced challenges with this unconstrained component as discussed in Chapter 1 Section 2.4.

As indicated in Chapter 1, Section 2.4, Section B of the O Level Biology 5008/2 was unconstrained when it was marked on paper and was constrained before it was marked on-screen. The decision to constrain all the examinations raises a number of questions. Can all answers to the question fit onto the space provided on the question paper? What criterion is used to determine the magnitude of the space provided for responses to the question? Are there no chances of shortening examination questions so as to provide minimum space and save on costs of paper, therefore, designing examinations for logistical efficiency of the technology and compromising validity? It was therefore important to study how the nature of examination questions and mark schemes influence quality control in the OSM environment.

In addition to improving logistical efficiency, the ZIMSEC also expected to reduce marking costs using the OSM technology. Mr. Esau Nhandara was quoted by DRS (2014: www.drs.co.uk) saying:

Another benefit is reducing costs. Given the significant time savings, ZIMSEC estimates savings of up to \$30,000 a day from labour costs alone. Further cost savings are anticipated as the new system becomes more embedded and improvements in the IT infrastructure are realised. Markers will also eventually be able to mark scripts securely from any location and will not be limited by having to attend a centralised marking facility.

This public pronouncement of estimated savings may force the examination authority to continue using the technology even if it becomes financially unsustainable, compromising the quality of marking. As discussed in Chapter 1, Section 2.4, examinations are marked to meet specific

timelines (Lessons learnt reports, June/November 2017; June 2015; Ofqual, 2013), exerting pressure on both the examiners and their supervisors, to mark and meet the deadline. This desire to save money may force the Council to set short training, standardisation and marking periods, compromising the quality of marking. It was, therefore, imperative to interrogate the standardisation activities, mechanisms that have been put in place to ensure adherence to mark schemes and to enforce the re-training of stopped examiners. The OSM technology presented challenges to the ZIMSEC.

In January 2014, there were media reports which said the ZIMSEC was facing challenges with the OSM technology and there was a likely delay in the release of the November 2013 results. Some official was quoted saying that infrastructure and technical problems were delaying the marking of three major subjects (Dube, 2014; NewsdzeZimbabwe, 2014). The nature of the challenges was not clear from the media reports. As discussed earlier in this chapter and in Chapter 1 Section 2.3, there are several challenges that militate against ICT in Zimbabwe, and these could have contributed to the challenges faced by the ZIMSEC in January 2014. Some of the challenges were associated with quality control in the marking of Commerce Paper 2 as discussed in Chapter 1 Section 2.4. These challenges could compromise the quality of marking, hence the need to interrogate the practice of quality control in the OSM environment. The researcher did not come across published research studies on the implementation of the OSM technology in Zimbabwe. However, numerous research studies were conducted in other countries, especially the UK and Hong Kong, as discussed earlier in this chapter. The contexts of UK and Hong Kong differ from Zimbabwean contexts, hence the need to explore the practice of quality control in the OSM environment.

3.6 Advantages of OSM

The OSM technology was basically designed to overcome challenges associated with PBM. Newgen (n.d: www.newgensoft.com/wp-content) and DRS (2015: www.drs.co.uk) concur that the OSM was introduced to overcome PBM challenges, and these are listed as follows:

- Handling of large numbers of paper scripts during marking.
- Risk of misplacing scripts during distribution.

- Labour intensive process of sorting scripts and allocating them to the right examiners.
- Errors that lead to inconsistent and delayed marking.
- Lack of real-time information about the marking process.
- Recruiting and managing large numbers of examiners.
- Completing marking and publishing results on time.
- Storing and preserving physical scripts to comply with retention policies, leading to space constraints.

Examination authorities have introduced technical systems and services to support the assessment processes, hoping to overcome the enumerated PBM challenges and derive the purported benefits of the technology. Several advantages of the OSM technology were enumerated in literature, which include:

- enhanced process control;
- improved access to wider examiner expertise;
- enhanced communication and support across the evaluator teams;
- richer data on examiner, candidate, item, component and paper performance;
- reduced time to result;
- raised marking quality/consistency; and
- reduced administrative error

(DRS, 2015; Ramakrishna et al, 2012; Roan, 2009).

Some of the benefits were confirmed in the research studies discussed earlier in this chapter. The confirmed advantages of the OSM include improved efficiency of marking; no sorting/flipping of papers, time is therefore saved; marking is faster (Coniam & Yan, 2016; Fowles, 2011). Ramakrishna et al (2012:15), however, warns that “...without close matching of the technical system and the assessment process, and without careful engagement of appropriate stakeholders, benefits remain theoretical...”, and calls for intensive research to validate the transition from paper to the OSM, arguing thus:

Although technology offers potential to broaden educational assessment beyond what traditional methods allow, there are inevitable concerns during a transition phase (where assessments exist in both paper- and computer-based modes) that their outcomes are not comparable....Comparability studies explore the possibility of differential effects due to the use of computer-based evaluation instead of paper-based evaluation. These studies help ensure that test score interpretations remain valid and that students are not disadvantaged in any way by taking a computerized evaluation instead of the typical paper evaluation.

The ZIMSEC is still in the transition phase with only 20 subject components being marked on-screen by November 2017 (Examination Circular Number 10 of 2015; Examination Circular Number 8 of 2015; Examination Circular Number 42 of 2013; Examination Circular Number 41 of 2013). As noted in Chapter 1 Section 2.1, subjects in the old curriculum were last examined in the June 2018 examination series (Examination Circular Number 2 of 2018), hence the need to study the practice of quality control so as to inform the OSM related decisions in the new curriculum. The OSM technology, however, is not without challenges. The next section discusses the challenges of the OSM technology.

3.7 Challenges of OSM

Pinot de Moira (2013:3) argues that the advantages of moving away from PBM depend on the efficiency of the OSM system. Inefficient OSM systems can therefore pose challenges to the marking of examinations. Ofqual (2013:12) posits that the OSM systems can be sources of new clerical errors, insisting that there have been a few cases where one examination board in the UK reported incorrect addition of marks and incidences where some pages of an answer booklet cannot be scanned or marked by the examiners. As discussed earlier in this chapter, Ofqual (2011:4) reported a failure of the OSM system used by the AQA to mark unconstrained answers. This resulted in 622 incorrect grades being awarded to the candidates. The problem only came to light when the examination authority was inundated with inquiries about the issued results. The magnitude of the challenge faced by the AQA could compromise the credibility of public examinations and, therefore, need to be mitigated.

Some challenges were highlighted in the research studies discussed in this chapter and these are eye strain; inability to revisit scripts; limitations in navigation and annotations (Coniam & Yan,

2016; Johnson et al, 2012a; 2012b). Some of the challenges reported in some of the research studies in Hong Kong are purely administrative and are context specific. The challenges associated with the OSM technology could impact on the quality of marking, hence the need to interrogate the practice of quality control in the OSM environment in the Zimbabwean context.

3.8 Conclusion

This chapter reviewed scholarly literature on the development and adoption of e-assessment in public examinations so as to place the OSM technology into the context of this study. The chapter then explored the nature of the OSM and how it works, its use by examination boards in other countries and in Zimbabwe specifically. The advantages and challenges of the OSM technology were discussed. E-assessment refers to any assessment activity that is designed, accessed and stored through the medium of information communication technology and came into being due to some of the reasons enumerated in literature. The numbers of students to be tested are growing, making paper-based assessments unsustainable; the need to assess valuable life skills that require the use of computers; and the need to bridge the gap between classroom (where there is wide use of technology by both teachers and learners) and assessment practices.

E-assessment dates back to the 1950s, starting with optical scanners for multiple choice examinations and developed to the varieties that exist today. The adoption of e-assessment was predicted by Randy Bennett of ETS, as progressing through a three generation model. Literature suggests that most assessment technologies are still stuck in the first generation type, where existing assessment processes are automated. The adoption of Bennett's generations assessments face challenges associated with high-stakes examinations. The technologies have, therefore, been limited to the assessment of lower cognitive skills. This study focused on the practice of quality control in the OSM environment which automates the existing marking process, but not in the manner predicted by Bennett. The chapter went on to give an overview of the OSM technology.

The marking process in the OSM environment can be summarised thus: Examinations are completed by candidates; examination scripts collected from examination centres; examination scripts scanned and images saved; images of examination scripts distributed to markers for

marking via a secure intranet or internet system; and marks and annotations by markers captured by the OSM system. The software consists of two main modules, i.e. the administrator module where all quality control parameters are set; and the marker module where script images are accessed and marked by examiners. The OSM technology is used by examination authorities in the UK, Hong Kong, Australia, China and the Caribbean Islands. The technology has also been used by African countries such as Nigeria, Ethiopia, Tanzania, Namibia, with Zimbabwe being the pioneer. The ZIMSEC implemented the OSM technology for the June 2012 examination session, with a view to improve the quality of marking; reduce marking costs; and improve the logistical efficiency of marking.

Several research studies were conducted on various aspects of the OSM technology in different contexts, with the results indicating that scores awarded on screen are comparable to those awarded on paper; whole script marking is as reliable as item level marking; examiners prefer face-to-face standardisation to online standardisation; standardisation methods do not influence the marking accuracy; attitudes of examiners towards the OSM do not affect their marking accuracy; the OSM is faster than PBM; subjective answers that require examiner judgement are awarded higher marks on paper than on screen; some examiners complain that the OSM strains their eyes. Literature also highlighted the advantages of the technology, that include enhanced process control; improved access to wider examiner expertise; enhanced communications and support across the evaluator teams; richer data on examiner, candidate, item, component and paper performance; reduced time to result; improved marking quality and many more. Some challenges of the technology were also highlighted. These include some clerical errors; eye strain; inability to revisit scripts; limitations in navigation and annotations; wrong grades awarded for unconstrained examinations. It was, however, noted that most of the research studies focused on research questions that did not address quality control issues, but discussed them here and there. The studies were conducted in contexts other than Zimbabwe, which is the first country to implement the technology in Africa. It was therefore important to explore the practice of quality control in the marking of O Level Biology examinations in Zimbabwe.

The next chapter discusses the case study methodology which was used to address research questions raised in Chapter 1 Section 4.

Chapter 4

Research Methodology

4.1 Introduction

This chapter articulates the qualitative single instrumental case study methodology that was used in this study. The research study was premised on the constructivist paradigm. An overview of research paradigms was given before the constructivist perspective was discussed so as to justify the choice to use it in this study. I positioned myself and declared my background and experiences that might have influenced my interpretation of the findings and thus accounted for rigour. The history of case studies was traced so as to position the study within the constructivist paradigm. The population, sample and sampling procedures, data collection and analysis techniques were outlined and data collection explicated. Issues of trustworthiness and ethical considerations were addressed.

4.2 Location of the study

This study was located at the ZIMSEC, an institution that is mandated by the Act of Parliament Number 17 of 1994 to conduct primary and secondary school examinations as described in Chapter 1 Section 2.1. The act enumerates several functions for the council, which include to organise and conduct examinations in subjects that form part of a course of primary or secondary education as the minister may in writing direct; consider and approve subjects for examinations; appoint panels or boards of examiners; approve and register examination centres; review rules and regulations relating to examinations; confer or approve the conferment of certificates, diplomas and other awards to persons who have passed examinations; enter into arrangements, whether reciprocal or otherwise, with persons or organisations inside or outside Zimbabwe for the recognition of certificates, diplomas and other awards granted in respect of examinations organised or conducted by the council; do all things necessary to maintain the integrity of the examination system in respect of primary and secondary education in Zimbabwe; and do any

other thing that the council may be required to do by or under this act or any other enactment. The research was premised in the constructivist paradigm as discussed in the next section.

4.3 Constructivist paradigm

The term paradigm was first used by Thomas Kuhn way back in 1962 to mean a philosophical way of thinking (Kivunja & Kuyini, 2017; Shah & Al-Bargi, 2013). Paradigms are basic philosophical systems or world views that guide research. Tracey (2013:38) describes paradigms as preferred ways of understanding reality, building knowledge, and gathering information about the world. Yazan (2015:135) concurs with Tracey (2013:38) and posits that inquiry is conceptualised and operated by the researcher's beliefs about the nature and production of knowledge, emphasising that a paradigm permeates the investigation process, from the selection of the phenomenon of interest that is put under scrutiny to the way the ultimate report is composed. Kivunja and Kuyini (2017:26) posit that a paradigm is the set of abstract beliefs and principles that shape the researcher's view of the world, how the researcher interprets and acts within that world. The authors also emphasise that a paradigm permeates the entire research process. These authors concur that a paradigm is a set principles and beliefs that guide the research process. In this study, a paradigm will therefore be considered as such; a set of principles and beliefs that guide the research process.

There is, however, a need to acknowledge the contradiction regarding the paradigms that generally guide educational research. Some scholars believe that there are four main paradigms in educational research which include positivism, constructivism (also called interpretivism), pragmatism and the critical paradigm (Kivunja & Kuyini, 2017; Mack, 2010). Other scholars point out that educational research is guided by three main paradigms, i.e. positivism, constructivism and pragmatism, which subsequently inform quantitative, qualitative and mixed methods research approaches respectively (Boeren, 2017; Shah & Al-Bargi, 2013; Atieno, 2009). Boeren (2017:65) posits that some scholars discuss paradigms in a more sophisticated way, resulting in names such as post-positivism, feminism, critical theory and many more, extending the positivist and constructivist divide. I also acknowledge the existence of many paradigms in educational research but tend to agree that three main paradigms abound in literature as guiding educational research, and these are positivism, constructivism and pragmatism.

It is therefore important to understand the basic philosophical assumptions that illuminate research paradigms as advised by Tracey (2013:38) and Mack (2010:5), who concur that the researcher's view of the constructs of social reality and knowledge affects how they will uncover knowledge of relationships among phenomena and social behaviour, and how they evaluate their own and others' research. The purpose of this study was to explore the practice of quality control in the marking of Ordinary Level Biology in the onscreen marking environment in Zimbabwe in order to propose a framework which can help to improve the practice. I read and understood the philosophical assumptions of positivism, constructivism and pragmatism in relation to this purpose. I chose the constructivist paradigm because it offers the best approach to accomplish the purpose of this study. The philosophical assumptions of the positivist and the pragmatic paradigms are discussed in relation to the case study methodology later in this chapter.

Paradigmatic assumptions can be divided into four main elements, i.e. ontology, epistemology, methodology and axiology. These receive attention in the following sub-sections.

4.3.1 Ontology of the study

The ontology of a research paradigm constitutes basic assumptions about the nature of reality, which vary from one paradigm to another. Kivunja and Kuyini (2017:27) posit that ontology is the philosophical study of the nature of existence or reality, of being or becoming, and the groups of things that exist and how they relate to each other. The authors state that ontology makes the researcher to ask questions such as: Is there reality in the social world? Is it objective or subjective? What is the nature of the phenomenon being studied? The authors further state that ontological assumptions guide the researcher's thinking about the problem, its significance and the best approach to answer research questions. This implies that the basic assumptions about reality determine the methodology of providing solutions to the problem under investigation.

The constructivist ontology assumes that individuals seek to understand and make sense of the environment in which they live and work and develop subjective meanings of their experiences with objects or things. Reality is therefore a subjective construct that is shared among participants and researchers seek to understand the people's idea of reality (Charmaz, 2017;

Creswell, 2014; 2009). Mack (2010:7) unpacks the ontology of constructivism insisting that reality is indirectly constructed based on individual interpretation and is subjective; people interpret and make their own meaning of events; events are distinctive and cannot be generalised; there are multiple perspectives on one incident; and causation in social sciences is determined by interpreted meaning and symbols. Creswell (2014; 2007) emphasises that the goal of research is to rely as much as possible on the participants' varied and multiple perspectives, looking for the complexity of views rather than reduce meanings to a few categories or ideas.

The ontology of this research, therefore, assumed that examiners and subject managers held and shared multiple realities about quality control in the OSM environment, created during the marking of O Level Biology (5008) examinations. I sought to understand how they interpreted OSM quality control, the factors that influenced their interpretations and how their interpretations of quality control varied with experiences, time and context. This ontology determined the epistemology of the study.

4.3.2 Epistemology of the study

The epistemology of a research paradigm relates to the theory of the nature of knowledge, its scope and the validity and reliability of knowledge claims (Patel, 2012:11). According to Kivunja and Kuyini (2017:27), epistemology is defined as what can be counted as knowledge within the world, its nature, forms and how it can be acquired and communicated to other human beings; the nature of knowledge and justification; and how we come to know the truth or reality. In considering the epistemology of a research project, researchers need to ask questions such as: Is knowledge acquired or personally experienced? What is the relationship between the knower and what needs to be known? What is the relationship of the researcher and what is known? (Kivunja & Kuyini, 2017:27). The authors posit that in order to answer these questions, the researcher can tap from four sources of knowledge, which include intuitive knowledge (where the researcher relies on forms of knowledge such as beliefs, faith and intuition), authoritative knowledge (where the researcher gathers data from people in the know, books, and leaders in organisations), logical knowledge (where the researcher emphasises reason as the surest way of knowing the truth), and empirical knowledge (where the researcher emphasises that knowledge is derived from sense experiences and objective facts that can be demonstrated).

The transactional epistemology of the constructivist paradigm guided this research as expounded by some scholars. Creswell (2014:37) posits that subjective meanings are negotiated socially and historically and that the researcher focuses on interactions among individuals and specific contexts in which people live and work. Mack (2010:8) summarises the constructivist epistemology as follows, knowledge is gained through a strategy that respects the differences between people and the objects of natural sciences and therefore requires the social scientist to grasp the subjective meaning of social action; knowledge arises from particular situations and is not reducible to simplistic interpretation; knowledge is gained through personal experience; and knowledge is gained inductively to create a theory. Creswell (2007:11) contends that the constructivist epistemology requires the researcher to get as close as possible to the participants, interact and spend time with them, and negotiate the meaning of the phenomenon being studied.

Authoritative knowledge on the practice of quality control in the OSM environment was, therefore, gained from sources (Kivunja & Kuyini, 2017:27) by interacting with people in the know, the Examiners and Subject Managers, and reviewing literature on quality control in the marking of public examinations (Chapter 2) and on the OSM technology (Chapter 3). I interacted with Examiners and Subject Managers to negotiate and create the shared interpretations of OSM quality control and, together, generated a framework that could guide the practice of quality control in the OSM environment. This epistemology informed the methodology of the study, as discussed in the next section.

4.3.3 Methodology of the study

Methodology refers to the research design, methods and procedures used to investigate a phenomenon. Kivunja and Kuyini (2017:28) state that the methodology articulates the logic and flow of the process followed to conduct the study, how to obtain the data, knowledge and understanding that will enable the researcher to answer the research question and contribute to knowledge about the phenomenon under study.

Patel (2012:15) warns that in the research methodology or design, it is important to be aware of different approaches to research such as qualitative, quantitative and mixed methods approaches,

emphasising that there exists a methodological debate that is informed by ‘paradigm wars’. Johnson and Onwuegbuzie (2004:14) posit that the protracted debate about quantitative and qualitative research paradigms produced purists on both sides, who argue that qualitative and quantitative research paradigms and their associated methods cannot and should not be mixed. Johnson and Onwuegbuzie (2004:14) further posit that the debate was divisive, so they proposed a philosophy that leads to mixed methods that attempt to put together insights provided by qualitative and quantitative research. Although the authors accuse qualitative and quantitative purists of being divisive, their philosophy has become a distinct philosophy with its own methodological principles.

According to Creswell (2014:41), researchers must choose one of the three approaches to research, meaning qualitative, quantitative or mixed methodology, as guided by the research paradigm. Creswell (2014; 2009; 2007) insists that the constructivist philosophy is considered as a typical qualitative research because of its ontological and epistemological assumptions. Depending on the purpose of the study, the researcher should further select a type of study within the qualitative approach. There are five types of qualitative studies that a researcher can choose from, and these are phenomenology, grounded theory, ethnography, narrative and case study (Creswell, 2014; 2009; 2007). After reading and understanding the five types of qualitative studies in relation to the purpose of this study, I adopted the qualitative case study methodology because it best suits the research purpose. It is important for qualitative researchers to position themselves in the study (Creswell, 2014; Tracey, 2013). The next section discusses the axiology of the study, before discussing the case study methodology in detail.

4.3.4 Axiology of the study

The axiology of a paradigm questions the role of values, and the constructivist philosophy acknowledges that research is value laden and biases are always present (Creswell, 2007:11). Sanjari, Bahramnezhad, Fomani et al (2014:2) posit that researchers should clarify their role in the research process because they are actively involved in all stages, from designing the concept to reporting, arguing that qualitative researchers are considered as the ‘instruments of choice’ for several reasons. Human instruments are responsive to the environmental stimuli; they can interact with the context; they put together different pieces of information simultaneously; they

can process findings quickly; they provide quick feedback; and can sense inappropriate responses. Creswell (2014:235) concurs with Sanjari et al (2014:2), that the qualitative researcher is the key data collection instrument, and that qualitative researchers should position themselves in the study by declaring their background and experiences and explain how these might have influenced the interpretations of the data. Creswell (2007:11) contends that the constructivist axiology implies that the researchers should describe in detail, the context of the study and continually revise questions from experiences in the field.

I, therefore, declare my personal background and experiences that might have influenced the conduct and interpretation of the findings of this study. I brought into this study my experiences and values acquired from teaching Biology in the Zimbabwean secondary schools and assessment at tertiary institutions. I also brought my experiences as an examiner and subject manager for O and A Level Biology 5008 and 9190 syllabi examinations respectively and as the question paper development manager with the ZIMSEC, experiences that have given me considerable knowledge about quality control in both paper-based and onscreen marking. I was a full-time employee of the ZIMSEC at the time of data collection, so I conducted a ‘backyard research’ (Creswell, 2014:237) where the data was collected from subject managers who were colleagues, and examiners, who were part time employees of the same organisation. The researcher always explained how these experiences and firsthand knowledge would have shaped the study. The researcher gave a thick description of the context in which Biology (5008) examinations were marked onscreen, and always revised the research questions as necessary in the field. The next section discusses the history of case study research in order to position this study within the qualitative methodology.

4.4 Case study research

4.4.1 History of case study

As mentioned earlier, this study adopted the case study design. Case studies date back to the 1900s when they were qualitatively used in the fields of anthropology, where journeys were described in detail and cultures were systematically studied (Harrison, Birks & Franklin et al, 2017; Starman, 2013; Johansson, 2003). An anthropological approach to case study was used between 1920 and 1950 in the Chicago School of Sociology to study university cultures using

field-based observation of groups so as to understand their social and cultural lives (Harrison et al, 2017; Johansson, 2003). The approach was also adopted in medicine, social work and psychology (with the works of famous psychologists such as Piaget and Freud), political science, sociology and education (Harrison et al, 2017; Starman, 2013). Case studies were purely qualitative then, with participant observation as the main method of data collection (Johansson, 2003:6), until the emergence of the positivist philosophy.

The positivist realist ontology assumes that there exists a world of material objects, the objects exist whether they are perceived or not, some truth about the objects can be known through senses, and the objects and their properties are independent of the researcher (Kivunja & Kuyini, 2017:31). This ontology results in an objective epistemology which assumes that knowledge is acquired through the application of reason and scientific research methods that are based on deductive logic, formulation and testing of hypotheses, calculations, mathematical equations and extrapolations in order to draw conclusions. Issues of generalisability, validity and reliability are of primary concern in scientific research (Kivunja & Kuyini, 2017; Creswell, 2014; Atieno, 2009). When the positivist philosophy emerged in the field of science in the late 1940s, quantitative methods dominated research in social sciences up to the 1970s. Surveys, experimental and quasi experimental designs with quantitative empirical results were preferred over qualitative designs (Harrison et al, 2017:3). Case studies were criticised for lack of rigour, inability to generalize and valuing practical knowledge over theory (Harrison et al, 2017; Yin, 2003; Flyvbjerg, 2006), leading to methodological divisions in social science research. The positivist philosophy informed quantitative methods while the constructivist and interpretive philosophies informed qualitative methods, with science philosophers such as Peter Winch and Georg Henrik von Wright criticising the methodological influence of natural sciences on social sciences (Harrison et al, 2017; Johansson, 2003). Grounded theory, however, brought a different dimension to the case studies.

In the late 1960s, Glaser and Strauss combined qualitative methods used in the Chicago School of Sociology and quantitative methods of data analysis and proposed the grounded theory approach, resulting in an inductive methodology that used detailed systematic procedures to analyse data (Harrison et al, 2017; Johansson, 2003). This encouraged qualitative researchers to

revive the use of case study in many disciplines. Prominent case study methodologists, Robert Yin, Robert Stake and Sharan Merriam wrote extensively about case study research (Harrison et al, 2017; Yazan, 2015; Starman, 2013; Yin, 2013; 2004; 2003; Stake, 2008; 2005; 1995; Baxter & Jack, 2008) resulting in comprehensive literature on the subject. Literature shows that case studies follow a general trend characterised by distinct phases, which are philosophical commitments, defining the case and unit of analysis, designing the case study, data gathering, data analysis and strategies of achieving trustworthiness (Harrison et al, 2017; Yazan, 2015; Starman, 2013), as guided by philosophical commitments or paradigms.

Some scholars have analysed the work of Yin, Stake and Merriam to come up with philosophical beliefs that guide case study research (Harrison et al, 2017; Yazan, 2015; Starman, 2013; Brown, 2008). Three paradigms have been identified as guiding the case study research and these are discussed in the following section. The discussion of such helped to position the current study in the constructivist paradigm.

Some researchers argue that Yin, Merriam and Stake subscribe to specific philosophies in their case studies. Yazan (2015:142) concurs with Gaya and Smith (2016:534) that Stake's and Merriam's case studies are premised on the constructivist paradigm which believes in multiple and subjective realities that are transactional in their construction and researchers seek to understand and interpret the participants' view of reality. Harrison et al (2017:4), however, contends that Stake believes in constructivism while Merriam's work is premised on pragmatism, where there is no definite philosophical assumption. The truth is what works at the time (Creswell, 2014; 2007), and researchers can use both quantitative and qualitative methods to solve the research problem.

This debate serves as evidence that case studies are mainly informed by three main paradigms. Recent literature suggests that the three world views to case studies are the positivist, constructivist and pragmatic views, culminating in quantitative, qualitative and mixed methodology approaches respectively (Harrison et al, 2017; Starman, 2013). Starman (2013:30) further argues that case studies are associated with qualitative research and methodology but they can also be quantitative or a combination of qualitative and quantitative approaches, depending

on what the researcher prefers. I chose to use the qualitative case study to explore the practice of quality control in the OSM environment because it is in line with the constructivist philosophy that I hold. The next section discusses the qualitative case study methodology.

4.4.2 Qualitative case study

Njie and Asimiran (2014:35) posit that the qualitative method is used to study phenomena in their natural settings with the aim to interpret the phenomena in terms of the meanings people bring to them. The authors argue that qualitative research unravels complex phenomena that cannot be amassed by research methods that rely on figures and absolutes. These sentiments are supported by Yin (2011:7-8), who posits that qualitative research enables the study of people's lives in real world conditions; representation of the views and the perspectives of the people in the study; description of the contextual conditions within which people live; contributing insights into existing or emerging concepts that may help to explain human social behavior; and strives to use multiple sources of evidence rather than relying on a single source alone. Writing about the qualitative case study, Stake (2005:447) posits that the researcher should gather data on the nature of the case, particularly its activities and functions, its historical background, its physical setting, other contexts such as economic, political, legal and aesthetic, other cases through which this case is recognized, and the informants through whom the case can be known. I could not manipulate any variables in the OSM environment, neither could I control nor predetermine any outcomes. The phenomenon of quality control in the OSM of O Level Biology, which was the focus of this study, could only be explored by qualitative methods so as to understand it in the Zimbabwean context.

Stake (2005:459) posits that qualitative case researchers have several conceptual responsibilities, enumerating them thus:

1. Binding the case; conceptualising the object of study;
2. Selecting phenomena, themes or issues;
3. Seeking patterns of data to develop the issues;
4. Triangulating key observations and basis for interpretation;
5. Selecting alternative interpretations to pursue; and

6. Developing assertions and generalisations about the case.

Stake (2005:460) emphasises that except for number one, the other steps are similar to those of other qualitative researches. In the context of this study, the object of the study was the practice of quality control in the OSM of O Level Biology (5008) examinations. Issues were identified and articulated as research sub-questions in Chapter 1 Section 4. Patterns were sought from the data collected by document review, face-to-face interviews and WhatsApp discussions, from the subject managers and examiners respectively. The data from different methods and sources were triangulated to ensure trustworthiness of the methods as described later in this chapter. Alternative interpretations were pursued, and assertions were made as described in the data analysis and presentation section. The case was defined and bound as detailed in the next section.

4.4.3 Binding the case

Cases can be defined in many ways. In the context of this study, a case is viewed as an object of study and not a methodological choice as defined by Stake (2005:440), who goes on to identify three types of case studies the researchers can do, and these are the intrinsic, the instrumental and the multiple case studies. The intrinsic case study is when the researcher wants to understand a particular case, focusing on a unit, a person or an institution (Njie & Asimiran, 2014; Baxter & Jack, 2008; Stake, 2005). The intrinsic case study is not chosen because it is representative of other cases or because it shows a particular characteristic or problem (Baxter & Jack, 2008:548). Njie and Asimiran (2014:37) emphasise that a case is intrinsic “...when one wants a better understanding of a particular case and is, within all its particularity and ordinariness, a case of interest...”. The instrumental case study is where a particular case is examined to provide insight into a particular issue (Njie & Asimiran, 2014; Baxter & Jack, 2008). Stake (2005:445) elaborates an instrumental case thus:

The case is of secondary interest, it plays a supportive role and it facilitates understanding of something else. The case is still looked at in depth, its context scrutinised and its ordinary activities detailed, but all because it helps us pursue the interest. The case may be typical of others or not.

This implies that the OSM of Biology examinations can be examined to illustrate the practice of quality control, making it a secondary concern. The third type is the multiple case study in which the researcher is interested in more than one case. Stake (2005:445) posits that a multiple case study is the instrumental case study that can be extended to many cases. Although Stake (2005:445) emphasises that there are no hard and fast rules that separate an intrinsic and an instrumental case study, the focus of this study could best be accomplished by the instrumental case study, and not the intrinsic case study.

This study adopted the single instrumental case design because quality control was a common practice for any subject marked on screen. Ordinary Level Biology, therefore, represented a typical examination that was marked onscreen where the practice of quality control was explored. The OSM of Biology examinations is a typical case because many other subjects have been marked onscreen since 2012. Table 4.1 shows the subjects that have been marked onscreen in Zimbabwe from 2012 up to 2017.

Table 4.1 Subject components marked onscreen

No.	Subject name	Subject code	Year first marked on screen
1	English Language paper 2	1122/02	2014
2	French paper 2	3011/02	2013
3	Mathematics paper 1	4008/01	2012
4	Mathematics paper 1	4028/01	2012
5	Statistics paper 1	4041/01	2015
6	Statistics paper 2	4041/02	2015
7	Integrated Science paper 3	5006/03	2012
8	Biology paper 2	5008/02	2015
9	Biology paper 4	5008/04	2013
10	Physical Science paper 2	5009/02	2015
11	Physical Science paper 4	5009/04	2015
12	Physics paper 2	5055/02	2015

No.	Subject name	Subject code	Year first marked on screen
13	Physics paper 4	5055/04	2013
14	Chemistry paper 2	5071/02	2015
15	Chemistry paper 4	5071/04	2013
16	Food and Nutrition paper 1	6064/01	2015
17	Commerce paper 2	7103/02	2013
18	Principles of Accounts paper 2	7112/02	2013
19	Business Studies paper 1	7116/01	2015
20	Business Studies paper 2	7116/02	2015

Purposive sampling was used to select the case (Creswell, 2014; 2009; Dawson, 2007). The object (the case) of study for this research is, therefore, the OSM of Ordinary Level Biology, the investigation of which provided insight into the practice of quality control (the issue of interest). There are different purposive sampling strategies that qualitative researchers can use, depending on the objectives of the study. Etikan, Musa and Alkassim (2016:3) concur with Palys (2008:697) that typical case sampling is used to select a case that helps to set the bar of what is standard. As mentioned earlier, O Level Biology represented a typical examination that was marked onscreen where the practice of quality control was explored. My background and experiences also influenced the selection of the case. As indicated in the axiology of the study, I have substantial experience and firsthand knowledge with Biology examinations. This influenced the selection of the subject because I would better understand the phenomenon of quality control when studying Biology than any other subject. The OSM activities were examined in detail in the Zimbabwean context to explore the practice of quality control in the OSM environment. The OSM activity was not extended to other subjects and therefore it remains a single instrumental case study.

Now that the case has been identified, it is necessary to bind it to avoid the pitfall of attempting to answer a question that is too broad or a topic that has too many objectives for one study (Baxter & Jack, 2008:546). Case studies can be bound by time and space, time and activity or by definition and context, so as to provide boundaries that will limit the scope of the study (Njie &

Asimiran, 2014; Baxter & Jack, 2008). This study was bound by activity and time. The activity is the OSM of O Level Biology (5008) examinations and the time is the period spanning from 2013 to 2017. As discussed earlier in Chapter 1, the Zimbabwe Ministry of Primary and Secondary Education reviewed the curriculum and brought in a new curriculum in January 2017. The first examination for the new curriculum was sat for in November 2018 with a new set of examinations earmarked for OSM when the Council is able to securely print their examinations.

The ZIMSEC had been contracting other organisations to print some of its examinations, (including all papers marked on computer screens that require special printing processes), with the Director of ZIMSEC, Dr. L. Nembaware, arguing that the outsourcing of printing services became a source of examination leakages (Mawonde, 2018: www.nehandaradio.com; personal experience). There was rampant malpractice and cheating in the November 2017 examinations, resulting in government ordering a re-sit of O Level English Paper 2, a component that was printed outside ZIMSEC and marked onscreen. The decision to re-sit was overturned when the parents of the learners dragged the Council to the High Court (Manayiti & Munyeke, 2018; Bulawayo24, 2018; personal experience). In a bid to curb examination leakages, ZIMSEC acquired a state-of-the-art printing machine worth \$5m, and installed it at their premises in Norton (Zhou 2017: www.bulawayo24.com; personal experience). In May 2017, the then Minister of Primary and Secondary Education, Dr. L. Dokora, reported that ZIMSEC was still to buy a printing press that was meant to preserve the security and integrity of the examinations. The minister added that the printing press would cost \$5 007 126, and the Council had paid a deposit of \$1m. In October 2018, Dr. L. Nembaware said that the printing press had been acquired and will be used to print the June 2019 examinations. This will enable the Council to print all examinations in-house and be able to contain leakages that rocked the country in November 2017, and resume OSM (Mawonde, 2018; www.news.pindula.co.zw; personal experience).

The practice of quality control in the OSM environment was, therefore, explored in the old curriculum where O Level Biology (5008) examinations were marked. The OSM of O Level Biology (5008) examinations for the period covering November 2013 to November 2017 created the context in which the practice of quality control was studied. Stake (2005:449) explains that

the case is embedded in a number of contexts, with the historical context being always of interest. Five years would give enough historical contexts to the OSM of O Level Biology examinations and the practice of quality control. The case is made of three major units of analysis, which include:

- Zimbabwe School Examinations Council, the organisation that examines O Level Biology, which created the context in which the subject was marked;
- subject managers who developed the examinations and supervised examiners; and
- senior markers and normal markers, who did the actual marking.

Data from these units was collected and analysed to get a clearer picture of quality control in the OSM environment.

The next section discusses the issues or research sub-questions that led to the development of the conceptual framework that guided this study.

4.4.4 Research sub-questions

Stakes (2008:142) claims that as any other research kinds, a case study has a conceptual structure organised around a small number of questions referred to as issues or thematic lines. After defining or binding the case, the researcher needs to select phenomena, themes or issues. The issues are usually intricately linked to political, social, historical and personal contexts (Stake, 2005:459). To answer the main research question, five issues relating to the practice of quality control were identified and formulated into research sub-questions as listed below:

1. How does training of examiners and standardisation activities influence the quality of marking O Level Biology in the onscreen marking environment?
2. How is the quality of marking O Level Biology monitored in the OSM environment?
3. How do O Level Biology examination questions and mark schemes inform quality control in the OSM environment?
4. What are the opportunities and challenges of quality control in onscreen marking of O Level Biology?

5. How can quality control in the OSM of O Level Biology examinations be framed to provide guidelines for its practice?

Scholarly literature was reviewed around these issues and a conceptual framework was assembled in Chapter 2 Figure 2.3. That conceptual framework guided the exploration of the practice of quality control in the OSM environment in the context of Zimbabwe. The next section defines the population of the study.

4.5 The population

Qualitative researchers should purposefully select participants, sites or documents that will best help them to understand the phenomenon under investigation (Creswell, 2009:178). I, therefore, needed to draw the most appropriate sample made up of individuals with the ability and opportunity to provide the most accurate information about the practice of quality control in the OSM environment. There was, therefore, a need to clearly define the population from which the sample was drawn. The population of this study was defined as guided by the systematic and organised specification (SOS) put forward by Asiamah, Mensah and Oteng-Abayie (2017:1607-1621). The authors posit that qualitative researchers draw relatively small samples from large populations made of fairly legible members. Asiamah et al (2017:1609) argue that a population is better defined as guided by the phenomenon to be explored, the research objectives and the context in which the phenomenon is explored. They further argue that the researcher should define a general population made of participants with at least one attribute of interest. The general population is then refined by removing participants whose inclusion in the research would violate the goals and context of the study. The refinement will yield a target population that is further refined into the accessible population (Asiamah et al, 2017:1611).

The purpose of this study was to explore the practice of quality control in the marking of Ordinary Level Biology in the OSM environment in Zimbabwe in order to propose a framework which can help to improve the practice. The OSM of Biology examinations was conducted in a context where there were ICT related challenges and intermittent power cuts among other challenges discussed in Chapter 1, Section 2.2.

The general population of this study, therefore, consisted of all personnel who participated in the OSM of O level Biology (5008) in the period 2013-2017, made a total of 100 people. The personnel were six subject managers, 90 examiners, two ICT technicians and two clerical support staff (personal experience). The target population was defined by the researcher's interest to select participants who could best share their experiences and thoughts about the practice of quality control in the OSM of Biology (5008) examinations. Etikan et al (2016:3) proposes a purposive sampling strategy called expert sampling, where experts in a particular field are selected to provide information. The authors argue that expert sampling is useful when the research is expected to take a long time before it provides conclusive results or where there is currently a lack of observational evidence. In the context of this study, expert sampling was used because there is no observational evidence. Examiners and subject managers were considered as the experts in this study. As discussed in Chapter 2, Section 7 to 10, and Chapter 3, Sections 4.1 to 4.2, examination scripts are marked by examiners (senior and normal markers) who work under the supervision of subject managers. The general population was therefore refined by eliminating ICT technicians and clerical support staff who had no relevant experiences with the practice of quality control, resulting in a target population of 96, made of 90 examiners and six subject managers.

According to Asiamah et al (2017:1612), the qualitative researcher can either draw a sample from the target population using qualitative sampling strategies, depending on its size and complexity or from the accessible population that results from the refinement of the target. In the context of this study, the target population was too large, making it difficult to define the population from which to draw the sample. As stated above, the target population consisted of 90 examiners (16 senior and 74 normal markers) and six subject managers who participated in the OSM of O Level Biology in the five years spanning from 2013 to 2017. It was therefore important to zero in on the last marking session of O level Biology in the old curriculum, which was November 2017, to define the accessible population. This session was selected for the reason that it was relatively recent, so the participants were likely to remember their experiences with the phenomenon of quality control in the OSM environment.

Asiamah et al (2017:1614), however, emphasise that the target population should be refined by taking out members who are unwilling to participate or will not be available to participate, while acknowledging that some qualitative studies can have large accessible populations. Asiamah et al (2017:1616) further suggest that the accessible population can further be reduced if it is large. In the context of this study, the target population was refined by the need to reduce it for reasons of size, without necessarily determining the availability or willingness of population members to participate. The criteria of participant willingness or availability were used to define the final accessible population, further reducing it when some members were not available to participate in the study.

After I received the ethical clearance from CEDU (Appendix N), I applied for access from the Director of the ZIMSEC (Appendix A). As soon as I received the permission to collect data, I requested the examiners team structures for the November 2017 Biology (5008) examinations from the subject manager. A total of 84 examiners had marked 5008/2 and 5008/4 in 2017, made up of 14 senior and 70 normal markers. Work schedules (November 2017) indicated that six subject managers, two ICT technicians and two clerical support staff participated in the marking exercise. One subject manager was directly responsible for Biology (5008) and the other five helped in training examiners and allocating script portions to the examiners. They also helped to monitor the quality of marking when requested by the responsible subject manager (interview responses; personal experience). Elimination of ICT and support staff resulted in the first accessible population of 90 expert participants.

The criterion of availability of participants was used to define the final accessible population. I requested the lists of examiners who were on the database at the time of data collection so that I could identify the prospective research participants. The examiners were on the team structures for Biology (4025) examinations in the new curriculum. Some examiners who marked Biology (5008) on screen were no longer on the examiner database at the time of data collection and were therefore not available as prospective participants. A total of 58 examiners made of 13 senior and 45 normal markers were identified on the team lists. The six SMs were still available to participate in the study. A total of 64 participants in the second accessible population were available to participate in the study.

Asiamah et al (2017:1616) propose a stepwise population refinement process (SOS) for the careful analysis of the population, leading to the determination of the sample. The population description is summarised in Figure 4.1.

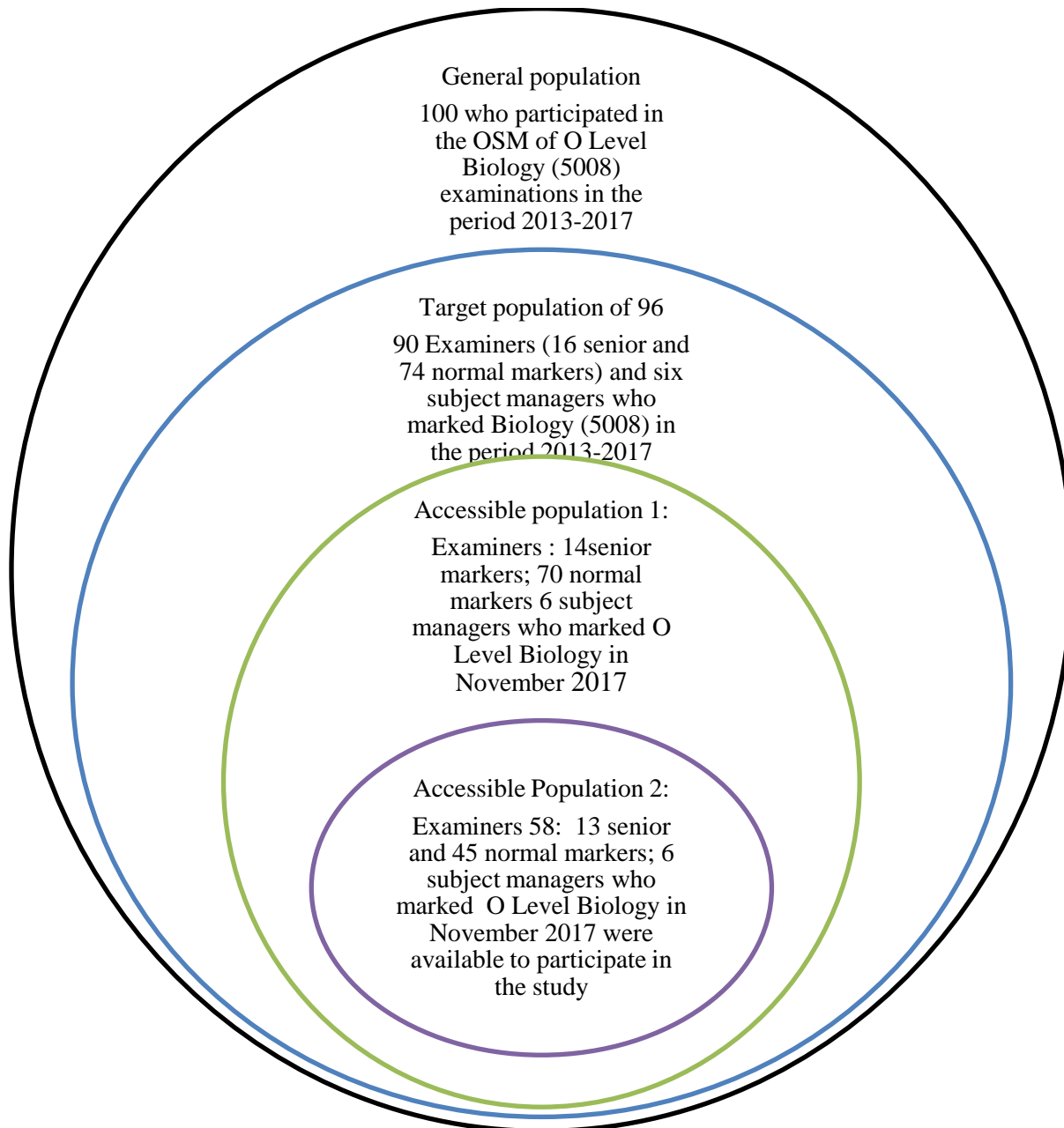


Figure 4.1: Target population of the study (Adapted from Asiamah et al, 2017:1611)

The sample of the study was therefore drawn from the accessible population 2. The next section discusses the sample and sampling strategies.

4.6 The sample

Gentles, Charles, Ploeg and McKibbon (2015:1776) posit that sampling in case studies is done twice, i.e. when the researcher samples the case and the data sources within the case. The case level sampling has already been discussed in Section 4.4.3 where the case was defined and bound by time and activity. This section discusses the sampling of data sources within the case, i.e. the sampling of examiners and subject managers. Purposive sampling strategies, where the researcher uses their own judgement to select the participants who have the best opportunity and ability to provide accurate information about the phenomenon under investigation, were used in this study (WHO, 2017; Palyis, 2008). There are many purposive sampling strategies that qualitative researchers can employ. As mentioned earlier, expert sampling was used to select examiners and subject managers who made the accessible population of the study. Another purposive sampling strategy was used to select the sample from the accessible population.

Creswell (2007:127) proposes the stratified purposeful sampling strategy, which is used to illustrate subgroups and facilitate comparisons. There are two categories of examiners who marked Biology (5008) examination scripts on computer screens, and these were senior markers and normal markers. Subject managers were full-time employees of ZIMSEC, who also participated in the OSM of O Level Biology (5008) examinations. They made a third category of participants. The stratified purposeful sampling strategy was, therefore, used to select the participants of this study. Data on the practice of quality control in the OSM environment were collected from the three distinct categories of participants to gather multiple perspectives and allow triangulation of the findings.

The intention of the qualitative researcher is not to generalise from the sample to a population, but to explain, describe and interpret the phenomenon. Therefore, sampling is not a matter of representative opinions, but of information richness (Guetterman, 2015:3). A total of 33 participants, made up of 11 senior markers, 18 normal markers and four subject managers participated in the study. The adequacy of the sample was assessed by data saturation, which is a point of information redundancy when new information would not add value to the study (Gentles et al, 2015; Guetterman, 2015). I stopped asking questions when responses from the

examiners resembled the ones from the SMs and document review. Table 4.2 summarises the sample of participants in the study.

Table 4.2: Interview participants

Type of Interview	Participants	Number
Face-to-face	Subject managers	4
	Senior markers	1
	Total	5
Focus group (WhatsApp Platform)	Examiners	
	Senior markers	11
	Normal markers	18
	Total	29

The senior marker who was interviewed face-to-face also participated in the focus group discussions, so a total of 33 participants shared their views on the practice of quality control in the marking of Biology (5008) examinations.

Another purposive sampling strategy was used to select documents for review. According to Palys (2008:697), criterion sampling is used to select cases that meet certain criteria, giving an example of a health researcher who selected men who had been clients of sex workers. In this study, documents were selected using the criteria that their content related to quality control in the OSM of examinations in Zimbabwe's old curriculum. Literature indicates that a representativeness of documents is not important in their selection (Triad, 2016; Ahmed, 2010; Bowen, 2009), with Ahmed (2010:4) emphasising that it is difficult for the researcher to determine if selected documents are representative of all documents written on the phenomenon being investigated. I requested and received documents that related to the OSM of examinations between 2013 and 2017. I however selected and reviewed one document that related to the OSM of examinations in 2012 because it contained important information that influenced the practice of quality control in subsequent examinations and it also helped to clear some contradiction in literature about the time when ZIMSEC adopted the OSM technology. I used criterion sampling to select documents that were relevant to the practice of quality control in the OSM environment

using a document review guide (Appendix D). Table 4.3 lists the documents that were selected and reviewed.

Table 4.3: List of documents reviewed

Category	Type of document	Number reviewed
Public records	5008 Syllabus	1
	Examination circulars	3
	Question Papers	
	5008/2	3
	5008/3	2
	5008/4	5
	Mark Schemes	
	5008/2	3
	5008/3	2
	5008/4	3
	Works schedules	2
	Reports	14
Personal documents	Memos	6
	Emails	6
	Minutes of meetings	4
Physical evidence	Training materials/guides	7
Total		55

The next section discusses the data collection methods.

4.7 Data collection methods and instruments

Researchers can get insights into a case using observation, interviews, document review and visual artifacts (Creswell, 2014; 2009). Observation was not relevant to this study because ZIMSEC had suspended OSM at the time of data collection, as discussed earlier. Instead, interviews and document analysis were used to collect data about the practice of quality control

in the OSM of Biology (5008) examinations. The next section discusses the document analysis method and how it was used in the study.

4.7.1 Document review

Documents provide useful information and there are several reasons for considering their use in a study. Documents are a record of past events, providing background information as well as historical insights. They can provide data on the context in which the research participants operate, and can indicate the conditions that impact on the phenomenon under study (Triad 2016; Bowen 2009). Documents provide information that can suggest an interview. The information from documents provides valuable knowledge about the phenomenon being studied and provides a means of tracking change and development as well as corroborating or verifying data from other sources (Triad 2016; Bowen 2009). Contextual data were very important in this case study. Documents provided background information about the OSM of examinations in Zimbabwe. The context in which ZIMSEC decided to adopt the OSM technology, the assessment framework that guided the examinations, the technological infrastructure, the human resource capacity and the quality control activities in the OSM of Biology (5008) examinations were provided in the documents. I analysed the documents, watching out for issues that needed clarification and conducted follow-up interviews with the examiners and subject managers. The documents also helped to track changes in the practice of quality control in the OSM environment between 2013 and 2017.

There are three basic types of documents that can be reviewed, such as public records, personal documents and physical evidence. Public records are the official records of an organisation's activities, for example, mission statements, annual reports, policy manuals, strategic plans, course syllabi, event organisation programs, internal correspondence and others. Personal documents are first-hand experiences of individual actions, experiences and beliefs. Physical evidence includes flyers, posters, agendas, handbooks and training materials (Triad, 2016; Bowen, 2009). The three categories of documents were collected and reviewed as indicated in Table 4.3. A form was designed to guide the review of the documents (Appendix D).

Bowen (2009:33) cautions that documents should not be treated as necessarily accurate or complete records of the events that occurred, urging researchers to determine the relevance of the documents to the research problem, the purpose and the conceptual framework of the study. The purpose of this study was to explore the practice of quality control in the OSM environment to propose a framework that could guide the practice. This purpose was accomplished by answering five research sub-questions in Chapter 1 Section 4 as guided by the conceptual framework in Chapter 2 Section 2.11. A summary of the questions answered by each data source is shown Table 4.4.

Bowen (2009:33) posits that the researcher should determine the authenticity, accuracy, credibility and representativeness of the selected documents. Bowen (2009:33) concurs with Triad (2016:3) that the researcher must evaluate the original purpose of the document by determining the target audience, whether the author was a first-hand source or used second-hand source. Triad (2016:3) posits that the researcher should consider the latent content of each document. The author explains latent content as the style, tone, agenda, facts or opinions that exist in the document. As I analysed some documents that were in my work file, I came across minutes of meetings that were held to commission marking exercises. However, all the minutes were not signed by the chairpersons, who varied from meeting to meeting. I approached some of the listed attendees and the chairpersons, who were still at ZIMSEC, to confirm the authenticity of the minutes. They confirmed that the minutes were true records of the meetings they attended, stating that there was a culture of not signing minutes at the organisation. Once the minutes were read and passed as a correct record in a meeting, they were then distributed to attendees straight from the secretary's office via email, without the signature of the chairperson. I also received some of the minutes of such meetings via email and they too were not signed by chairpersons. I also confirmed the authenticity of a communication that was written to the software provider (SWP) by the then marking administration manager (MAM). There was no date on the communication but it related to the results of a review meeting for the November 2013 examinations. Fortunately, the communication had been sent by email and the MAM confirmed the date, which tallied with the period of the review meeting.

I could not establish the representativeness of the documents that I selected and reviewed. I knew from personal experience that a specific section within ZIMSEC generated and received important documents related to OSM of examinations since its inception in 2012 up to 2017 when the last examinations were marked on screen. I wrote a request letter to the head of the division, who authorized the supervisor of that section to provide me with the documents. I made several follow ups with the supervisor who kept on promising to give me the documents. When I realised that time was running out on me I decided to use the documents that had accumulated in my work file during the OSM era and those provided by the subject manager for Biology (5008). Even if I were to be provided with the documents, I still would not know how many were withheld by the section and the subject manager, and how many I missed as I worked in the OSM environment.

Document analysis is efficient as it involves selection rather than data collection; documents are readily available in the public domain; they are cost effective; they are not affected by the research process (unobtrusive and non-reactive); documents are stable, exact and cover a wide range of events and settings over a long time (Bowen, 2009:31). As discussed in Section 4.7.2, data were collected at a time when the Zimbabwean economy was deteriorating, with prices of fuel hiked. Document analysis provided readily available data at almost no cost. Minimal costs were incurred in photocopying and scanning the documents that provided information about quality control in the OSM environment over a period of five years, which could be forgotten by subject managers and examiners. I kept going back to the documents without fear of offending, irritating or inconveniencing them like the human participants. Document analysis, however, has its own limitations.

Some documents may contain insufficient detail; others may not be accessible because they are deliberately blocked; the selection of the documents might be biased; and the documents were not created for research purposes (Bowen, 2009:31). The document review guide assessed issues of credibility and authenticity of the documents. Ahmed (2010:5) posits that the researchers need to decide what inferences to make about issues other than factual assertions in documents, suggesting that there is need to interview key informants who might give their perceptions, meanings and interpretations of the documents. To mitigate some of the limitations of document

analysis data collected were triangulated with data from interviews. Where the document contained insufficient detail, I would ask follow-up questions to the subject managers. In most cases the interview responses added some missing information or confirmed the information in the documents, especially the syllabus, the question papers, mark schemes and reports. The subject managers mentioned and gave me the item designation summary (IDS), a document that I had not thought of reviewing, even though I had copies of it on my office computer. As discussed earlier, I could not ascertain that I got all the documents that related to the OSM of the examination.

Data were also collected by interviews.

4.7.2 Interviews

The purpose of the interview is to explore the views, experiences, beliefs and motivations of individuals on a phenomenon to gain a deep understanding thereof (Gill, Stewart, Treasure & Chadwick, 2008:291). Oltman (2016: www.uknowledge.uky.edu) posits that in the qualitative paradigm, interviews are considered as the primary method that allows the researcher to enter into the perspective of the participant. Interviews were used to collect data about quality control in the OSM of Biology (5008) from the examiners and subject managers.

Dawson (2007:28) identifies three types of interviews, which are unstructured interviews, mainly used for life history studies; semi-structured interviews where the researcher collects specific information and therefore asks the same questions in each interview; and structured interviews mainly used in quantitative surveys. Bolderston (2012:68) states that semi-structured or unstructured interviews are normally used in qualitative research, emphasising that the semi-structured format allows the researcher to set the agenda but following the participant's thoughts and exploring tangential areas. According to Coniam (2011a:1046), one of the main advantages of using semi-structured interviews is that although they do not reveal as much of the whole picture as unstructured interviews, they provide an initial framework for defining categories and the subsequent recording of salient details in the analysis, thus saving a great deal of time in the labour-intensive context of qualitative data analysis. The author further argues that semi-structured interviews avoid too much interview fatigue for the interviewees, while still ensuring

that they feel free to say whatever concerns them. This study adopted the semi-structured interview to benefit from its advantages (Coniam, 2011a:1046).

Creswell (2014:242) and Bolderston (2012:68) concur that qualitative researchers can conduct interviews by email, face-to-face, focus groups, online focus groups or by telephone. In an accounting research project to understand how sustainability reports are prepared and assured in the Australian and New Zealand contexts, Farooq and de Villiers (2017:8) conducted 50 interviews as follows: 43 by telephone, one computer-based audio (Skype) and six face-to-face. They noted that the majority of participants preferred the telephone interviews because they had busy work schedules, so they provided their mobile numbers. The computer-based interview was of poor quality due to erratic internet connectivity. The authors noted that the Skype call was poor despite the fact that they were in New Zealand and the participant was in Australia, countries with good internet connectivity. They concluded that the internet-based technology should not be relied upon for qualitative interviews. Internet-based technologies would pose more challenges in Zimbabwe where the internet was patchy and there were intermittent power cuts that limited internet access, as discussed in Chapter 1 Section 2.3. The WhatsApp platform however proved to be the best option to access the examiners, after a careful consideration of the economic conditions that prevailed in the country.

The subject managers who were colleagues as indicated in the axiology of this study were interviewed face-to-face because they were easily accessible. Telephone interviews seemed appealing and were chosen after a thorough consideration of the prevailing economic situation in Zimbabwe. The Zimbabwe Energy Regulatory Authority (ZERA) increased fuel prices by 150% in January 2019, resulting in the erosion of salaries (Langa, 2019; Ncube & Langa, 2019). The ZERA continued to announce fuel prices that were at par with the value of US\$1. By October 2019 fuel prices were pegged at ZW\$14.97 and ZW\$16.64 for petrol and diesel respectively (Nyathi, 2019; Personal experience). It was reported that by October 2019 a full tank of fuel cost more than the salary of a cook and more than half the salary of a medical doctor (Nyathi, 2019: www.zimeye.net).

In reaction to the economic hardships, Zimbabwe's two largest teachers' unions issued a joint statement communicating the teachers' plans to go on strike at the beginning of February 2019, arguing that they were incapacitated by the eroded salaries. The teachers demanded that their salaries should be paid in United States dollars or be hiked by 500% in bond notes (Manayiti, 2019; Marawanyika & Sguazzin, 2019; Sibanda, 2019). Teachers' unions reported that their members heeded the strike call and at least 80% of teachers were not reporting for duty by the 6th of February 2019. The unions reported that the state security agents and politicians were visiting schools, threatening to fire teachers who failed to report for duty and that some teachers had been picked up by police for questioning (Maravawanyika & Sguazzin, 2019: www.bloomberge.com). However, the Public Service Commission (PSC) dismissed the strike as a non-event.

The PSC said they had received reports of isolated cases of teachers who had absconded from duty and would deduct some money from their salaries (Matiashe, 2019: www.newsday.co.zw). The Ministry of Primary and Secondary Education seemed to have realised that the majority of teachers were actually on strike and called for a meeting with the two teachers' unions. The unions called off the strike without giving reasons and urged teachers to report at their work stations on 11 February 2019 and wait for further commands and directions. The union leaders emphasised that they reserved the right to regroup, re-strategise, engage and prepare for other disabling forms of industrial action should their demands be ignored. Some teachers' unions had, however, chosen dialogue with government instead of the industrial action. A leader of one such union said the strike called by the two large unions was technically illegal, so they had no option but to call it off (Razemba, 2019; Marawanyika & Sguazzin, 2019). This is evidence that teachers' unions in Zimbabwe were divided.

It later turned out that the two large teachers' unions had called off the strike on the understanding that the government would meet them separately to address their concerns (Ndlovu 2019: www.bulawayo24.com). By 22 February 2019, it was reported that the teachers' unions had requested a meeting with the PSC to discuss the salary adjustment. The PSC however refused to meet the unions, arguing that the commission would only meet the Apex Council, and not unions or associations. This irked the teachers' unions who said that they would meet their

members soon to map a way forward (Mhlanga, 2019; Ndlovu, 2019). The prevailing situation required a careful consideration of appropriate data collection methods.

The majority of the O Level Biology (5008) examiners were teachers. The few who were not teachers might also not be readily available at their work stations in the wake of the eroded salaries. The prevailing economic situation rendered focus group discussions and face-to-face interviews for examiners inappropriate because they were not financially sustainable. Like other workers in Zimbabwe, I could not afford to fuel my car most of the time; neither could I rely on the erratic power and subsidised public transport that was unreliable. Furthermore, it might not have been prudent to convene focus group meetings with teachers at a time when they were allegedly being watched by state security agents. The economic hardships and possible security threats affected the participants in the same way they affected me. To avoid major changes in the research design, I planned for telephone interview for senior and normal markers. Data collection instruments were designed for face-to-face interviews and telephone interviews for subject managers and examiners respectively. The ethical clearance from CEDU was applied for on the strength of the same. However, an unexpected challenge cropped up at the time of data collection.

Massive power cuts, extending for 18 hours per day crippled the economy (Kuwaza, Kwinjo & Gonditii, 2019; Samaita, 2019). The Zimbabwe Electricity Supply Authority (ZESA) owed millions of dollars to the Mozambican and South African power utilities, reducing the power supplied to the country. The ZESA reportedly projected a loss to the tune of ZWD1.4 billion (Tazviinga, 2019: www.theindependent.co.zw), destroying the hope of paying up the debts. This made phone calls almost impossible to make. Cell phone numbers were not reachable most of the time, mainly because the batteries were flat and the network was poor (personal experience). My two cell phones were also not usable most of the time because they too had no power. Despite the patchy internet and power cuts, I opted to use WhatsApp; a web based messaging application, to conduct focus group discussions with the examiners. The WhatsApp will be discussed in detail later in this chapter.

Oltman (2016: www.qualitative-research.net) posits that interviews are negotiated accomplishments for both the interviewer and the interviewee and are shaped by the contexts and situations in which they take place. The author advises that the researcher should consider the interviewer and interviewee contexts when deciding to use the face-to-face interview or the focus group discussions. The following section discusses the face-to-face interviews and how they were used in this study.

4.7.2.1 Face-to-face interview

Face-to-face interviews are conducted one-on-one with the participant, hence the need to consider the interviewer and interviewee contexts as argued by Oltmann (2016: www.qualitative-research.net). The author argues that for the interviewee, face-to-face interviews allow the researcher to capture non-verbal language and cues and is less likely to have technological problems except with recording devices. Oltmann (2016: www.qualitative-research.net), however, posits that face-to-face interviews raise some challenges for the interviewer, enumerating them thus:

- Time and financial costs that arise from the need to travel to meet the interviewee at a place of their choice.
- Often limited by geographical distribution of participants.
- The interviewer can be endangered depending on the location and time of the meeting.
- Note taking can be obtrusive, resulting in the interviewer missing important issues.

These factors were considered in the decision to use face-to-face interviews for subject managers, who were colleagues. The subject managers were likely to provide rich information on the practice of quality control in the OSM environment because they were the custodians of the examination process, from test design to grading (personal experience). They were also likely to provide documents that were relevant to the practice of quality control, enhancing the quality of data collected. There was no need to travel long distances to interview the subject managers, even if they were to choose to be interviewed out of their offices. The interviews were, therefore, not time- and cost-intensive. I had initially planned to seek consent from the subject managers to record the interviews to reduce writing and the chances of missing important issues. However,

the plan was discarded due to time constraints that emanated from the deteriorating economic conditions discussed earlier. Notes were taken during the interviews, which were then written up and sent to the subject managers for verification. Face-to-face interviews have contextual issues that relate to the interviewees as explained above.

Oltmann (2016: www.qualitative-research.net) concur with Block and Erskine (2012:432), that although there are low dropout rates in face-to-face interviews, with the attendance rate pegged at 70%, the participants may feel pressured to be available for the interview once they have agreed to participate. Oltmann (2016: www.qualitative-research.net) lists participant contexts that are unfavourable due to the following reasons:

- Face-to-face interviews undermine confidentiality: It will be difficult to hide the identity of the participant from the interviewer; anonymity will depend on the integrity of the interviewer and data protection.
- The interview is invasive to participants (often in their homes or offices), compromising the privacy of participants.
- There are power imbalances in face-to-face interviews; because the participant can see the researcher, they may feel socially pressured to respond to the interviewer.
- The participant may conform to social pressures and under-report on topics they see as controversial.

On the whole, face-to-face interviews raise ethical issues that the researcher addressed in the trustworthiness and ethical issues sections. Some of the challenges inherent in the face-to-face interviews were reduced by the WhatsApp discussions. The next section discusses the use of WhatsApp as a data collection tool.

4.7.2.2 WhatsApp Focus groups

WhatsApp is a messaging application that can be downloaded on smart phones and has a variety of functions like text messages, images, audio files, video files, and links to web addresses (Qudsia, Farooq & Muhammad et al. 2017:39). According to Shahid (2018:14), WhatsApp was launched by Brian Acton and Jan Koum in November 2009, replacing short message services

(SMS) by providing additional information. The additional information included ticks that indicate whether a message has been sent, delivered or read by the recipient. Other features include sending text messages, audio notes, videos, location files, and many more (Shahid, 2018:14).

WhatsApp is widely used across the world by almost a billion people on their mobile phones (Boughton, 2019; Shahid, 2018). At the time of writing, WhatsApp was also widely used as a means of communication in Zimbabwe, with a population of 16.7 million, an internet penetration of 50% with close to 100% of the population owning mobile phones (www.theindependent.co.zw). It was reported that WhatsApp was by far the cheapest means of communication, with mobile network operators providing data bundles specific to WhatsApp (Karombo, 2017; Mangudhla, 2016). The examiners were therefore interviewed by focus group discussion on the WhatsApp platform.

WhatsApp is increasingly being used for research purposes. The UNDP (2018:2) used WhatsApp to conduct qualitative surveys on the needs, perspectives, fears and local conflict dynamics of host communities and Syrian refugees in Lebanon. Qudsia et al (2017:39) used the platform to explore the potential of WhatsApp as an instructional strategy for 4th Year MBBS students in Ophthalmology. Tarisayi and Manhibi (2017:34) used WhatsApp to study the interaction of Heritage Studies teachers on a WhatsApp group to address challenges in the implementation of the new curriculum in Zimbabwe. I used the platform after considering its use as a research tool.

The WhatsApp platform has several advantages. The platform has scale and speed, enabling the collection of data from a large sample within a short time because of its features that allow users to create and manage groups of contacts (Boughton, 2019; UNDP, 2018). The platform is accessible real-time and by logging onto a web browser, allowing qualitative researchers to gather and organise data effectively on their phones and computers (Boughton, 2019: www.info.angelfishfieldwork.com). In this study, data were collected from 29 participants in three days, despite the fact that cell phone batteries had no power for the greater part of the day due to power cuts that extended for 18 hours every day. Sometimes I got instant responses if

some participants happened to be online when I posted the questions, confirming that WhatsApp is accessible real-time. Those who were not on-line were still able to respond to the questions later, as long as they had data bundles, offering some convenience to the participants.

WhatsApp is cost effective. WhatsApp Messenger is downloaded for free and the messages are sent for free as long as there is Wi-Fi or data bundles available (Boughton, 2019; UNDP, 2018). In the Zimbabwean context, the users need to purchase data bundles for WhatsApp from mobile network operators. As the economic situation deteriorated and prices skyrocketed, the price of data bundles also increased to significant levels. Weekly WhatsApp bundles were pegged at a minimum of ZW\$1 in March 2019 and by the time of writing the minimum cost was ZW\$5 (www.thezimbabwean.com; personal experience). The cost of data bundles had risen to substantial amounts by the time of data collection, so I had to carry the cost of data bundles.

I joined two chat groups for Biology 4025/2 and 4025/3. As soon as I introduced myself and the research, one examiner insinuated that I needed to provide some airtime first. The examiner wrote "...Cushion first..." in apparent reference to airtime before I could say much about the study. I intended to send airtime to participants who would have answered my questions at the end of the data collection period. I had to send ZW\$10 (R20 at that time) to everyone who responded to my questions in the evening of the first day after realizing that the response rate was too low. The participants must have informed each other that 'cushion' was being paid to those who had responded to research questions. More and more responses started coming on the second and by the third day I had reached the saturation point. On the second day I posted a message on the two groups informing the participants that I had sent airtime to mobile numbers that had responded to my questions and those who had not received the money should come to my inbox. Some participants went on to answer just one question in one or two words and came to my inbox to inform me that they had not received airtime. I sent them the money hoping that they would respond to more questions, but some never did.

According to the UNDP (2018:3), the WhatsApp platform limits the power and interference of the researcher in people's stories. While the researcher still asks questions, they cannot steer the narrative through follow-up questions or prompts, giving people space to talk about issues that

matter to them. The authors emphasise that there is no personal relationship between the researcher and the participant that could produce social desirability or silencing effect. In this study the WhatsApp platform reduced the power imbalances that could have reduced the accuracy of the responses from the examiners. As indicated in the axiology of this study, I was a full-time employee of ZIMSEC and had worked with some of the Biology (5008) examiners as the subject manager for the subject. Although I knew some of the examiners by name, I could not tell who they were on WhatsApp because they did not use their real names. I only came to know their names when I sent airtime money to their mobile numbers. All the same, I could not link their responses to their names. The WhatsApp platform offered the examiners some degree of privacy that cannot be afforded by face-to-face interviews.

The UNDP (2018:2) states that WhatsApp platform reduces the power of gatekeepers. From my experience, WhatsApp reduces the power of gatekeepers when the researcher is dealing with individuals or has created a group for the purpose of collecting data. In this study I could not deal with individual examiners, neither could I create a group, given the context in which the data was collected. I had to join a group that was formed for the purpose of communicating issues that related to the marking of Biology (4025) examinations, so I badly needed the gatekeeper who was the subject manager. It is the subject manager who told me about the existence of the groups and introduced me to the examiners before my mobile number was added to the two groups. Like any data collection method, the WhatsApp platform has its own limitations.

The UNDP (2018:3) posit that it may be difficult to access phone numbers depending on the country and data context. The authors propose that the researcher can make a data sharing agreement with a mobile network operator; collect phone numbers through local stakeholders or invite the participants to subscribe to the research by providing a link on to all registered numbers. It was however not difficult for me to get mobile numbers for the examiners since the contact details for Biology (5008) 2017 examiners, including mobile numbers, were readily available at ZIMSEC. Although I had accessed mobile numbers from the examiner database, I did not use them because I joined already existing groups.

According to the UNDP (2018:3) the WhatsApp creates sampling bias that the researcher cannot account for in who chooses to respond to the questions and who does not. This was true for this study. Although I had accessed the mobile numbers for the examiners who marked Biology (5008) examinations in 2017, I was not sure if all of them were still in use; neither could I tell if the numbers were on WhatsApp, forcing me to join already existing groups. Each group was made of 58 participants, which means I accessed 116 examiners, including some who had been newly trained to mark Biology (4025) examinations. I also noticed that a few mobile numbers were on both groups. There was no way that I could tell if all examiners in the accessible population were on the WhatsApp groups, some either changed their numbers or were not marking anymore. Examiners who had no smart phones were automatically eliminated from the discussion. I could not control the number of responses per question; neither could I control the number of senior and normal markers who responded to the questions. This resulted in responses skewed towards senior markers. The ratio of senior to normal markers was 13:45 in the accessible population and but it rose to almost 11:18 on the WhatsApp platform. There was no time at which all examiners responded to a question, very few responded and at times I had to solicit responses from the senior markers.

On one group I even got some digressions from the questions on the first day when some examiners chose to talk about social issues that were current in Zimbabwe, like a soccer coach who had been fired from a team. Except highlighting the question and soliciting for responses, there was nothing much that I could do to make the examiners stick to the research issues only. It seemed as if there were no ground rules that prohibited the examiners from discussing issues not related to the purpose of the group, so as a guest member I could not insist on them sticking to the questions lest I offended them. I however noticed that the digressions were reduced on the second and third days. I attributed this to probing by the subject manager who politely reminded the examiners on the group to "...please answer Mrs. Masiri's questions....." This experience made me think that joining a group that was created for other purposes was not the best option. It could have been better if I had created a group for examiners who were willing to participate in my study. I however had lost valuable time trying out the telephone interviews that I had planned, and joining existing groups was the next best option under the circumstances. The other group remained focused on the research questions up to the third day.

According to the UNDP (2018:4), sample sizes between 2000 and 3000 should give respondents between 340 and 510 respondents, giving a response rate of 17%. As indicated earlier, the accessible population of this study was made of 58 examiners and 29 responded to the questions I posted on the WhatsApp groups. The response rate was, therefore, 50%. This confirmed that the majority of the examiners who marked Biology (5008) examinations were still marking in the new curriculum. After three days I closed the discussion and logged onto the web browser on my computer and downloaded the charts for management and analysis. I requested to stay on the groups for one month so that I could make follow-ups as I analysed the data, after that I thanked the participants and exited the groups.

Data was collected by document review, face-to-face interviews and WhatsApp group discussions, in that order. The interview schedule initially designed for telephone interviews was revised to focus it on group discussions and to include questions relating to issues raised in documents and face-to-face interviews. Table 4.4 summarises research sub-questions and the data collection instruments that provided the data.

Table 4.4: Research sub-questions answered by data collection instruments (Appendices)

Research sub-question	Data collection Instrument
1. How does training of examiners influence the quality of marking O Level Biology in the onscreen marking environment?	
Training	Document analysis form: Appendix D Interview schedule for subject managers: Appendix E: Question 2; 3; 4 Interview schedule for examiners: Appendix F: Question 1; 2; 3 Findings from documents – question 1: Appendix K Face-to-face interview write up: Appendix G WhatsApp chats write up: Appendix H
Standardisation	Document analysis form: Appendix D

Research sub-question	Data collection Instrument
	<p>Interview schedule for subject managers: Appendix E: Question 3; 4</p> <p>Interview schedule for examiners: Appendix F: Question 2; 3</p> <p>Findings from documents – question 1: Appendix K</p> <p>Face-to-face interview write up: Appendix G</p> <p>WhatsApp chats write up: Appendix H</p>
2. How is the quality of marking O Level Biology monitored in the OSM environment?	<p>Document analysis form: Appendix D</p> <p>Interview schedule for subject managers: Appendix E: Question 5; 6</p> <p>Interview schedule for examiners: Appendix F: Question 4; 5;</p> <p>Face-to-face interview write up: Appendix G</p> <p>WhatsApp chats write up: Appendix H</p>
3. How do O Level Biology examination questions and mark schemes inform quality control in the OSM environment?	<p>Document analysis form: Appendix D</p> <p>Interview schedule for subject managers Appendix E: Question 7</p> <p>Interview schedule for examiners: Appendix F: Question 6; 7; 8</p> <p>Face-to-face interview write up: Appendix G</p> <p>Appendix I: Question paper review</p> <p>Appendix J: Mark scheme review</p> <p>Document analysis form: Appendix D</p> <p>WhatsApp chats write up: Appendix H;</p> <p>Question paper review: Appendix I</p> <p>Mark scheme review: Appendix J</p>
4. What are the opportunities and challenges of quality control in onscreen marking of O Level Biology?	
Opportunities	Interview schedule for examiners: Appendix

Research sub-question	Data collection Instrument
	F: Question 10 WhatsApp chats write up: Appendix H Face-to-face interview write up: Appendix G Interview schedule for subject managers: Appendix E: Question 8
Challenges	Document analysis form: Appendix D Interview schedule for examiners: Appendix F: Question 10 WhatsApp chats write up: Appendix H; Face-to-face interview write up: Appendix G
5. How can quality control in the OSM of O Level Biology examinations be framed to provide guidelines for its practice?	Document analysis form: Appendix D Interview schedule for examiners: Appendix F all items WhatsApp chats write up: Appendix H Face-to-face interview write up: Appendix G Appendix I: Question paper review Appendix J: Mark scheme review Interview schedule for subject managers: Appendix E: all items

The next section discusses data analysis.

4.8 Data presentation and analysis

Data analysis is the process of making meaning from the data collected. According to Creswell (2014; 2009), qualitative studies are characterised by inductive and deductive data analysis. The author explains that in the inductive analysis, the researchers build patterns, categories and themes from the data, working back and forth between the themes and the database until a comprehensive set of themes has been established. In the deductive analysis, the researchers search for evidence from the data that supports each theme, and determine if there is need to

collect additional information (Creswell, 2014:234). Qualitative researchers therefore analyse data as they collect it. This type of analysis is called thematic analysis.

Nowell et al (2017:2) define thematic analysis as a method of identifying, analysing, organising, describing and reporting themes found within a data set, arguing that thematic analysis can be used across a wide range of epistemologies and research questions. Creswell (2014:245) proposes six stages of implementing thematic analysis arranged in a linear fashion. Nowell et al (2017:4) also proposes another six stages of implementing thematic analysis in a linear fashion, but emphasising on activities that enhance trustworthiness of the findings. However, Creswell (2014:245) concurs with Nowell et al (2017:4), that data collection, data analysis and write-up of findings occur concurrently in qualitative research, and that thematic analysis is an iterative and reflective process that involves a constant moving back and forward between phases. The concurrent activities of data collection and analysis enable the researcher to decide that the saturation point has been reached. In this study, the six stages put forward by Creswell (2014:245) and Nowell et al (2017:4) were adapted and used to implement thematic analysis in an iterative manner as well. Table 4.5 summarises how the six stages of thematic analysis were used.

Table 4.5: Six-stage thematic data analysis (adapted from Creswell 2014:245; Nowell et al 2017:4)

Stage	Activities
1. Preparing for data analysis	Optically scan documents, sort and arrange data by source – documents, face-to-face-interviews, and WhatsApp discussions; summarise interviews (see Appendix G & H) and contact participants for confirmation; create a matrix of interview responses (WhatsApp); store data in organised files on a password protected computer and locked drawer.
2. Familiarising with data	Read data to get a general meaning, document thoughts about the potential themes or codes, write notes on the margins of transcriptions.
3. Data coding	Organise data according to categories of issues raised by the participants about the practice of quality control in the OSM

Stage	Activities
	environment: watch out for expected codes based on literature such as examiner training and standardisation, monitoring of marking and questions and mark schemes, and unexpected codes; compare codes across data sources – documents and interviews; discuss codes with a colleague (a PhD student) and two participants (subject manager and senior marker).
4. Identifying themes	Use the codes to generate a thick description of the setting in which Biology (5008) examinations were marked in the period 2013– 2017; generate a small number of themes about quality control in the OSM environment – reflecting the multiple perspectives of subject managers, senior markers and normal markers, support the themes with evidence, compare the themes across data sources– documents and interviews, discuss the themes with a colleague, summarise the themes and send to two (subject manager and senior marker) participants for verification.
5. Data presentation	Present data narratively: that involved detailed descriptions of the themes about quality control in the OSM environment as presented by multiple perspectives of subject managers and examiners,
6. Interpretation of the findings	Articulate lessons learnt about the practice of quality control in the OSM environment, including personal interpretation from experiences, comparisons with information from literature, questions raised by the data; develop a diagrammatic framework for quality control in the OSM environment was constructed from the findings.

The idea of creating a matrix of interview responses was borrowed from Coniam (2011a:1046), who argues that, in his study, a detailed approach to tabulating responses provided a complete picture of the data by basing the findings on a matrix with comments on major areas provided by each marker. The author felt that the matrix enhanced the findings and militated against any tendency to make selective choices of illustrative quotations. In the context of this study, the response matrices were created for examiners on the WhatsApp groups and for subject managers.

In addition to reducing the biases that emanated from my background and experiences that were closely related to Biology (5008) examinations, the response matrices helped me to condense the responses into a user-friendly format. The questions posted on the WhatsApp platform were not answered in an orderly manner, the responses were mixed up. I put responses to a question together as I designed the matrix.

Thematic analysis has some advantages. It is a flexible approach that can be modified to meet the needs of many studies and provides an analysis framework for rich and yet complex data; novice researchers can easily grasp and use thematic analysis; by forcing the researcher to take a structured approach, thematic analysis is useful for summarising key elements of a large data set (Nowell et al, 2017:2). In this study I realised that I was overwhelmed by WhatsApp responses, some of them spontaneous and others irrelevant. The six-stage analysis and the response matrices helped me to work through the data, picking out information that was relevant to the practice of quality control in the OSM environment. Thematic analysis has some disadvantages. The flexibility of thematic analysis can lead to inconsistencies and lack of coherence when developing themes from the data; limited literature about thematic analysis may erode the confidence of novice researchers in the trustworthiness of thematic analysis (Nowell et al, 2017:2). Guidance was always sought from the supervisor throughout the study, in addition to reading more about thematic analysis and taking actions that enhanced the trustworthiness of the study. The next section discusses the trustworthiness of the study.

4.9 Trustworthiness

Trustworthiness relates to the quality and value of the study. Literature shows that trustworthiness can be judged by four criteria, namely, credibility, transferability, dependability and confirmability that were put forward by Lincoln and Guba way back in 1985 (Korstjens & Moser , 2018; Nowell et al, 2017; Houghton, Casey, Shaw & Murphy, 2013).

4.9.1 Credibility of findings

Credibility refers to the confidence that can be placed in the research findings to judge if the results represent the participants' views (Korstjens & Moser 2018:121). There are several methods of improving credibility and hence, trustworthiness. In this study, credibility of findings

was enhanced by triangulating the data sources, member checking and a thorough definition of the population of the study so as to shed more light on the appropriateness of the study sample (Korstjen & Moser, 2018; Asiamah et al, 2017; Nowell et al, 2017; Creswell, 2014; Houghton et al, 2013).

According to Creswell (2009:175), qualitative research studies are characterised by multiple sources of data. In the context of this study, data were collected from the subject managers by face-to-face interviews, from senior and normal markers by WhatsApp discussions and by document review. Responses were compared across data sources to enhance credibility. Interview responses were compared to data collected from documents. Responses from the senior markers were compared with responses from the normal markers and subject managers. When the responses were not supported by other sources I searched for scholarly literature that might support them. Suspicious data were discarded, such as confused and unclear statements, especially on WhatsApp.

Credibility was also be enhanced by member checking. Boyce (2006:6) suggests that, soon after the interview, the researcher should reconstruct the account and submit it to the participant for accuracy and improvement. Creswell (2014:251) discourages the use of raw interview transcriptions for member checking, but advises that researchers use polished or semi-polished products such as the final report or themes respectively. However, the interview summaries suggested by Boyce (2006:6) would help to establish accuracy of responses right at the beginning. It would also be prudent to establish the accuracy of themes from a few participants during data analysis. In this study interview accounts were reconstructed and sent back to the subject managers for accuracy and improvement (Appendix G). The reconstructed accounts were hand-delivered to the subject managers and discussed. Due to time and logistical constraints, the WhatsApp discussion write-up (Appendix H) could not be sent to examiners for verification. Categories and themes on the practice of quality control in the OSM environment were verified with senior markers who were invited for some other activities at ZIMSEC and subject managers. The member checking process did not lead to any changes in the themes.

The population and sample of the study were thoroughly defined to enhance the credulity of the study. Asiamah et al (2017:1607) argue that the population is the primary source of data and can, therefore, influence the credibility of research findings, proposing that a proper definition of the population is critical because it guides others in appraising the credibility of the sample, the sampling techniques and the outcome of the research. The population of this study was carefully defined using the SOS framework proposed by Asiamah et al (2017:1616) to improve credibility of the findings. The sampling procedures were also discussed to allow others to appraise the study.

The credibility of the documents was assured by assessing the researcher's subjectivity and the authors' motive for writing the document (Triad, 2016; Bowen, 2009). To reduce subjectivity, a document review guide was designed (Appendix D). According to Ahmed (2010:4) the authenticity should be established in cases where the documents were full of errors and were not making sense; there was no internal consistence in terms of style, tone and content; there were several versions of the same document; the document had been in the hands of persons with vested interests in it. As discussed earlier, the authenticity of minutes of meetings was verified with the listed participant. Otherwise there was no reason to doubt the authenticity of the majority of the documents that were accessed for review. Most documents were discarded because they had no content that was relevant to the practice of quality control in the OSM environment, not because they were suspicious. Ahmed (2010:4) advises researchers to ensure that documents are free from distortions; that they are prepared independently and before-hand; or that the documents are not prepared specifically for the researcher. The documents that were reviewed in this study were actually prepared for the OSM of examinations marked in the period that the technology was used, 2012-2017. I had to select documents that were relevant to the period under study, 2013-2017. There was no reason for me to suspect that anyone could create a document specifically for this study, given that OSM was suspended in 2018 and people were busy with new tasks. Transferability is another criterion for judging trustworthiness of a study.

4.9.2 Transferability

Transferability refers to the extent to which the results of qualitative research can be applied to other contexts with different participants. The researcher enables the transferability judgement by

the reader through the thick description of the behaviour and experiences of the participants as well as the context so that they become meaningful to the reader (Korstjens & Moser, 2018; Nowell et al, 2017). The results of this study may be applicable to the practice of quality control in the OSM of other subjects in the Zimbabwean contexts. However, a thick description of the context in which the OSM of O Level Biology (5008) examinations occurred in the period 2013-2017 was presented, so that readers in similar contexts can transfer the findings to their own situations. A thick description of the themes as depicted from the multiple perspectives of the subject managers, senior and normal markers was also presented. Transferability to other contexts depends on the judgement of the readers of this study.

4.9.3 Dependability and confirmability

Dependability refers to the stability of the findings over time and involves the evaluation of the findings, interpretations and recommendations of the study to ensure that they are supported by data collected from the participants (Korstjens & Moser, 2018:121). Confirmability involves taking action to ensure that the findings are the results of the participants' experiences and views, not the preferences of the researcher (Korstjens & Moser, 2018; Nowell et al, 2017). The findings of this study were presented in narratives exemplified by verbatim statements by the participants and documents. The language of the research report was edited by an expert so as to eliminate language errors that would create misunderstanding, misconceptions and misrepresentations (Appendix M).

Houghton et al (2013:14) advise that the researcher should outline decisions made throughout the research process to provide a rational basis for methodological and interpretative judgments made, arguing that this enables the readers to discern the means by which the interpretation was reached. The case study methodology was thoroughly discussed and documented as supported by the constructivist paradigm. Choices of data collection instruments were explained and justified. The data collection instruments were designed and attached as appendices. The data analysis and presentation strategies were also documented. Records of raw data were kept (Nowell, 2017:3).

Bias could emanate from the background and experiences that were relevant to the practice of quality control in the OSM environment, as declared in the axiology of the study. Constant

reflection on these experiences with the practice of quality control and the first-hand experiences of the participants helped to bring out a good interpretation of results (Korstjens & Moser, 2018; Houghton et al, 2013:15). A reflective diary was maintained, where the conceptual lens, assumptions, preconceptions and values, personal challenges experienced during the research were examined, and their effects on research decisions were explained (Korstjens & Moser, 2018; Houghton et al, 2013:15; Maramwidze-Merrison, 2010). Personal challenges experienced during the research were examined, and their effects on research decisions were explained in the relevant section of this study. An autobiographical reflection of the research experience was presented in Chapter 7.

During the face-to-face interviews, the researcher allowed the participants enough time to complete their answers before asking the next question or probing. The interviews lasted for about an hour to avoid fatigue that might impact on the quality of the data (Bolderston, 2012:72). I had planned to tape record the face-to-face interviews and later transcribe them, but I realised that the economic context at data collection did not permit tape recording and transcription. Short notes were taken to minimise distractions, chances of missing responses were reduced by write ups described earlier. The questions in face-to-face interviews and on WhatsApp were kept open to explore the participants' perspectives on the practice of quality control in the OSM. Follow-up questions and rephrasing the questions helped to get the actual perspectives of the participants (Farooq & de Villiers, 2017:22), hence eliminating personal preferences. Ethical issues that related to the conduct of this study were discussed.

4.10 Ethical considerations

Sanjari et al (2014:1) posit that researchers face ethical challenges at all stages of the study, from designing to reporting. These challenges include access and informed consent, privacy and confidentiality, the researcher's dual roles and power relations between the researcher and the participants (Sabar & Ben-Yohashua, 2017; Dongre & Sankaran, 2015; Sanjari et al, 2014). The next sub-sections account for the manner in which I positioned myself in matters of ethics.

4.10.1 Informed consent

Cohen, Manion and Morrison (2007:52) describe informed consent as one of the most important ethical consideration, where competent participants voluntarily participate in a research study, after the researcher has made sure that they are fully informed about the research projects. Sabar and Ben-Yehoshua (2017:413) argue that informed consent is viewed as an agreement between researchers and their researched population, to ensure that they have the relevant details before they agree to participate in the study. The researcher should seek official access and acceptance to the research cite and the participants, and should also not disrupt normal activities at the research cite (Cohen et al, 2007:56).

Universities and research institutions lay down principles and guidelines for conducting research in an ethically appropriate manner, requiring researchers to obtain approval from ethical committees (Sabar & Ben-Yehoshua, 2017;413). Ethical clearance (Appendix N) was therefore first applied for in the College of Education at the University of South Africa in which the study is housed, through the assistance of my supervisor. Once the ethical application was approved the necessary steps to gain access to the field from the Director of the ZIMSEC (Appendix A) and informed consent (Appendix B; Appendix C) from the participants were taken. Maramwidze-Merrison (2016:159) identified three stages of gaining access to an organisation. In the first stage the researcher seeks permission to get into the organisation to conduct the study. The second stage involves the researcher building relationships in order to gain access to the people and information. The third stage the researcher need to identify the potential participants, contact them and gain their commitment to participate in the study.

In the first stage I wrote a letter (Appendix A) to the Director of ZIMSEC to articulate the purpose of the study and to seek official access to the organisation. I visited the Director's office to discuss the study with him before delivering the letter. Once I was granted access, I visited the office of the Assistant Director Examinations Administration and the Marking Manager, who were directly involved in the marking activities to establish rapport with them. These two officials granted me the permission to access the lists and contact details (team structures) of the examiners who marked Biology (5008) examinations in 2017 as well as important documents that relate to OSM. According to Sabar and Ben-Yehoshua (2017:409), the researcher's

experiences and first-hand familiarity makes access into the field much easier. As mentioned in the axiology of this study, I was a full-time employee of ZIMSEC at the time of data collection, and was familiar with the reporting systems and business processes within the organisation. This familiarity allowed me to seek audience with the Director with ease. The first-hand experience with the OSM of examinations offered me an idea of some documents that could be requested from specific offices. I also had files of documents that I had acquired over the years as I worked in the OSM environment. I however faced some challenges in accessing the relevant documents. As discussed earlier the relevant section did not provide me with any documents until I realised that time was running out on me. I decided to use the documents that had accumulated in my work file during the OSM era and those provided by the subject manager for Biology (5008).

In the second stage of gaining access I used the team structures to determine the accessible population as describe in Section 4.6, so that I could build relationships and recruit participants, leading to the third stage of gaining participants' commitment (Maramwidze-Merrison, 2016:159). There are several methods of recruiting participants. Farooq and de Villiers (2017:13) enumerate several ways of recruiting participants: participants can be recruited by written invitations; the researcher could visit the participants personally; letters could be sent to the prospective participants and then followed up by telephone; the participants could be contacted by telephone. Farooq and de Villiers (2017:13) emphasise that the researchers should tailor their approach according to the needs of their study and the type of participants. In this study, I personally visited the subject managers in their offices to invite them to participate in the study. According to Farooq and de Villiers (2017:14), pre-interview conversations allow the researcher to address participants' concerns, create interest, build rapport and explain the interview style. The authors emphasise that the participants' interest can be aroused in three ways: the research topic has to relate to the participants' work; the participants need to know the value of their contribution; and the participants need to know how the findings will be shared. In the context of this study, these issues were addressed by the information sheet (Appendix B). The information sheet was distributed to subject managers, who were given time to read. I then followed up on them to set the date and time for the interviews. The four subject managers who were interviewed signed consent forms before participating in the interviews.

As mentioned earlier, the contact details of senior and normal markers were obtained. It was, however, not easy to access examiners due to extended power cuts that prevailed in Zimbabwe. I tried calling the numbers on the team structures but most of them were not reachable. Sending emails was not an option because the examiners could not access the internet without electricity. Worse still, they would need data bundles to access the emails and I could not afford to give them airtime at such an early stage of data collection. I only managed to give the information sheet to two senior markers who had been invited for an activity at ZIMSEC. One of them volunteered to be interviewed that very day. The senior marker signed a consent form before participating in the interview. The other one said he would prefer to be interviewed the next time he came to Harare, so I let him be. I discussed my plight with the supervisor who advised me to consider the WhatsApp platform. I then discussed the possibility with the subject manager who told me about the two groups for Biology 4025/2 and 4025/3, where the majority of examiners had marked Biology (5008) examinations on screen. I then submitted the information sheet to the subject manager for posting onto the groups.

I was added onto the two WhatsApp groups, where I sought consent from the group participants. It was not possible to gain consent from all participants on the groups, only one from each group granted consent. I therefore assumed that the examiners who did not answer the questions posted on the groups had either not marked Biology (5008) examinations or were not willing to participate in the study. Evidence of the consent is on the chats printout.

Bolderston (2012:73) concur with Sabar and Ben-Yehoshua (2017:413), who argue that consent is not a formal and singular moment in the research but is an ongoing process deriving from the ethics of care than from rights. The authors insist that the researcher should inform the participants of their right to withdraw from the research at any time; to refuse to answer certain questions; and to be informed about any harm that might be caused to them. The participants of this study were informed of their rights by the information sheet provided to them before data collection. They were informed of disruptions in their normal activities when they set aside time to participate in the interviews. To minimise disruptions in the normal activities of the subject managers and examiners, the interviews were conducted at times that were convenient to them (Bolderstone, 2012:69). I negotiated the time allocated to each interview with the subject

managers (Farooq & de Villiers, 2017:15). This implied that the duration of the interviews varied from one participant to another but did not exceed one hour. The duration of the WhatsApp group discussions was three days. This was determined by the cost of data, the need to meet the deadline for submission of the research report and the overwhelming responses that came through the platform within that short space of time. In addition to seeking access and consent the privacy of the participants was also protected.

4.10.2 Privacy of participants

Researchers need to keep participant information away from everyone except the primary research team. This is accomplished by using codes or pseudonyms in the report, so that the participants remain anonymous (Saber & Ben-Yehoshua, 2017; Creswell, 2014). In this study all the participants were assigned codes to denote their ranks as follows: Subject manager – SM; senior marker-SNR; and normal marker-NM. They were then assigned numbers from one to 33 to identify them individually, e.g. SM1, SNR7, NM11 etc. To avoid deductive disclosure, Saber and Ben-Yehoshua (2017:413) advise that the researcher should avoid detailed description of the participants. In this study there was no need to describe individual participants' characteristics because their roles were collective. Roles and responsibilities were described in detail for subject managers, senior and normal markers. Bolderston (2012:73) posits that the participants should be made aware that verbatim quotations in publications can lead to deductive disclosure, even though identifiers are removed. This implies that there could still be chances of deductive disclosure in this study and participants were informed of the possibility. The participant information sheet (Appendix B) informed them about how the data would be stored to keep them confidential. The role of the researcher in qualitative research raises ethical issues that need to be addressed.

4.10.3 The role of the researcher

As declared in the axiology of the study, I was a full-time employee of the ZIMSEC at the time of data collection. Power imbalances (Saber & Ben-Yehoshua, 2017; Creswell, 2014) were likely to arise during the discussions with the examiners. To put the participants at ease, the purpose of the study was explained to the examiners before collecting data. Leading questions were avoided; personal impressions were not shared; only questions on the interview schedule were

asked and probed (Creswell, 2014; Bolderston, 2012). As discussed earlier, a reflective diary was maintained, where the conceptual lens, assumptions, preconceptions and values, personal challenges experienced during the research were examined, and their effects on research decisions were explained (Korstjens & Moser, 2018; Houghton et al, 2013:15; Maramwidze-Merrison, 2010).

The WhatsApp platform greatly reduced the influence of power imbalance between me and the examiners. The platform offered the examiners some privacy by allowing use of usernames that were not real names. They could choose not to respond to the questions without feeling bad about it because of the distance. To reduce bias emanating from the researcher's role, steps were taken to ensure the trustworthiness of the study as discussed in Section 4.9. The next section summarises and concludes the chapter.

4.11 Conclusion

This chapter gave an overview of research paradigms and selected the constructivist paradigm as the philosophy that guided this study. The ontology, epistemology, methodology and axiology of this study were articulated in line with the constructivist paradigm. The chapter highlighted that there are three main paradigms that guide educational research, namely positivism, constructivism and pragmatism. The three paradigms lead to quantitative, qualitative and mixed methods respectively. As guided by the constructivist paradigm, the ontology of this research assumed that examiners and subject managers hold and share multiple realities about quality control in the OSM environment, created during the marking of O Level Biology (5008) examinations. This study sought to explore how they interpret OSM quality control, the factors that influence their interpretations and how their interpretations of quality control varied with experiences, time and context. This ontology led to the transactional epistemology and the qualitative instrumental case study methodology that guided this study.

The history of case study research was traced so as to position this study in the qualitative domain, in line with the constructivist paradigm. The study adopted the instrumental case study, where O Level Biology was studied to illuminate the practice of quality control in the OSM environment. It was established that researchers need to define and bind the case to clarify the

boundaries of the study. This study focused on the OSM of O Level Biology (5008) examinations from 2013 to 2017.

The population of the study was carefully defined using the SOS framework to improve the credibility of the findings. Purposive sampling strategies were used to select the participants (comprising of subject managers, normal and senior markers who marked Biology examinations in 2017). A total of 33 participants agreed to take part in the study. Documents were also selected by purposive sampling for review. Data were collected at a time when the Zimbabwean economy was deteriorating. Interviews and document analysis were chosen as the most appropriate methods and justified. Subject managers were interviewed face-to-face, senior and normal markers were interviewed by focus group discussions on the WhatsApp platform. Data collection instruments were designed and attached to the report as appendices.

Data were analysed by a six-stage thematic analysis adapted from Creswell (2013:245) and Nowell et al (2017:4) and presented in the form of a narrative involving thick description of the context in which the OSM of Biology examinations occurred in the period 2013-2017. The themes were described in detail to depict the multiple perspectives of the participants. The role of the researcher in qualitative research raises issues of ethics and trustworthiness. Trustworthiness was assured by a thorough definition of the population and sample, detailed description and justification of data collection methods, triangulation of data sources and member checking. Ethical issues that related to the conduct of the study were addressed to reduce bias that was likely to emanate from power imbalances between the participants and myself. Relevant ethical documents were designed and attached at the end of the report as appendices. An ethical clearance application was obtained from the UNISA College of Education. The access permissions were sought from the director and the participants. Despite the challenges encountered data was successfully collected to answer the research questions.

The next chapter presents the findings of the study.

Chapter 5

Data presentation and analysis

5.1 Introduction

The purpose of this study was to explore the practice of quality control in the marking of Ordinary Level Biology in the OSM environment in Zimbabwe in order to propose a framework which can help to improve the practice. This purpose is in tandem with the research question and sub-questions presented in Chapter 1, Section 4 and Chapter 4 Section 4.4. Data is presented in themes that emerged from the research sub-questions. The themes include building the capacity of examiners, monitoring the quality of marking, test design issues, opportunities and challenges of quality control in the OSM environment.

5.2 Data processing and presentation

In accordance with the methods in Chapter 4, data were collected through document review and face-to-face and focus group interviews on the WhatsApp platform. Tables 4.2 and 4.3 summarise the interview participants and documents that were reviewed respectively.

Two subject managers (SMs) were interviewed face-to-face using the interview guide. The other two SMs were interviewed face-to-face and by telephone and they mainly responded to the follow-up questions about the issues raised in the document analysis, face-to-face interviews and the focus group discussions, although the first two SMs also responded to a few follow up questions. The one senior marker (SNR) who was interviewed face-to-face also participated in the focus group discussions on the WhatsApp platform. Twenty-nine examiners, made up of 11 SNRs and 18 NMs participated on the two WhatsApp group discussions.

A total number of 33 participants shared their experiences on the practice of quality control in the OSM of Biology (5008) examinations. The participants were assigned numbers from 1 to 33 and their ranks were assigned codes as follows: Subject manager – SM; senior marker – SNR;

and normal marker – NM. The participants were identified by a combination of the code and the number in the narrative, e.g. SM1, SNR9, NM11, etc.

As discussed in Chapter 4 Section 7.1, three categories of documents were analysed, namely public records, personal documents and physical evidence. The specific documents in each category were presented in Table 4.3. A total of 55 documents were analysed and the findings were written up by research sub-question (Appendix K) to come up with themes.

The data collected from the document review and interviews provided information about the context in which O Level Biology (5008) examinations were marked on screen between 2013 and 2017, as described in the next section.

5.3 The context of OSM in Zimbabwe

The information about the context of the OSM of Biology (5008) examinations was provided mainly by the documents that were reviewed and confirmed by the interview participants. The data gathered provided insights into the framework that guided the assessment of O level Biology, the technological infrastructure and human resource capacity provided for the OSM. The next sub-section describes the assessment framework for O Level Biology.

5.3.1 The assessment framework for Biology (5008)

The O level Biology (5008) syllabus provided the framework that guided the design and marking of Biology examinations in Zimbabwe. The practice of quality control in the OSM environment was explored in the context of the aims, assessment objectives and the assessment scheme in the syllabus.

5.3.1.1 Aims and objectives of the syllabus

The syllabus enumerated the aims of the O Level Biology course thus:

The aims of the syllabus are to help learners to:

1. develop interest and curiosity in science;
2. develop concepts and skills that are relevant to the study and practice of biology;

3. appreciate and enjoy biology and its methods of enquiry;
4. develop creativity, initiative and skills of enquiry;
5. develop good practices for health and safety;
6. develop accuracy and precision, objectivity and integrity;
7. recognise the usefulness and limitations of science;
8. apply the scientific method in other disciplines and in everyday life;
9. appreciate the beneficial and detrimental effects of the applications of science;
10. recognise that the study and practice of science are inter-related and are subject to economic, technological, social, political, ethical and cultural influences
(O Level Biology Syllabus: 5008:2).

The aims were translated into the three categories of objectives, namely, knowledge and understanding; handling information and solving problems; and experimental skills. The skills that the candidates were expected to demonstrate in each assessment category were specified. Examples of the assessment objectives in each category are indicated in Table 5.1.

Table 5.1: Assessment objectives and skills for O Level Biology (5008)

No.	Assessment objective	Examples of skills
1.0	Knowledge and understanding	Pupils should be able to demonstrate knowledge and understanding of: scientific instruments and apparatus, techniques and operation and aspects of safety; biological units, terminology, symbols and conventions; scientific quantities and how they are determined; etc
2.0	Handling information and solving problems	Pupils should be able to demonstrate, in familiar and unfamiliar situations, their ability to: Extract information relevant to a particular context from data presented in diagrammatic, symbolic, graphical,

No.	Assessment objective	Examples of skills
		numerical or verbal form; use data to recognise patterns, formulate hypotheses and draw conclusions; translate information from one form to another; communicate logically and concisely; etc
3.0	Experimental skills	Pupils should be able to: Follow instructions for practical work; plan, organise and carry out experimental investigations; select appropriate apparatus and materials for experimental work; use apparatus and materials effectively and safely; etc

A closer look at the assessment objectives suggests that the first objective required the candidates to recall basic biological facts and phenomena; the second objective required them to apply biological concepts in solving problems; and the third objective required them to conduct experiments as they investigated biological phenomena. The questions set for the examinations should have assessed the skills regardless of the marking mode. The objectives were weighted across the examination papers as indicated in Table 5.2.

Table 5.2: Weighting of assessment objectives

Objective	Paper	Weighting
1.0. Knowledge and understanding	1 & 2	55%
2.0. Handling information and solving problems	1 & 2	45%
3.0. Experimental skills	3 & 4	100%

Table 5.2 shows that the first two objectives were examined in Paper 1 and Paper 2 while the experimental skills were examined in Paper 3 and Paper 4. The syllabus introduction stated that less emphasis should be placed on the factual recall of material and more emphasis should be placed on the understanding and application of the scientific concepts, principles and skills. This

emphasis was therefore expected to reflect in the weighting of the skills in question papers. The weighting of the skills, however, placed more emphasis on the factual recall of the material as indicated in Table 5.2. This contradiction could have caused variations in the tests that were designed for Biology (5008) examinations. When asked about the contradiction, SM3 said that *I knew about the contradiction. I preferred to use the objective weighting indicated in the syllabus.* SM1 added that the contradictions were carried forward to the syllabi in the new curriculum, and thus stated: *There were many errors in the old curriculum syllabi, unfortunately the errors were passed on to the new syllabi.*

Despite the contradiction in the emphasis, the O Level Biology (5008) syllabus provided the criteria for assessing the course. The examinations were supposed to be set and marked as guided by the syllabus. The syllabus also prescribed the examination papers for the course in a detailed assessment scheme.

The next section describes the assessment scheme for the Biology (5008) course.

5.3.1.2 The assessment scheme

The syllabus presented an assessment scheme that prescribed four papers. An analysis of the question papers showed that they were identified by codes (Biology 5008 question papers, 2013–2017). The assessment scheme is summarised in Table 5.3.

Table 5.3: Assessment scheme for Biology (5008)

Paper	Code	Paper type	Duration	Marks	Paper weighting
1.	5008/1	Theory: 40 compulsory multiple choice questions	1 hr	40	30%
2	5008/2	Theory: Section A: (40 marks) a number of compulsory short answer and structured questions of variable mark value.	2 hrs	100	50%

Paper	Code	Paper type	Duration	Marks	Paper weighting
		Section B: (60 marks) of five free-response questions of twenty marks each; candidates choose 3.			
3	5008/3	Practical examination: two questions worth 20 marks each.	1 hr 30 min	40	20%
4	5008/4	Alternative to practical: written paper of four compulsory short-answer and structured questions designed to test familiarity with practical laboratory procedures.	1 hr	40	20%

The candidates were required to sit for Papers 1 and 2 and either Paper 3 or Paper 4. An analysis of the assessment scheme shows that Paper 2 was the main component of the examination since it contributed 50%. The syllabus went on to specify the content of the O Level Biology course. Twelve topics were to be covered in two years (Form 3 and Form 4). Each topic was broken down into learning objectives, content and activities with guiding notes (O Level Biology syllabus 5008:8-31).

Document review showed that Biology examinations, 5008/3 and 5008/4 were marked on screen for the first time in the November 2013 examination (Examination Circulars Number 41 and 42 of 2013; Lessons learnt report, November 2013). Another paper, 5008/2 was migrated to the OSM platform in the November 2015 examination (Examination Circular Number 8 of 2015; email, 25 February 2015; Lessons learnt report, June 2015). The examiners raised concerns that they could not compare the candidates' responses to the supervisor's report because the scripts were segmented into individual items and were anonymous (email, 8 May 2013). The examiners therefore requested that the practical examinations (5008/3) be marked on paper to avoid prejudicing the candidates (personal experience; Lessons learnt report, November 2013).

The SMs confirmed that papers 5008/2 and 5008/4 were marked on screen, adding that 5008/3 was marked on screen once in one session only. They stated in this regard:

The practical papers could not be e-marked because the candidates' responses were compared to those on the supervisor's report (SM1).

Practical examinations for O Level Physics, Chemistry and Biology were marked on screen once and had to be returned to manual marking (SM4).

The examiners on the WhatsApp platform were asked to indicate the papers they marked on screen and they indicated 5008/2 and 5008/4. These were indicated as follows:

2 and 4 (SNR5).

Paper 2 (NM11).

I marked papers 2 and 4 in 2016 (SNR10).

The two papers, 5008/2 and 5008/4 were marked for long enough to provide data about the practice of quality control. Most of the findings presented in this chapter relate to the practice of quality control in the marking of the two papers as guided by the assessment framework presented in Sub-section 5.3.1.

The next section presents findings on the human and material resources that the ZIMSEC provided for the OSM of Biology (5008) examinations.

5.3.2 Technological infrastructure

On this aspect, review of the June/November 2014 lessons learnt report showed that the ZIMSEC established a scan centre (bureau) at head office where scripts were scanned for OSM in every examination session. It was apparent that scanners were bought from the OSM software provider (SWP). Four new scanners were not shipped from the UK for the November 2013 examinations. There were sufficient scanners and computers for the June 2014 and June 2015 examinations

(lessons learnt report June/November 2014) probably because the examinations had fewer candidates compared to the November examinations (personal experience). There was enough evidence to suggest that the scanning process for every examination was slowed down by numerous challenges. It was reported that the equipment was not up to date (antivirus and updates) for both June and November 2014 sessions (Lessons learnt report, June/November 2014). Three scanners reportedly broke down in the June 2017 examination session forcing the staff to work day and night in two shifts (Minutes of meeting, 30 June 2017). A total of 16 scanners were required to achieve the scanning task on time in November 2017 but the ZIMSEC provided a maximum of 13. There were no spare parts for the scanners. One new scanner was found to be faulty upon receipt (Lessons learnt report, June/November 2017).

There was a need for two additional guillotines but they were not procured for the November 2013 examination. The scanning was delayed when the single guillotine broke down (Lessons learnt report, November 2013). The two guillotine machines were bought for the June and November 2014 examinations. However, the old one was not serviced. The SWP encouraged ZIMSEC to have a clear plan for guillotine usage and to check if there was need for spares or servicing prior to the exam session (Lessons learnt reports, June/November 2014). The ZIMSEC, however, allowed the guillotines to break down again. A review of the minutes of meetings indicated that there was only one guillotine machine in the scan centre. The supervisor pleaded for the purchase of a bigger guillotine in the two meetings. The directorate informed the meetings that an industrial guillotine was in the budget and would be purchased (minutes of meetings, 3 July 2015; 2 December 2016). This was evidence that the ZIMSEC did not maintain the guillotine machines. The scripts were scanned and saved onto the script marker.

Document review (email, 18 May 2014; Administration routines; 2010) further indicated that the OSM software was made of two components, the script marker and the e-marker. The script marker was used to capture script details by centre and candidate number during scanning. The scanned script images were saved in the script marker, segmented into portions and exported to the e-marker where examiners could access and mark them (email, 18 May 2014). The segmentation process was done by a third party in India, creating challenges for the ZIMSEC (email, 18 May 2014). It was evident that the script marker had three modules, the administrator,

the senior marker and the marker modules (Administration routines, 2010; Senior Marker Quick Reference Guide, n.d).

The ZIMSEC had no servers of its own and it had to borrow from the SWP. An email to the OSM project manager by the SWP indicated that the latter had loaned servers to the ZIMSEC in previous examinations and hoped to do the same in November 2013. The ZIMSEC had indicated that it would want to buy its own servers for future marking sessions but would rather not use the new servers in a November examination in case there were problems. The servers were, however, not cleared by the UK customs, creating a major challenge for the November 2013 examination. It was also indicated that the SWP was testing software delivery by cloud. The ZIMSEC indicated that it perceived the political and security challenges on the cloud, so it would rather continue using servers (email, 27 August 2013). Despite the ZIMSEC's reservations about the cloud, the OSM software was eventually delivered on cloud (Lessons learnt report, June/November 2017). The ZIMSEC made arrangements for the internet connections between Harare and the marking venues which were hired (Lessons learnt report, November 2014).

The ZIMSEC hired computers for OSM from other institutions as evidenced in several documents. An internal memo from the marking administration manager (MAM)) indicated that the ZIMSEC had requested a local university to provide 400 computers for marking one subject. The computers had to be networked and with internet access. The university had not responded by the date of the memo (Memo, 11 April 2013). The examinations were, therefore, marked from a central venue using the hired computers. The institutions could not provide enough computers, forcing the ZIMSEC to group the subjects into sessions. A communication between the ZIMSEC staff and the SWP (email, 12 September 2013) indicated that the November 2013 examinations were marked in three sessions in two universities. The November 2014 examinations were also marked in three sessions, which spilled into January 2015, hosted by two universities (E-marking programme, 23 October 2014). The ZIMSEC lost two weeks of marking time when one university hosted a political party congress in December 2013 (Lessons learnt report, November 2013), implying that the Council could not access the venues on time.

The next section presents the findings on the human resource capacity for OSM.

5.3.3 Human resource capacity

Several departments within the ZIMSEC collaborated to implement the OSM of examinations with the major roles played by the Examiner Records (ER), the Test Development, Research and Evaluation (TDR & E – mainly SMs and the question paper development manager [QPDM]), and Information Systems (IS) departments (Marking administration schedule, Nov-Dec 2014). A close analysis of work schedule and some other documents (Item Designation Summary [IDS] 5008/2, June 2016; Marking Administration Schedule, Nov-Dec 2014; OSM Work Schedule, November 2014; SWP emails, 8 April 2013, 27 August 2013, 22 October 2013, 17 December 2013; Monitoring Marking Phase Walk Through, n.d) showed that the quality control activities began before the actual marking of scripts and continued until all the scripts had been marked. The activities were shared between the ZIMSEC and the SWP as indicated in Table 5.4.

Table 5.4: Quality control activities in the OSM environment

Activity	By who	Purpose of activity	When
Creation and submission of IDS	SMs & QPDM	Define mark scheme identification and validation parameters	Before marking
Creation of the mark scheme	SWP		
Check mark schemes for accuracy	SMs		
Provide soft copies of blank scripts to SWP	QPDM & IS	To create script identification and validation parameters	Before marking
Provide information about components, centres and candidates	ER & IS	Link the scripts to the examination centres and the	Before marking

Activity	By who	Purpose of activity	When
(3Cs) to the SWP		candidates	
Loading of 3Cs data onto the system	SWP		
Scanning scripts	ER, IS & SWP	To generate script images for marking by examiners	Before marking
Standardisation meetings	SMs & examiners	To enhance mastery and consistent application of the mark scheme	Before marking
Uploading the standardised mark scheme	IS	Link the scripts and the mark scheme	Before marking
Computer Training examiners	SWP & SMs	Familiarise the examiners with the OSM.	Before marking
Setting quality control parameters	SMs	Monitoring the quality of marking	Before marking
Set seed	Senior markers	Set marking standards	Before marking
Monitoring quality of marking			
Amending quality control parameters	SMs	Adjust the seed parameters appropriately	During marking
Retiring seeds	SMs	Delete marks from wrong seeds and returning them to the marking queue	During marking

Activity	By who	Purpose of activity	When
Deletion of marks	SMs	Delete marks awarded by poor markers and returning scripts to the marking queue	During marking
Invoking marking quality review	SMs	Review quality of marking real-time or later	During and after marking

The scripts were scanned by the temporary staff supervised by permanent ZIMSEC officers in the ER department. The scan centre manager expressed concern over the high turnover of the temporary staff where new people were recruited in every examination. The supervisor argued that new people needed to be trained, wasting valuable time for scanning (Bureau Supervisor's report, 17 November 2014; minutes of meetings, 3 July 2015; 2 December 2016; 30 June 2017). The same concern was also raised by the SWP who warned that the extended scanning periods would delay the marking as well (Lessons Learnt Reports, June 2013; June/November 2014; June/November 2017. There was no apparent effort to address the concern as evidenced by the fact that it was raised from 2014 right up to 2017 when the last examinations were marked on screen.

The marking was done by the examiners who were supervised by the SMs. The document review showed that the marking exercise was conducted by the examiners who were divided into categories, i.e. senior and normal markers (Marking Administration Schedule, Nov-Dec 2014; E-marking Programme, November 2014; E-marking Training Programme, December 2012). In addition to supervising the normal markers, the senior markers also marked the scripts (Quality of Marking Reports: 5008/2 & 5008/4 November 2016; Senior Marker Quick Reference Guide n.d.). On the OSM platform, the SMs were called administrators and they set the marking and quality control parameters; supervised the senior markers; and monitored the progress of the marking exercise (Marking Administration Schedule, Nov-Dec 2014; Administration Routines, 2010; Commence Marking Phase, n.d.; Marking Monitoring Walk Through, n.d.). The quality of

marking would therefore be determined by the capacity of the SMs to work in the OSM environment.

It was evident from the document review that there were capacity building initiatives by the SWP. In 2012 and 2013 ZIMSEC personnel were trained as administrators and examiner trainers (E-marking Training Programme, December 2012; Emails, 6-8 December 2013). In 2014, a training programme was designed for the ZIMSEC staff in the scan centre. The supervisory teams were trained to train others; information technology technicians were trained to manage the technological infrastructure; engineers were trained to set up and service the scanners and to run pre-live tests. The training documents were updated and distributed to the relevant people in the scan centre (Bureau Training Programme, 29 October 2014). In the same year another training programme was launched for the administrators, senior and normal markers. The training materials that included computer-based training for the senior markers were prepared and loaded onto the system. The senior markers and normal markers were trained (ZIMSEC Work Package, October 2014; Marking Administration Schedule, Nov-Dec 2014). The training initiatives launched before June 2014 were, however, rated as superficial by the ZIMSEC.

There was evidence that the ZIMSEC was disgruntled that the SWP retained the control of the OSM software and were doing most of the key activities. In a communication to the SWP in 2014, soon after the review meeting of the November 2013 marking session, the marking administration manager (MAM) raised a concern about the SWP's control of the OSM technology (email, 18 May 2013). The MAM wrote thus in the email:

The Council feels that it has gained superficial knowledge of the system and most work is done by SWP behind the scenes. SWP is not keen to transfer some functions of the technology, like mark scheme creation, data uploads and all work that is done after you receive 3Cs data, blanks, marking guidance, handling and uploading of examiner files. The feeling is that ZIMSEC has to be involved in these processes so that we are able to trace back all processes and be able to trouble-shoot rather than depend on SWP for all and sundry. For that reason ZIMSEC wishes to see a systematic involvement of our IT staff in the use of these functions with a view to cede that responsibility to ZIMSEC as a

cost saving and capacity building measure.....If SWP does not give ZIMSEC reasonable ownership rights, then it means SWP will do consultancy work for ZIMSEC forever and that is expensive and it defeats the Council's idea of a cost effective option.

This statement implies that the ZIMSEC was paying consultancy fees to the SWP and that seemed unsustainable unless there was meaningful skills transfer. The MAM outlined several challenges that bedeviled the OSM of the November 2013 examinations and in conclusion called for a refocus of the training efforts. The conclusion read thus:

It is requested that training that shall obtain hence forth focus more on detail and back office skills in order to equip Bureau Operators and Administrators with diagnostic skills to be able to deal with problems on-site without having to call or email SWP every time there is a problem.

The communication provided evidence that the ZIMSEC adopted the OSM technology to enhance the efficiency and quality of examinations at minimum cost and expected some degree of autonomy in the use of the technology. The SWP seemed to have ignored the communication and continued to receive and upload the data about the components, centres and candidates (3Cs) and create mark schemes in the system (Lessons Learnt report, June 2015). The ZIMSEC officers were not trained to clip and load mark schemes onto the software until after the November 2014 examination. The SWP loaded the mark schemes and discussed the required changes with individual SMs as usual (Lessons Learnt Report, June/November 2014). It was therefore apparent that the SWP were determined to do consultancy work and provided on-site support to the ZIMSEC.

The interview responses, however, indicated that ZIMSEC information technology technicians were later trained to clip and upload the mark schemes. The SWP sent the engineers and other personnel to provide onsite support to ZIMSEC. A team of seven people came to provide support for the November 2014 examinations. It was apparent that the ZIMSEC bore the cost of the onsite support in terms of flights and accommodation (Email, 1 October 2014; Lessons Learnt Reports, November 2013), providing further evidence that the conduct of the SWP increased the

cost of running the OSM technology. This could be the reason why ER department opted to run the June 2015 examination without onsite support from the SWP.

Summarising the minutes of a teleconference for the June 2015 examinations, a lady representing the SWP wrote thus in an email:

Bureau – ZimSEC have agreed to run the June series for 2015 with no SWP onsite support. They feel that they have the sufficient skill sets to run the series with only remote support from SWP-Fantastic news (email, 25 February 2015).

The SMs however gained requisite skills to carry out the quality control activities with minimum support from the SWP. The statements in the Lessons Learnt Reports (June/November 2017) read “...minimal support was required for SMs and administration teams during marking....” The quality control activities were planned.

There was evidence of planning for each examination, with the ZIMSEC working closely with the SWP (Email, 27 August 2013; Memo, 11 April 2013). Teleconferences were held for the planning purposes and lessons learnt reports were compiled to evaluate the exercise (Lessons Learnt Reports, June/November 2017; June 2015; June/November 2014; June 2013; email, 25 February 2015). Meetings were held to commission the marking exercises (Minutes of Meetings, 30 June 2017; 2 December 2016; 3 July 2015). A review meeting was held for the November 2013 examination session with the purpose to critically evaluate the marking exercise in order to inform the planning and implementation of subsequent sessions (2013 post-mortem meeting programme, 14 May 2014). However, the SWP raised concerns that the ZIMSEC withheld some information from them.

The ZIMSEC had indicated to the SWP that 5008/4 and other subjects would be migrated to OSM in the June 2013 examination. Probably fearing for the costs, the ZIMSEC did not include the components on the June examinations as agreed and did not communicate the new position to the SWP. In earlier examinations, the servers were at the ZIMSEC Head Office, but for some reason, the ZIMSEC moved them to the marking venue without informing the SWP, who

complained that the servers were reconfigured and could not be accessed from the UK. The SWP pleaded with the ZIMSEC to communicate changes to agreed activities and timelines (Lessons Learnt Report, November 2013). This was evidence that the relationship between the SWP and the ZIMSEC was not healthy for the adoption and use of the OSM technology to enhance the efficiency and quality of examinations. The unhealthy relationship could have worsened the challenges that disrupted almost all marking exercises as indicated in the lessons learnt reports.

Despite the planning, it was apparent that the ZIMSEC always seemed not prepared for the OSM exercises as evidenced in reports. In 2014, the SWP raised concerns that scanning was delayed because the working space was not immediately available; scanners were not available for inspection by the visiting engineer who had to extend his stay in Zimbabwe; scanning was interrupted by other activities, with the SWP pleading with the ZIMSEC to allow the staff to focus on the scanning process until it was complete (E-marking Project Handover Report, 13 November 2014). The same challenges of guillotines not serviced, non-procurement of spare parts and inadequate scanners, high turnover of the temporary staff and power cuts were highlighted in reports and meetings, with no apparent action being taken (Lessons learnt report, June/November 2017; June/November 2014; Minutes of Meetings, 30 June 2017 & 2 December 2016; E-marking Project Handover Report, 13 November 2014). The 3Cs data were submitted late in some examination sessions (Lessons Learnt Report, June/November 2017; November 2013; Emails, 8-16 April 2014).

The decisions about important issues and milestone dates took too long to be made. There was a need to barcode the scripts for easy script identification and validation at scanning but that was never done (Lessons Learnt Reports, November 2013, June/November 2017). In 2013, the ER office wrote to the TDR & E, seeking a policy position on inline additional pages (ILAP) for constrained components and answer booklets for an unconstrained component for the June 2013 examination (Memo, 11 April 2013). It was evident from the memo that the SWP had made several enquiries about the same issues before the date of the memo. The SWP made another enquiry about the same issues in May (Email, 5 May 2013). The ER office had to write another follow up email to TDR & E (Email, 27 May 2013). A decision was probably never made, as evidenced by the June 2013 lessons learnt report where the SWP made suggestions for dealing

with the challenges that emanated from the ILAP and the unconstrained answer booklets. The challenges are presented in the findings of the third research sub-question later.

It was apparent that the work schedules were either not approved on time or not followed. The ER office circulated a work schedule for the November 2014 examination session with OSM activities starting on 1 October. The key activities were outlined by start and finish dates, some of which were marked 'TBC' (to be confirmed), implying that decisions about the dates had not been made. It was evident that the work schedule had been designed earlier and had to be updated (OSM Work Schedule, November 2014). The scanning of Biology 5008/4 scripts (with a candidature of 22058) was subsequently delayed, as indicated in a report (OSM project handover report, 13 November 2014). None of the scripts had been scanned by the date of the report. This could have led to the scanning challenges that were reported in the lessons learnt report which highlighted that the marking of Biology 5008/4 was interrupted because of script locations that were empty/invalid (Lessons Learnt Report, June/November 2014). The ER office had to retract the marking programme that they had circulated for the same examination session because they had been advised to reduce the marking period for some groups of subjects (Memo, 23 October 2014). This meant that the relevant people had to reconvene, design another programme and start seeking approvals again.

The TDR & E division was evidently headed by different people, most of them in acting capacities most of the times, as indicated by communications addressed to different people and meetings chaired by different people (Minutes of Meetings, 30 June 2017; 2 December 2016; 3 July 2015; Memo, 11 April 2013; Email, 5 May 2013; Memo, 17 October 2013). This could be reason why decisions took too long to be made, compromising the practice of quality control in the OSM environment.

The practice of quality control was therefore influenced by the context that included the assessment framework in the Biology (5008) syllabus; the technological infrastructure that included a scan centre, hired computers and marking venues; planning meetings and teleconferences; review meetings and lessons learnt reports; and challenges relating to the context as explained above.

The next section presents findings on the first research sub-question.

5.4 Capacity building of examiners

Sub-question 1: How does training of examiners and standardisation activities influence the quality of marking O level Biology in the onscreen marking environment?

To enhance the quality of marking Biology (5008) examinations, the ZIMSEC recruited and trained the examiners and organised standardisation meetings where the mark schemes were standardised. The next section presents findings on the recruitment and training of examiners.

5.4.1 Examiner recruitment and training

Some examiners were recruited and trained for PBM while others were recruited and trained during the OSM period. There was evidence that the ZIMSEC advertised for the position of examiners in the newspapers, indicating a transparent recruitment procedure. The transparent procedure could have enhanced the recruitment of the right calibre of examiners for the Biology (5008) examinations. Migration of examiners from PBM to OSM could be an indication that marking competences were independent of the marking mode. The evidence was provided by examiner responses.

Examiners on the WhatsApp platform indicated that they responded to newspaper adverts and applied to be selected and trained for marking. The examiners responded thus:

I applied after seeing a newspaper advert for markers (NM13).

I was selected through manual marking, and then trained later for e-marking (NM11).

I applied for marking during e-marking (NM12).

I think all examiners who were marking Paper 4 manually were trained for e-marking at CUT (NM27).

It was evident that the examiners who responded to the adverts were shortlisted and invited for training. The trainee examiners sat for a test and then marked dummy scripts during training, as indicated by a normal marker who said *we also wrote a test. O level paper* (NM13). Some examiners were not sure if the test was used for selection while others said it was used. They stated in this regard:

Yes, the test was used as well as proper marking of more than 15 dummies (NM11).

Not sure if the test was used for selection but it was part of the training (NM13).

This could be an indication that the purpose of the test was not emphasised to the trainees. SM3, however, confirmed that the trainees set for a test before they were trained to mark; *we gave them a paper where they would answer either the whole or part of it. This is done to check content and competence in the subject*. Probed on how the test was used the SM3 answered thus:

The trainers would mark the test. After the training, the trainees were given dummies to mark under test conditions. The test scores and the dummies were used to assess the trainees. They were awarded grades A to D. A – leadership material; B Independent marker; C – requires strict supervision; and D – dismissal/demotion.

The procedure outlined by SM3 shows a well organised training that could equip the examiners with the requisite skills that could enhance the quality of marking. The assessment and awarding of grades provided criteria for the selection of examiners into the relevant roles and made sure that poor markers were not selected as examiners. The marking of dummies also enhanced examiner competences and, hence quality of marking, as confirmed by the examiners who were probed on the dummies, what they were and how they influenced the quality of marking. A SNR explained that dummies were photocopies of real scripts that were selected from sample candidates by the chief examiner (principal marking supervisor – PMS) of the paper, emphasising that dummies represent a range of candidates' responses, explaining thus:

It is envisaged that all ranges of candidates' performance are catered for, i.e. good, mediocre, as well as poor. Dummies play a major role in standardising the mark scheme so that it is in unison with the real situation on the group, in terms of how candidates perform. Examiners, for the first time, are familiarised with the marking scheme as well as getting standardised in terms of how they approach candidate answers of various abilities (SNR33).

Normal markers expressed their own perspectives and experiences with dummies, and they had this to say:

Dummies are selected examination scripts from various centres. I am not sure who prepares them. Their role is to enable consistency in marking. It is the dummies that all e-markers and the senior supervisors discuss and explore all possible responses to an item before live marking begins. They basically show the good responses and the general misconceptions made by candidates to items (NM11).

Dummies are scripts that are marked by each marker soon after discussing the mark scheme. Their purpose is to ensure that all markers have almost the same consistency when marking live scripts, and also discuss possible marking points that sound correct while they are wrong (NM12).

Dummies are selected scripts from various centres. They are prepared by ZIMSEC officials appointed to do so at any given time. This can be the principal supervisor and his team or the SM... the dummies are there to brainstorm the marker so as to identify more marking points before the start of live marking. They promote critical thinking in the evaluation of the mark scheme to ensure validity (NM17).

The examiners concurred that dummies were practice scripts that promoted consistent application of the mark scheme. The above verbatim responses indicate that, through the dummies, the examiners were trained to mark scripts for a wide range of candidates; poor to very good, and from different contexts, as indicated by a variety of centres. The dummies therefore ensured consistency of marking across contexts and candidates' competences, enhancing quality

of marking. The examiners indicated that the first training was specific to marking in general and not to OSM or PBM, evidenced in their verbatim responses as follows:

Training was mainly marking of dummies, no computers (NM11).

We were trained for marking in general at Belvedere but when we went to Chinhoyi University for e-marking we were trained on how to use the computer and marking using the ZIMSEC portal (MN14).

We were trained for marking in general, and then when e-marking was introduced we were trained for e-marking (NM15).

Marking in general, later e-marking (NM16).

The responses confirm that PBM markers were migrated to OSM, providing more evidence that marking competences were independent of marking mode, according to the participants. Asked to comment on the relevance of the training to OSM, the examiners said the training was relevant:

The training was relevant because the main aspect of marking rests on adherence to the mark scheme, which was mainly done manually, then e-marking was mainly the marker's ability to use the computer effectively (NM1).

Use of dummies for training addressed the issue of quality and uniformity in marking (NM12).

These responses emphasise the importance of dummies in enhancing the mastery and consistent application of the mark schemes across contexts and candidates' competences. They also provide more evidence that marking competence was independent of the marking mode. The examiners thought that the training that they received enabled them to mark examinations on paper and on

screen. The examiners were then specifically trained to mark onscreen during live marking to familiarise with the computer mode, as indicated by SNR10 who had this to say:

We were trained for e-marking at a coordination (standardisation) meeting. We went through pre-live training that did not record marks. Pre-live training offered us computer training and familiarisation with e-marker icons.

Some examiners felt the OSM training was not adequate, expressing feelings that suggest the OSM training was rushed: *it was difficult due to the fact that we were trained through standardisation rather than being given separate time for training (NM17); there was inadequate orientation/training of those who were to set seeds (SNR10).*

The allusion to inadequate training to set seeds, coming from a SNR, implies that the rushed training compromised the quality control mechanism. The examiner NM17 might have a point when they suggested a separate training session for OSM. A separate training would give the NMs and SNRs adequate time to train in the OSM mode without rushing to meet the deadlines of marking live examinations. The document review (E-marking Programme, 2014; Marking Administration Schedule, Nov-Dec 2014) confirmed that the normal markers went through the computer-based training and pre-live marking soon after the standardisation meeting. The pre-live training seemed to enhance the examiners' capacity to mark onscreen by familiarising with the icons and the scripts. Probed on pre-live marking, the SMs responded thus:

Examiners work in the training mode before they mark live scripts. All examiners (senior and normal markers go through pre-live marking. The examiners mark scripts for their subject or for any other subject (SM1).

Practice by pre-live training enhances quality of marking. There are no seeds in the training mode so there is plenty opportunity to practise and familiarise with candidates' responses (SM2).

The pre-live marking is in the training mode. The examiners familiarise with the e-marking icons and the scripts. If there are any segmentation errors on the scripts the examiners pick them and they are corrected before live marking begins. There are no seeds in the training mode, and the marks are not recorded (SM4).

The SM1's response indicated that the pre-live marking could be practiced using scripts for any subject. This was conformed in training material for markers and work schedules where the subjects were given pseudonyms MK1, MK2 etc (OSM work schedule, November 2014; Marker quick reference guide, n.d.). It could, therefore, have been prudent to schedule a separate OSM training, as suggested by NM17, using the pseudo scripts and then use real scripts for pre-live marking at the live session. Seeds could also be included in the training mode to familiarise the markers with the quality control mechanism and to train the SNRs to set quality seeds. This would enhance both the quality of marking and the quality of seeds and avoid lamentations expressed by NM17 and SNR10.

The SMs concurred with the examiners that there was no special recruitment of the examiners when OSM started; some examiners were recruited and trained during the OSM period, and training on e-marking was done during live marking. They aired their views thus:

There was no special training for examiners. All examiners from manual marking were transferred to e-marking with their roles. For some subjects it turned out that older markers had no computer skills. Younger markers had to take on senior roles because of their technological competencies. The examiners were trained on the job. New examiners were trained on e-marking as they came in (SM1).

There was no special recruitment and training, all examiners involved in manual marking were migrated to e-marking. The training for manual marking was sufficient for e-marking, the only difference was on the marking tools (SM2).

The SM responses confirm the examiners' assertion that the training they received was general and not specific to PBM or OSM. This was more evidence that the SMs and the examiners

thought that marking competence was independent of marking mode. There was evidence that the SNRs were trained to mark on screen by SMs. Asked about their role in the OSM of O Level Biology, SM1 responded, *I trained senior markers on quality control, seed setting*; while SM2 responded, *I trained and supervised senior markers. I was a co-subject manager*.

The SMs therefore needed to be trained to work in the OSM environment.

5.4.2 SM training

In response to the question, *how were you trained to perform your role in the OSM environment*, some SMs said they were not formally trained for the roles in the OSM environment. They explained themselves as follows:

There was no formal training, I learnt by discovery. Groups of SMs were learning as they were working. At first we presumed that script portions were missing, until we discovered that they were never missing but had not been loaded onto the system (SM1).

I joined ZIMSEC when e-marking had already started. I learnt from colleagues as we worked. They provided me with notes that they had compiled from practice. Experienced colleagues would assign me some tasks and assist me work through them (SM3).

I joined ZIMSEC in December when people were preparing for marking. I never received any formal training. I learnt from colleagues and by discovery. I would click icons, read and do. We compiled our own guide that we used for e-marking (SM4).

These responses confirm the assertion that the ZIMSEC staff gained superficial knowledge from the SWP as indicated in Section 5.3.3. SM2, however, said *I was trained by SWP*. The document review indicated that SM1 and SM2 were invited to an OSM training workshop designed for the administrators and SNRs in December 2012 (E-marking Training Programme, December 2012). The training, however, ran concurrently with the live marking as indicated by the SMs and examiners. A statement in the training programme reads thus:

The trainees are advised to appreciate that the training is being conducted during a live marking session with set timelines. The delivery of the December 2012 marking should, therefore, take precedence in terms of thrust. Should there be need for an additional training session to accommodate any training gaps; the issue will be escalated to the relevant authorities... For security reasons, system access privileges will be restricted to designated personnel only.

This statement shows that the training was not given the importance it deserved and that some trainees did not have hands-on experience during the training. There were two possible reasons for the contradiction between SM1 and SM2. It could be that SM2 was given system rights during training while SM1 was not, or SM1 was overwhelmed with live marking duties since he was the SM for one of the pioneer subjects (personal experience). I was also invited to the same training session where I was trained as an administrator and a SNR trainer (E-marker Training Programme, December 2012). The SNR trainers and the SNRs for the November 2012 live examination were trained in the same session in less than two hours. I was provided with quick reference guides and Power Point presentations for use. I trained the SNRs for the Biology (5008/4) and other subjects in the June and November 2013 examinations. I was one of the lucky few who were given system rights which gave me the opportunity to practice with the SNR and administrator roles. From my own experience, the training of trainers was time constrained and rushed. The SWP support teams were allegedly not well versed in the OSM system as well. In his email to the SWP (email, 18 May 2014), the MAM wrote thus:

The onsite support team does not always display comprehensive knowledge of their own system but have to rely on further instructions from developers in the UK. This limits the functions of both ZIMSEC staff and the SWP support team. The Council views this as a big compromise and expense and the desired ideal is to facilitate a direct skills transfer between SWP and ZIMSEC staff and not through third parties who sometimes do not have adequate knowledge.

The MAM was also not amused by the training that ran concurrently with live examinations and wished for a separate training session, writing thus in the same email:

Skills transfer happens during live session and there is need to conduct a more detailed system appreciation training session to broaden the knowledge base for bureau operators off-peak....There is need for SWP to appreciate the ZIMSEC environment is different from theirs and there would be exceptions to the manner in which business is conducted and that it is not possible to replicate the SWP environment.

The statements by the MAM suggested that the SWP was trying to train ZIMSEC staff to use the OSM technology in the way it was used in the UK. This could be the reason why the bureau staff opted to run the June 2015 examination alone and the SMs compiled guiding notes from their practice and managed to operate with minimum support from the SWP (Email, 25 February 2015; Lessons Learnt Report, June 2015). By retaining the key OSM rights and compressing the training for both examiners and ZIMSEC staff, the SWP did not give the ZIMSEC the opportunity to adapt the OSM technology to the local context, lest the former lost control and business. This was more evidence that the SWP was out to do business while the ZIMSEC was looking for a technology that could enhance the efficiency of marking at minimum cost, leading to challenges in the adoption and use of the software.

Document review and personal experience, however, indicated that the SWP designed and supplied the ZIMSEC with the training materials that covered all aspect of the OSM technology. Table 5.5 presents the training materials that were identified through document review.

Table 5.5: OSM training materials provided to ZIMSEC

Document name	Target group	Purpose
Quality control	SMs	Describe quality control by seeds and percentage double marking
Marker quick reference guide	All markers	Provide examiners with guidance on marking scripts.
Senior marker quick reference guide	SMs Senior markers	Provide SMs with content for training senior markers Provide senior markers with guidance on senior

Document name	Target group	Purpose
		marker roles
E-marker administrator walk-through	SMs	Train SMs on controlling marking quality in the OSM environment
Parameter calculator	SMs	Provide guidance on setting quality control parameters (seeds and percentage double marking).
E-marker administrator training manual	SMs and other administrators	Provide guidance on all aspects of the OSM technology, from planning, scanning, and marking to exporting of marks.

These materials were probably not used as evidenced by some claims that there was no formal training, leading to the compilation of guiding notes by the SMs. The training materials were probably not very relevant to the ZIMSEC context and were largely ignored. All the same, the SMs managed to train the examiners to mark on screen and the senior markers to monitor the quality of marking in the OSM environment. To enhance the mastery and consistent application of the mark scheme, ZIMSEC organised standardisation meetings for the examiners.

5.4.3 The standardisation process

Two face-to-face meetings were held for standardisation purposes. These were pre-standardisation and standardisation meetings. Standardisation was known as coordination by SMs and examiners (personal experience; E-marking schedule, November 2014).

5.4.3.1 Pre-standardisation meeting

Senior markers met to set the marking standards for each examination. The SMs were asked to explain the purpose of the pre-standardisation meeting and they responded thus:

The meeting is conducted face-to-face and is meant to prepare for the big meeting. Major issues relating to the examination are discussed; problems that might arise during

marking are also discussed, such as malpractice cases; Seeds were not discussed; Mark schemes are discussed and edited before uploading onto the system (SM1).

Senior markers coordinate the mark scheme generated at item writing; they check to see if the marking guidance is correct (SM2).

The senior marker SNR10, in the face-to-face interview, responded thus:

Leadership discuss and fine-tune the mark scheme; Errors and omissions are corrected; they also mark dummies; the meeting lasts two days for leadership (SNR10).

The quality of marking, therefore, depended on the ability of the SNRs to set standards. The seniors could have set good or poor marking standards, enhancing or compromising the quality of marking. The duration of the pre-standardisation meetings for the November 2014 examination was shorter than the two days that were indicated by SNR10. As mentioned earlier, a document relating to the November 2014 e- marking session (Memo, 23 October 2014) communicated an instruction to reduce the marking duration for OSM. The instruction read thus:

We have been advised to come up with a programme aimed at reducing the e-marking period for the second and third sessions by rescheduling the coordination and other related activities for your components to a period within the first session (6-16 December 2014). May I therefore kindly request you to meet and agree on a viable programme detailing the check-in and check-out dates, the numbers involved and the revised dates?

Subjects had been divided into three marking sessions, and 5008/4 was slated for the third session that was supposed to run from 6-12 January 2015 (Memo, 23 October 2014). The 5008/4 pre-standardisation meeting was scheduled for one day on 4 January 2014 and so was the standardisation meeting on 5 January 2014. The SMs were instructed to reschedule the standardisation meetings to December 2014. This implied that the examiners attended the standardisation meetings by 16 December 2014 and then marked on screen in January 2015. It is questionable if they still remembered the mark scheme and the marking points after the

Christmas and New Year breaks. The long break between standardisation and live marking could result in the examiners forgetting some marking points, leading to inconsistent application of the marking scheme, and hence poor quality of marking.

The SM3 who was directly in charge of Biology (5008) examinations, however, concurred with SNR10 that the pre-standardisation meeting was two-days long:

Senior markers would check in earlier than the rest of the markers. They would coordinate the mark scheme in one day; they would also mark and discuss the dummies the next day. The normal markers would be checking in the second day. The whole coordination period would take about five days.

SM3 emphasised that 5008/4 was much easier to standardise than 5008/2. The senior markers for the former would therefore meet for a shorter period. This could explain why the 2014 meeting was scheduled for one day.

The two days did not seem to suffice for the first OSM of 5008/2. In a report on the November 2015 marking exercise, PMS for 5008/2 indicated that the time allocated to the meeting was inadequate, writing thus: *more time need to be given to the PMS/DPMS & BMS to carry out the exercise.*

DPMS and BMS referred to the deputy principal marking supervisors and belt marking supervisors respectively, whom together with the PMS, were known as senior markers (personal experience). The above statement therefore implies that senior markers were not given adequate time to go through the pre-standardisation meeting, leading to quality control challenges that were highlighted in the report. The challenges are reported in the appropriate section of this chapter later.

5.4.3.2 Standardisation meeting

The standardisation meeting was attended by all examiners to discuss the mark scheme edited by senior markers. The standardisation meeting was described thus:

The examiners mark dummies and come up with a standard for each question. The normal markers make their input to the mark scheme at the coordination meeting (SM2).

Normal markers are provided with an edited mark scheme from the leadership meeting for a discussion. The input of examiners to the mark scheme is generally reduced. They mark dummies. New answers from the dummies are added to the mark scheme before it is closed. The meeting is usually two and a half days long (SNR10).

The quality of the standardisation meetings depended on the quality of standards set by the SNRs in the pre-standardisation meeting. The good or poor standards were passed on to the normal markers, enhancing or compromising the quality of marking. The standardised mark schemes were uploaded onto the OSM system by the information technology technicians, as elaborated below:

Mark schemes are edited and uploaded onto the system (SM1).

The senior markers would provide me with the edited hard copy of the mark scheme. I would type in the additional answers onto the original soft copy. I would give the edited soft copy to IT technicians who would segment the mark scheme to suit the item designation summary before uploading it onto the system (SM3).

Probed on the item designation summary, SM2 and SM3 referred me to a document that was created by SMs and sent to the SWP who would define parameters for identifying and validating the mark scheme as presented in Section 5.3.3. A review of the document showed that it states the number of marks allocated to each item and how each item would be marked. The IDS also defined the marking method, either ticking or direct entry of marks into boxes (IDS 5008/2, June 2016). I also had firsthand experience in designing the item designation summary when I was a relief SM for Biology (5008) in 2013 and 2016. As the QPDM I also coordinated the design of the IDS for all components marked on screen and forwarded them to the SWP. Any errors in the IDS could result in marking errors that compromised marking quality.

In the face-to-face interview, SNR10 indicated that the standardisation activities enhanced mastery of the mark scheme and increased the speed of marking, saying: *coordination promoted internalisation of the mark scheme by active participation. The activities also increased marking speed.*

The examiners on WhatsApp were asked to share their experiences of standardisation meetings, focusing on adequacy of time, relevance to e-marking and the extent to which the meetings enhanced quality of marking. Some participants felt that the standardization time was adequate:

Time was adequate considering that everyone grasped the e-marking idea within the specified time, and the software was easy to use. Use of dummies and seeds for training addressed the issues of quality and uniformity (NM12).

Time for standardisation was adequate. If I am not mistaken there were less seeds and they were increased to improve the quality of marking, thanks to these meetings (NM27).

The NM12 thought that seeds were used for standardisation and training. Another normal marker thought the same, writing thus:

Whilst I agree that the marking and discussion of seeds do great to enhance quality marking, I had reservations on the number of dummies/seeds used for standardisation. I think more time and more seeds should be used if we are not to undermark or overmark some candidates (NM11).

Seeds were never used for training and standardisation as presented by SM4 earlier on pre-live training in Section 5.4.1. As mentioned earlier, I superintended over the Biology (5008) examinations and had firsthand experiences with standardisation and pre-live training. There were no seeds in the training mode. NM11, however, makes a valid suggestion that more time should be created for practice with the seeds, suggesting thus: *if more time is taken on standardisation, no marker should be stopped during marking.* Extended standardisation periods

would allow examiners to mark more dummies and capture more answers from the candidates. Others felt that the standardization time was not adequate.

The examiners raised concerns that the correct answers that were missed at standardisation could not be added onto the mark scheme once marking had started. The examiners wrote thus:

With e-marking, there is no room for inclusion of new ideas that might arise during live marking, thus there is need to take more time exploring all possible responses to a question (NM11).

Correct points that do not appear on the marking scheme are ignored. If you try to mark them correct on a seed, you will be stopped from marking (NM12).

I probed the examiners to suggest how additional answers could be added after marking has started and NM12 responded: *that will be tricky, I think. Altering the mark scheme would mean remarking all the marked scripts. To avoid that more dummies are supposed to be used rather than leaving other valid points.*

There was no consensus among the SMs about additional answers. SM2 insisted that additional answers were added to the mark scheme even after marking had started: *additional answers that were missed at the standardisations will be communicated to examiners. Marking will be temporarily stopped to add the new answers. IT technicians will load the answers onto the system if they are readily available.* SM3 gave a contrary response; *we never stopped to add answers to the mark scheme, probably because the examiners never raised such answers. We could add the answers if we so wished of course. The IT guys were available to do that.* Probed on the need to remark some scripts after adding new answers, SM3 responded; *mmm, that will be tricky. We may need to assess the impact of the new answers on the performance of the candidates. In order to remark, the SM would have to delete all the marked scripts and send them to the marking pool and that will be cumbersome.*

The contradicting responses indicate that there was no standard way of dealing with correct responses that were missed at standardisation meetings. I supervised the marking of Biology (5008) examinations in November 2013 and June 2016. I do not remember the examiners bringing new answers during marking. It could be that the examiners did not come across additional answers or they simply ignored them. If new answers were indeed ignored as alleged by NM11 and NM12, then the quality of marking was compromised. The actual marking commenced after the standardisation meeting. The quality of marking was monitored through a variety of ways.

The next section presents findings on the monitoring of marking.

5.5 Monitoring the quality of marking

Sub-question 2: How is the quality of marking O Level Biology monitored in the OSM environment?

Document review showed that the ZIMSEC could use three approaches to quality control; the seeding, the percentage double marking and the S-Process (Quality control document, n.d; Parameter calculator user guide, n.d; Senior marker quick reference guide, n.d). The ZIMSEC, however, used seeding and percentage double marking. Seeds were used to control the quality of marking O Level Biology (5008) examinations (interview responses; E-marking programme, November 2014; D-grade reports: A-G, 29 January 2016; PMS report, January 2016; Quality control overall report: 5008/2, November 2015; Quality control overall report: 5008/4, November 2015).

The SMs were provided with a guide that they used to set seed parameters in the system [Parameter calculator user guide, n.d (Appendix L)]. Two types of seeds were used to control the quality of marking and these were the qualification and the seeds.

5.5.1 Qualification

The qualification seeds were defined as the number of seeds presented to a marker at the start of each day. The seeds could be limited to a number of days or could be marked every day of the

marking period. The parameters for the qualification seeds were set on the calculator (Appendix L). The SMs could set the qualification tolerance, i.e. the acceptable range of deviation from the senior marker's mark (quality control document, n.d). The examiners who failed the qualification seeds were stopped from marking the question. From my experience with PBM as an examiner and SM, the examiners were required to mark ten live scripts soon after the standardisation meeting. They were assessed by senior markers, who would either allow them to mark their allocation or request them to mark some more practice scripts. The SMs concurred that qualification replaced the ten live scripts used to assess the examiners in paper-based marking. Probed on qualification, one SM elaborated thus:

Qualification seeds replaced the first 10 live scripts marked by examiners in manual marking. The seeds were marked for the first two days. Examiners were not informed about the duration of the qualification so that they would continue to exercise caution during marking (SM1).

Probed on why the qualification seeds were only marked for two days, the SM responded thus:

Fast examiners would have finished marking in the first two days; qualification would therefore not be useful thereafter (SM1).

This statement suggests that e-marking is fast, if examiners can exhaust their marking allocation in two days. Another SM described how qualification seeds worked, elaborating thus:

Qualification seeds ensure that all markers fully understand the mark scheme before they are allowed to mark live scripts. Examiners are given a prescribed number of portions of every question to mark correctly (8-10). Examiners are stopped for wrong marking and restarted. The stopped examiners discuss the failed seeds with senior markers before they are reactivated (SM2).

The qualification, therefore, acts as the first line of defense against poor markers. Qualification parameters could be changed during marking, as indicated by the senior marker who participated in the face-to-face interview:

There were too many qualification seeds in N2015; 10 that were reduced to six. Qualification seeds were marked at the start of each day (SNR10).

The majority of examiners for 5008/2 failed qualification seeds as indicated in the PMS report. The PMS for the paper thought that lengthy responses led to the failure of the qualification seeds, writing thus: ‘...because of the lengthy responses which cascaded to massive failure to qualify for the actual marking....’ (PMS report: 5008/2, November 2015 examination). Senior marker SNR10 concurred with the PMS, that 5008/2 examiners failed qualification seeds, citing qualification as a challenge in the OSM of the paper.

In 2015 there was no marking for the first 3 days; examiners were failing the qualification seeds (SNR10).

Compared to the first 10 live scripts, the qualification was a better assessment process, given that it was automatic. There were no chances of the system allowing deviating examiners to proceed to mark live scripts. However, the quality of the qualification seeds depended on the quality of the marking standards set by the SNRs in the pre-standardisation meetings and passed on to the examiners in the standardisation meetings. Good marking standards would enhance mastery of the marking scheme, and hence good quality qualification seeds. Once the examiners passed the qualification seeds, the system presented them with live scripts, with seeds appearing at set intervals. The next section presents findings about seeds that were used to monitor the quality of marking after qualification.

5.5.2 Seeds

A close analysis of the quality control calculator guide showed that seed parameters were set at question level, not at component level because some factors may vary depending on the complexity of the mark scheme and the marking tolerance. There were three input variables that

the SMs were required to key into the calculator, and these were the number of parts to be marked, number of (normal) markers and the length of actual marking period. The parameters included the percentage of the seeds (the number of seeds that were presented in every 100 clips of scripts, presented in pairs); the size of the seed window (the rolling number of seeds that would be used to determine ongoing marking quality); and the maximum number of seeds a marker could fail (Parameter Calculator Guide, n.d.).

A review of the Quality Control Document (n.d.) shed more light on the setting of seed parameters. Besides the seed window, seed percentage and seed failure, the SMs were required to set the minimum number of seeds per question required for quality control to be enabled, referred to as the seed bank size. It was evident that seed parameters could be changed by inputting new variables. There were chances of lowering or increasing the seed percentage, reducing or increasing the frequency at which the seeds appeared to the examiners. The quality control parameters could therefore be manipulated by the SMs to suit whatever they wanted, unless there was a governing policy. There was no evidence of any policy governing the quality control parameters.

The SMs confirmed that ZIMSEC used a seed window of 10 and a seed failure range of 30%, setting correct marking of seeds at 70%. One SM2 explained thus:

For a seed window of 10, examiners have to correctly mark 7 out of 10 seeds. At 5%, there is a pair of seeds for every batch of 40 script portions. If an examiner fails one seed they are allowed to continue marking. If they fail the pair, they are stopped. If they fail three, they are also stopped. All the failed seeds are presented in a report when the examiner is stopped. The discussion with senior markers then begins (SM2).

There was evidence that seeds were set by senior markers. Responding to the question, *who set the seeds*, the SMs had this to say:

Senior markers set the seeds soon after the pre-live training. Seeds are set in the live mode (SM1).

Three senior markers sit together to select portions for seeds (SM2).

Two or more senior markers sit together as they set seeds (SM3).

To reduce the chances of setting wrong seeds, two or three senior markers set seeds for one item, and then they move to the next item (SM4).

The senior markers on the WhatsApp platform also indicated that they set the seeds in groups to enhance quality. SNR5 wrote, *seeds need to be carefully done by a group of senior markers not by an individual to reduce chances of making obvious errors.* To support SNR5, SNR7 emphasised that it was frustrating to realise that they had been stopped by wrong seeds, writing thus: *zero tolerance to wrong seeds. ...Seeds need to be carefully selected; poor seeds lead to unnecessary stoppages; may be frustrating upon realisation.*

The quality of marking largely depended on the competence of the SNRs. As presented earlier, the SNRs set the marking standards by editing the mark schemes and marking dummies; they pass on the marking standards to the examiners through standardisation meetings; they set the qualification and the seeds. It was therefore important to earnestly build the capacity of the SNRs to carry out these quality control activities.

The participants were asked to describe the criteria for selecting seeds. The SMs responded thus:

Not tricky; should be legible; something worth discussing; no blanks, there must be something written; reasonable; and legible (SM1).

Not tricky/complicated to reduce stoppages; it should have correct answers as discussed in the guidance; should be legible (SM2).

Straight forward; wrong or correct answers; selected from a wide range of candidates' responses. We needed some seeds where the examiners awarded no marks (SM3).

In the face-to-face interview, SNR10 was asked to share his experience with seeds, emphasising on the quality of the seeds and he responded thus:

Seeds set in 2015 were not appropriate, they included ineligible scripts and vague language, 60-80% of seeds were good; 20-40% were problem seeds. Senior markers were not trained to set seeds.

SNR10 indicated that he started OSM as a normal marker and was later promoted to a SNR in 2017. As presented earlier, the SNRs were trained to perform their OSM roles by SMs. I trained 5008/4 SNRs to set seeds in the November 2013 examination. It could be that 5008/2 SNRs were trained to set seeds in 2015 when the paper was first marked on screen and there was no training in 2017 when SNR10 was promoted. In a follow-up interview on the assertion that senior markers were not trained to set seeds, the SMs indicated that it was not necessary to train the senior markers to set seeds because seed setting is a marking activity. The SM4 responded thus:

They did not need any special training for seed setting. Seed setting is marking, seeds are set in the live mode. It involves viewing a script and deciding if the responses make a good seed. If not, the examiner clicks the skip icon. To enhance quality senior markers would pair up to select the seeds.

SM3 also explained:

It's a decision-making process, whether to award a mark or not. If the script is a suitable seed, the examiner marks and clicks the seed icon to set the script as a seed. Once they have gone through coordination and pre-live training, then they should be able to set quality seeds, unless they have not mastered the mark scheme.

The responses from the SMs confirm that SNR10 was really not trained to set seeds. It is apparent that the SNRs were trained to set seeds once, when the component was first marked on screen. It could have been necessary to train the SNRs to set seeds at every session to refresh their memory and to train newly promoted SNRs.

Some unsolicited responses on the WhatsApp platform shed more light on seeds. One such response came from NM12: *seeds also ensure adherence to the marking scheme*. Another spontaneous response from NM11 showed that fewer seeds regularly showed up and could be recognised by markers:

Experience with e-marking shows that some seeds have a pattern of showing up, or are kind of recognisable during live marking. In this case the marker takes due care on such scripts, or in some cases the marker deviates from the mark scheme which is obviously not recognisable if they happen not to be stopped and in this case quality marking is compromised.

The SMs confirmed that some seeds showed up at intervals so regular that examiners could identify them and take due care when marking them. One SM responded thus:

The number of seeds is determined by the parameters put in the calculator. Few seeds can be predicted; examiners see the seeds and mark them correctly. A higher percentage of seeds would be better for quality control (SM1).

SM4 indicated that seeds recurred because they were over-used: *over-used seeds tend to recur until examiners can recognise them. Such seeds are retired/deleted*. Probed on what she meant by over-used, SM4 said the system presented more seeds to deviating examiners. Given that the seed window is fixed in size, the same seeds would be presented to the same examiner several times. This could explain why individual examiners marked huge sums of seeds in the November 2016 marking session. An analysis of the automatically generated quality of marking reports for Biology 5008/2 and 5008/4 for the November 2016 examinations showed that some examiners marked large numbers of seeds. The maximum numbers of seeds marked by individual examiners were 2980 and 859 for 5008/2 and 5008/4 respectively. If the seed window for a question was 10, then each seed appeared at most 280 and 85 times respectively. Examiners were likely to recognise seeds that appeared many times to them, defeating the whole purpose of setting the seeds.

NM19, in a spontaneous response, concurred with SM2 that increasing the number of seeds would improve the quality of marking, writing thus:

Seeds need to be increased in number to ensure proper quality control. They impede marking speed yes, however, if more time is allocated to the marking period it will improve us as markers.

Increasing the number of seeds would also increase the possibility of stopping deviating examiners. However, there are adverse consequences that come with a large seedbank. More seeds would be presented to the examiners, increasing their marking load (Quality Control Document, n.d.) and the marking period. Increasing the number of seeds to more than 5% was an unlikely option for SMs, who were pressured to meet set deadlines. Justifying the suggestion to increase the time for marking, NM19 elaborated thus:

This may help avoid strain caused by the process, considering that it is a sedentary job with little movement. If more time is allocated alertness and level of concentration increase, hence increased quality of marking.

This statement implies that the OSM process exerted some strain that could lead to poor marking. Despite the strain, extending the marking period was an unlikely option for SMs as explained earlier, the examiners had to bear the strain in order to mark and meet set deadlines, compromising the quality of marking.

There was evidence that all examiners failed seeds during marking. An analysis of the quality of marking reports (5008/2, November 2015; 5008/4, November 2016) indicated that all examiners, including senior markers, failed seeds. The reports showed the total number of seeds failed by each examiner. The maximum number of seeds failed was 577 and 184 for 5008/2 and 5008/4 respectively. The examiner who failed 577 was a senior marker.

Eight examiners absconded the November 2015 marking session and were awarded the D-grades for misconduct at the marking venue. An analysis of the D-grade reports indicated that the quality of marking was assessed by the number of seeds failed by the examiner. The eight examiners were named A to H. The senior markers wrote thus about the examiners:

Examiner B had a very poor understanding of the mark scheme. He scored 16.16% seeds failed. Examiner E had a fairly high % of failed seeds, 8.6% which is fairly above the average of 5.07. Examiner C interprets the marking scheme well but had a fairly high percentage seeds failed, 6.6% which is above the average of 5.07.

The D-grade reports provided evidence that quality of marking was being monitored indeed. As presented in Section 5.3.1, SM3 indicated that examiners were assessed and awarded grades during training and at marking. The examiners awarded D-grades were to be demoted or dismissed. Four senior and four normal markers were awarded D-grades and were dismissed from marking teams. I received the D-grade reports from the PMS and dismissed the examiners when I was managing the subject in June 2016.

The D-grade report form required the assessment of the interpretation and application of the mark scheme, commitment and general attitude to the marking exercise and any other behaviour that could compromise the quality of marking, e.g. missing deadlines, completion of mark sheets and absenteeism (D-grade reports A-H, 29 January 2016). The possibility of earning a D-grade could discourage behaviours that could lead to poor marking. The reports about SNRs were, however, worrisome. Given that the SNRs set the marking standards and supervised the normal markers, they would be expected to be exemplary in their conduct and marking competency. For them to misinterpret the mark scheme, fail to mark seeds, and abandon the marking exercise was unexpected. A close analysis of the form, however, showed that it was designed for PBM and was not adapted to OSM, as evidenced by the assessment of completion of mark schemes, when the marks were automatically captured by the OSM system.

The examiners were stopped from marking when they deviated from the acceptable tolerance range. The senior markers were supposed to discuss the failed seeds with stopped markers. Systems were in place on how to deal with the stopped markers.

5.5.3 Dealing with the stopped markers

The SMs indicated that the senior marker module allowed the SNRs to see examiners who had been stopped for failing seeds and they could open the seeds for discussion with the markers. The examiners could also approach the SNRs as soon as they were stopped. Asked about the stopped markers SM1 responded thus:

Markers were stopped by qualification and seeds. Some stopped markers approached senior markers. Others would just sit until the administrators noticed that they were not marking. That's when they would start approaching the senior marker for discussion.

Asked if there was any mechanism of enforcing the discussions, the SM1 responded thus:

No, there was no mechanism for enforcing discussions between senior markers and stopped markers. They might even activate an examiner without discussion.

The responses by SM1 provide evidence that seeds only identify poor markers but discussions between the SNRs and the stopped markers was the actual quality control activity. The SNRs who activated stopped markers without discussing the failed seeds deliberately compromised the quality of marking. SM2 concurred with his colleague that discussions were initiated either by SNR or stopped markers, explaining thus:

The senior markers have a window for stopped markers, which they should frequently check. The window displays all stopped markers and senior markers select examiners that belong to their team and invite them for discussions. Normal markers should also report that they have been stopped. Some markers do not report that they have been stopped but opt to start a new question. A new question means qualification seeds, which

will delay marking. It also means good markers are depleting portions and the stopped markers will be paid for very few portions.

The examiners who abandoned the marking exercise in the 2015 examination were probably frustrated by frequent stoppages. As indicated earlier by SNR7, it was frustrating to be unnecessarily stopped by seeds, and by SNR10 who said examiners wasted two to three days marking qualification seeds. The examiners were requested to share their experiences of the discussions between the senior markers and stopped markers. SNR6 concurred with SMs that discussions were either initiated by SNR or stopped markers, writing thus:

Constant checking of the stopped marker platform or the marker alerts the senior that they had been stopped. The latter is common. The leader is under pressure to mark since he/she has a big allocation to finish.

This response suggests that some SNRs concentrated on marking more than on monitoring the quality of marking. SNR10 described the logistics of the discussions: *senior markers would open failed seeds and discuss with normal markers. Problem seeds would be highlighted by senior markers for action by the SMs.* This response suggests that SNRs set some faulty seeds that stopped markers. This was evidence that that some SNRs were not competent with seed setting, yet the quality control mechanism depended on the mark awarded by the SNRs on the seeded scripts. Faulty seeds frustrated examiners as indicated by SNR7. Some faulty seeds would not be identified when the SNRs activated the markers without discussing the failed seeds.

NM12 indicated that the discussion was only necessary when they did not understand why they had been stopped, writing thus:

Discussion done if the marker doesn't understand why they failed the seeds. If you revisit the failed seeds and see your mistakes then there is no need for discussions, but if you don't see your mistakes then you discuss with your senior.

NM25 decided to come to my inbox to share her experiences of the discussions in private and wrote thus:

Well, I was a first-time marker and got hung up on this one seed that got me every time. So, I was constantly doing the walk of shame to my supervisor in order to continue marking. Finally, I gave up on the question and focused on better ones.

NM25 concurred with SM2 that some stopped markers abandoned the question that stopped them and started on new ones. Asked if the senior marker discussed the marking points before activating her, NM25 responded thus, *yes, every single time. I felt like a school kid in the last class.* NM25's experiences suggest that she stopped approaching the senior marker for discussions because she was ashamed of her poor performance. As mentioned by SM1, there was no mechanism of enforcing the discussion between stopped markers and SNRs. This meant that the quality control mechanism could be bypassed when stopped markers decided to start a new question and when SNRs decided to activate stopped markers without discussing the failed seeds.

As presented earlier in this section, the document review indicated that all examiners failed some seeds (Quality of Marking Reports 5008/2 & 5008/4, November 2016). The examiners were asked if SNRs were also stopped by seeds and they said yes. Probed on why SNRs were stopped by seeds when they had two coordination meetings, the examiners responded thus:

Some seeds caused that too (SNR22).

Some seeds were wrong (NM30).

In some (rare) cases, the seeds might be the ones with errors, resulting in markers being stopped (NM12).

These responses suggest that wrong seeds unnecessarily stopped examiners, confirming SNR7's frustration with wrong seeds. The seed that constantly stopped normal marker NM25 could have been wrong as well. The findings on dealing with wrong seeds are presented later in this chapter.

5.5.4 Permanently stopped markers

The document analysis indicated that the senior marker window allowed the SNR to permanently stop an examiner from marking a particular question (Permanently Stopped Marker Report: 5008/2, November 2015; Senior Marker Quick Reference Guide, n.d.). The Permanently Stopped Marker Report for the 5008/2 November 2015 examination had one marker who was stopped from marking question 6(a). The marker had been stopped nine times by seeds before he was permanently stopped from marking the question. The SMs were asked to describe the criteria for permanently stopping an examiner from marking a question. SM1 responded thus:

The icon has only been used accidentally. It can only be used when markers are marking from home and there are no discussions among the markers. Deviation had been minimised by marking a series of the same question with the same response.

This response suggests that marking from a central venue facilitates examiner discussions that could enhance the quality of marking. This is contrary to the finding that all examiners, including senior markers failed seeds, to the extent of delaying marking of the November 2015 5008/2 examinations. NM2 suggested that very poor examiners could not mark beyond the qualification, saying this:

This option has always been clicked by error. The administrator then reviews to see who stopped the marker. A really bad marker who has not understood the mark scheme should be stopped from marking. Such an examiner would have falsified information about their subject competence and would therefore not understand the mark scheme. Such an examiner would not normally qualify to mark; they would be stopped by qualification seeds, so they would not proceed to mark live scripts. Markers were permanently stopped purely by error.

This response points to flouting of recruitment procedures. As presented in Section 5.4.1, examiners were assessed on the subject content, trained and assessed on their ability to mark. A marker can only falsify their subject competence if they did not go through the selection and recruitment process. SM3 suggested a different reason for not permanently stopping poor markers; *the icon was never used because the SM had an option of deleting scripts marked by an examiner when they doubted the quality of the marking.*

A review of available documents confirmed that the SMs could delete script portions marked by suspicious examiners. The administrator module allowed the SM to select a marker and delete scripts marked by the examiner at a particular time range. However, the SM only deleted marks upon request by the SNR (Marking Monitoring Phase Walk Through, n.d.). Commenting on addition of more answers to the mark scheme, SM3 earlier indicated that deleting marked scripts was cumbersome, providing evidence to suggest that poor markers were allowed to mark and not permanently stopped from marking certain questions. SM4 suggested that poor markers could actually be permanently stopped, saying this:

The icon was used to stop bad markers from marking particular question. Such markers would have partially attended the standardisation meeting. Some examiners write university examinations during the marking period, so they leave the coordination meeting midway through.

The responses suggest that examiners could be permanently stopped from marking when they had not understood the mark scheme and they continually deviated from the seeds. For some reason, the SNRs never permanently stopped examiners, yet evidence abound that examiners failed huge sums of seeds. This could be a result of technological incompetence on the part of both the SNRs and the SMs who trained them.

As mentioned in Section 5.5.3, the responses on stopped markers suggested that senior markers set wrong seeds, known as suspect seeds (OSM Work Schedule, November 2014; Marking Monitoring Phase Walk Through, n.d.).

The next section presents findings on how wrong seeds were dealt with.

5.5.5 Dealing with suspect seeds

Probed on why senior markers set wrong seeds, SM4 indicated that markers set wrong seeds because they had not understood the marking scheme, saying: *Senior markers set bad seeds because they misunderstood the mark scheme, not because they were not trained to set seeds. They too, failed the seeds that they did not set.*

SM4 acknowledged that SNRs misunderstood the mark scheme, yet they were the pacesetters for the marking standards. If SNRs could misunderstand the mark scheme, what more the normal markers? SM1 earlier said that some young normal markers assumed senior roles because they were more technologically apt than older SNRs. It could be that the younger markers who were promoted to senior roles had limited marking expertise. There was need to extend the standardisation period as suggested by some examiners. There might also be some need to reconsider the recruitment and training procedures for SNRs in the OSM environment.

Examiners were also probed on why senior markers set wrong seeds, SNR6 responded, stating that *senior markers are chosen by humans who have their own weaknesses*. I asked if SNR6 was suggesting that some SNRs were not supposed to be SNRs in the first place, in which case SNR8 responded thus:

No, but fatigue maybe, during marking. Plus mamouse aisazikanwa nemasenior akawanda saka vaigona kubaya pasipo nemistake (moreover most senior markers were not versatile with the computer mouse, so they could have clicked wrong icons by mistake).

The response confirms SM1's statement that some old senior markers had no computer skills, leading to the promotion of younger markers with computer skills. SNR7 responded to SNR8 thus, *Uuuuu! I think it was just poor choice of seeds...and implications*. SNR8 suggested another reason for setting wrong seeds, *probably seeds were done by individuals thereby making them wrong in some way*. Normal markers weighed in as follows:

May be some are no longer classroom teachers and are not in tandem with the current data (data is colloquial for content) (NM30).

Maybe it's due to human error (NM29).

The responses provided ample evidence that SNRs set wrong seeds that were identified and flagged as suspects. NM30 apparently suggested that the ZIMSEC recruited examiners who were classroom teachers, concurring with SM3 who stated that the ZIMSEC lowered the teaching experience from five to three years for prospective markers to capture younger markers who were technologically competent. The teachers probably needed the five years teaching experience to become competent markers after recruitment and training. The ZIMSEC might need to find ways of striking a balance between technological competence and marking expertise, lest quality is compromised.

The examiners were asked to explain how the wrong seeds influenced the quality of marking. SNR8 responded thus:

Wrong seeds compromised standards since you will have to drift away from the standard marking points to suit the requirements of the wrong seed, so that you are not stopped from marking.

NM32 came in, writing thus:

The wrong seeds made us feel incompetent, yet we would be doing our best. Wrong seeds cause unnecessary stoppages. And besides, you end up marking with fear of being interrupted rather than fear of doing injustice to the learner.

NM25 supported NM32, stating, *it feels unnatural to mark wrong answers as correct and vice versa.*

These responses suggest that the examiners wrongly awarded marks or penalised candidates to suit the requirements of the wrong seeds. The wrong seeds therefore compromised the quality of marking.

There was evidence that the senior marker module allowed the SNRs to escalate wrong seeds to the SMs. The wrong seeds were marked as suspect seeds in the administrator module where the SMs would frequently delete them (OSM Work Schedule, November 2014; Marking Monitoring Phase Walk Through, n.d.). The participants were asked to describe how the wrong seeds were dealt with. The SMs responded thus:

Wrong seeds were identified in the discussions of seniors and stopped markers. They were retired, effectively deleting and sending them back for marking (SM1).

Retire the seeds and send them back to the marking pool. Retiring a seed deletes the mark given to the candidate by the senior marker. When the seedbank approaches minimum, the senior markers are activated in the seed setting mode. When the seedbank goes below minimum the system automatically stops all markers. The SMs should therefore closely watch the seedbank so that markers are not stopped (SM2).

Bad/wrong seeds were escalated to SMs by senior markers as suspect seeds and retired (SM4).

A SNR concurred with the SMs that wrong seeds were indeed retired by the later. SNR10, however, indicated that suspect seeds were not immediately deleted in the November 2015 examination session, saying this:

Some were retired or deleted after some time. The SM was not readily available to assist in the 2015 marking session. Senior markers raised the issue of bad seeds with the SM, who did not solve the problem. Managers for other subjects helped to deal with the bad seeds; they deleted the seeds and seed setting was re-done

This response suggests that the examiners did not receive the support that they needed for effective quality control of marking in the November 2015 examination, leading to the challenges raised by the PMS in his report. NM20 suspected that the OSM technology sometimes accepts wrong marking because marking proceeded without seeds. Writing on the WhatsApp platform, this marker had this to say, *guys am worried about e-marking on the fact that it comes a time when even wrong answers will be accepted, like there is a loophole somewhere.*

Probed on why he thought wrong answers were accepted, NM20 responded thus:

There are times when seeds are not available. I believe at those times wrong answers can be marked correct. I realised that after qualification the speed increases and concentration reduces.

I followed up with a SM about the examiner's suspicion. They asserted that there was never a time when there were no seeds in OSM system. SM4 responded thus:

There was never a time when marking proceeded without seeds. Examiners were not always stopped from marking when they failed seeds. They were only stopped if they failed more than the tolerant range of 30%. They marked up to 70% of their allocation without being stopped. A closer look at the quality of marking reports would show that even good markers failed some seeds but they would be within the tolerance range. The seeds were an efficient quality control system as evidenced by exemplar scripts that were printed by the SM for grading. No scripts had been wrongly marked.

SM4 added that NM20 did not understand how seeds worked. As presented in Table 5.4, the SMs could adjust quality control parameters during marking. While it could be possible that NM20 did not understand how seeds worked, it could also be possible that there were times when seed parameters were set at low levels, such that deviating examiners were not stopped any more. If a pair of seeds was presented for every 40 scripts (5%) as stated by SM2, then 38 scripts (95%) were marked without quality control. The Parameter Calculator Guide (n.d.) emphasised

that a higher seed percentage would probably increase the quota size (the number of script portions marked by each examiner). The SMs would, therefore, not take the risk of increasing the marking load and extending the marking period, hence the 5% or even lower. NM20's suspicion could be pointing to a real loophole in the seeding approach to quality control.

The quality of marking was also controlled by allowing examiners to escalate problems to senior markers and administrators.

5.5.6 Escalated problem scripts

An analysis of available documents showed that all markers could escalate problem scripts to SNRs who had a problem resolution tool on their module where they could park (defer resolution) or resolve the problem. In resolving the problem, the senior marker could view the comments made by the examiner, explaining why the script was escalated, open the candidate's script, mark it or send it back to the examiner with some comments/instructions. Some problems could not be resolved by senior markers and these were escalated to the SMs who would resolve them in the administrator module (Senior Marker Quick Reference Guide, n.d.). The problems escalated to SMs were marked as rescan requests in the administrator module. The SM could either accept or reject a request. The request was accepted by clicking the tick on the selected problem. A return-to-base (RTB) dialogue box would then appear displaying a 'yes' and a 'no'. Clicking a 'yes' meant that the script would be returned to the scan centre for killing (deleting from the script marker) and subsequent preparation for marking on paper. Clicking a 'no' meant that the script was returned to the scan centre for rescanning. Rejecting a request returned the problem script to the senior marker with a comment on why the script had been returned (Marking Monitoring Phase Walk Through, n.d.). The problem resolution tools offered the OSM technology opportunities to enhance quality of marking.

In 2013, the SWP encouraged the SMs to keep on top of the rescan request queue and liaise with the scan centre manager to evaluate whether approving a rescan is possible or a problem script should be killed (email, 17 December 2013). An increased number of killed scripts implied an increased number of scripts marked on paper, bringing in the challenges associated with the PBM. The nature of the software, however, increased the number of killed scripts.

In 2014, the MAM wrote an email to the SWP expressing a concern that the script marker did not allow the bureau operators to edit entries made in error. The MAM indicated that if an operator killed a script by error, the transaction could not be reversed; if an operator erroneously indicated that a candidate had additional pages, the entry could not be reversed (email, 19 May 2014). The MAM pleaded with the SWP to design a function that allowed operators to undo bad operations, writing thus in the same email:

The desired idea is for the system to prompt the operator (e.g. are you sure you want this candidate to have additional pages), to allow operators to revise their entries.

The inability of the script marker to undo the commands increased the number of the scripts that needed to be escalated to SMs as rescan requests and for killing. Increased PBM in the OSM environment would compromise the quality of marking by bringing in challenges associated with the paper mode. The examiners were asked to describe the kind of scripts they escalated to the senior markers, and these were the responses:

The candidates' response seems correct but is not on the mark scheme. If you try to mark them correct on a seed, you will be stopped from marking (NM12).

Where the handwriting is not legible (NM18).

Script without answers/blanks (NM14).

NM14 did not seem to understand that blanks were not supposed to be escalated to SNRs. The document review indicated that blank scripts were passed by clicking the 'not attempted' icon (Marker Quick Reference Guide, n.d.). In the face-to-face interview, SNR10 indicated that blanks were as many as the candidates who registered for every two questions in Section B of 5008/2, where the candidates chose three out of five questions. He also confirmed that the examiners would click the 'not attempted' icon to pass the scripts, increasing the marking load for examiners for no additional payment. Blanks could also arise when the candidates did not

write responses in some spaces. If all blank scripts were to be escalated to the senior markers, they would be overwhelmed by the problems which needed to be resolved. NM11, however, thought that some problems were not escalated to senior markers and said, *whilst in principle the marker can escalate issues, it is difficult to ascertain if that happens in practice.*

The examiner raised an important point that was not addressed by the data gathered in this study. Examiners could escalate scripts that they could not mark. What happens if the examiner chooses not to escalate a script they have failed to mark? Did the system allow the examiner to choose another script when they have not escalated a problem script? These questions need to be answered in another study. The examiners were asked about the action taken by senior markers on the escalated problems. NM11 responded and pursued his argument thus:

They are obviously rectified but worry is on those that may not be escalated and the marker is not stopped because the scripts might not be seeds. What quality control is there for marked scripts using the e-marking?

This statement shows that NM11 thought that escalating the problems to the senior markers was not an effective way of controlling quality in the OSM environment. As stated earlier, the data collected in this study did not answer the issue of problem scripts that were not escalated. There were no responses from the senior markers about the escalated problems on the WhatsApp platform. Probed on the escalated problems, SM1 indicated that most of the escalated problems from SNRs emanated from scanning, explaining thus:

Scripts with duplicate or missing pages could not be marked on the system. Some images were blurred. Such scripts were pulled out and marked on paper. Scanning was done overnight, the people probably got tired. I intended to visit the scan centre to investigate how such scripts arise but e-marking was suspended before I did that.

This response confirms the MAM's concern about the script marker's inability to undo erroneous commands (email, 19 May 2014). It could also explain why the escalated problems were set to appear as rescan requests in the administrator module and why the SWP recommended that SMs

keep on top of the rescan requests (email, 17 December 2013; Monitoring Marking Phase Walk Through, n.d.). It was apparent that most of the problem scripts could not be identified and validated in the system. Duplicate and missing pages trace back to printing. There was need to control the quality of printing in order to minimise problems at scanning. There was need for the ZIMSEC to adequately staff the scanning bureau to avoid overworking and tiring out staff by overnight shifts. Improved scanning would reduce killed scripts and enhance the quality of marking.

SM4 indicated that the SNRs were not allowed to leave the marking venue before resolving all escalated problems. This was expressed thus:

I would check that there were no parked problems before clearing senior markers. Otherwise the system would not export marks if there were unresolved problems. Some of the scripts could not be marked because they were a combination of responses from several candidates. Such scripts probably confused the system and required to be returned to base.

The combination of scripts for several candidates indicated that the night shifts alluded to by MS1 were taking a toll on the bureau staff, emphasising the need to adequately staff the scanning bureau and eliminate night shifts. Asked on how they resolved the escalated problems, SM4 responded thus:

We would view scripts escalated by senior markers in the administrator module. Some of the scripts had no problems at all. Some blurred text required the examiners to use the zooming tool on their screen and mark. We would return such scripts to senior markers with instructions on how to resolve them. Scripts with real problems were returned to the scan centre by clicking the 'return to base' icon, which sent the scripts to the scanning room.

This response indicates that both SNRs and normal markers escalated some problems that they could solve using tools on their screens. It could be that the examiners were either not trained to use the marking tools available to them or they were not trained to identify problem scripts that

needed to be escalated. They ended up escalating scripts that seemed to cause the slightest challenge. The participants were asked about the kind of feedback that was provided to examiners.

The next sub-section presents the findings on examiner feedback.

5.5.7 Examiner feedback

There was no consensus about the feedback provided to examiners on the quality of their marking. Some SMs indicated that there was no feedback provided by the OSM technology to examiners about the quality of their marking, responding thus:

Unless the marker requested, no feedback was given. The quality of marking report was visible to the administrator (SM) only (SM1).

There was no feedback, except that senior markers could see that some examiners had been stopped by the seeds (SM3).

There was no feedback directly to examiners. Quality of marking reports were only visible to the SM in the administrator module (SM4).

Some SNRs concurred with the SMs that the OSM system did not provide the examiners with the feedback about the quality of their marking. The senior markers wrote thus on WhatsApp:

No feedback was possible. No one knew how many good portions they had marked except the counter on how many you had marked for each part question (SNR10).

There was no feedback that could be given (SNR9).

In the face-to-face interview SNR10 responded thus:

There was no feedback on the number of seeds marked or failed. The system would show the number of portions marked, including seeds. The examiner would not know their marking progress and information about the quality of their marking.

It was apparent that the examiners did not get any other type of feedback apart from stoppages. As indicated by SM1, some senior markers did not discuss failed seeds with stopped markers. The examiners therefore went through live marking with no feedback about the quality of their marking. Feedback on the quality of marking was likely to help examiners improve their marking. SM2, however, insisted that the system provided feedback to examiners and said: *the examiner window displays the quality of marking. They can check on their quality of marking. They were trained to use the icon.* A response from NM11 also suggested that the system provided feedback to examiners: *despite that option being available, surely there wasn't enough time to constantly check for that at all.*

If indeed the OSM technology provided feedback to examiners, it is evident that NM11 never looked at the feedback because he had no time. The lack of consensus on the responses highlights the shortcomings of the interviews in this study where observation was not possible. A glance at the examiner window would have put the matter to rest.

The next section presents findings on the use of the automatically generated reports to monitor the quality of marking.

5.5.8 Automatically generated reports

Analysis of the Monitoring Marking Phase Walk Through (n.d.) document and personal experience showed that reports could be viewed and downloaded in the administrator module for the purpose of monitoring the progress and quality of marking and assessing the examiners. A bouquet of automatically generated reports for the 5008/2 and 5008/4 2015 November examination was reviewed.

Both the script marker and the e-marker generate reports that can be viewed in the administrator module. The Script Marker Reports were used by the scan centre personnel to manage the

scripts. The e-Marker reports were used by the SMs to monitor the quality and progress of marking. Similar reports were identified for the two Biology components 5008/2 and 5008/4 for the November 2015 examinations. More than 20 reports were identified with some in purely technical language and beyond the scope of this study. The reports with data about the subject component, examiners and candidates were analysed. The findings are summarised in Table 5.6.

Table 5.6: Automatically generated reports (5008/2, 2015; 5008/4, 2015)

Report	Data displayed	Possible uses
Component statistics	-number of registered centres; registered candidates; scanned scripts; absent candidates; markable scripts; scripts returned for rescanning; scripts marked; scripts not marked; unexpected scripts; scripts taken out of the e-marker for manual marking (killed); candidate whose marks were exported	assess marking progress; can give an indication of pirate candidates in the component; candidates whose details were wrongly captured can be traced; to account for scripts in the component; to determine possible missing scripts; to determine number of scripts that were killed and marked on paper; indicate the popularity of the subject
component part completion	Number of scripts per question; number of registered candidates; type of marker, problems that have not been resolve; scripts marked per question, scripts not marked per question	assess marking progress at question level; determine the overall speed of examiners; indicate SNRs' progress and competences in resolving problems
Marking progress	total number of parts to be marked; number of parts available for marking and parts marked	assess marking progress; assess overall speed of examiners
Quality of marking overall	All examiners invited for marking; the quota allocated to	Assess the quality of marking for each examiner; identify normal

Report	Data displayed	Possible uses
	each examiner; quota actually marked by each examiner; number of seeds marked and number of seeds failed by each examiner	markers who can be promoted to SNRs in future examinations; identify markers who need close monitoring or training
Marker portions remarked	List all examiners and the parts they marked; list examiners whose script were remarked; indicate changes in marks after remark	Assess examiners' marking competence; evaluate the overall quality of marking for the session; assess the credibility of scores for each component
Markers not in quota groups	Lists all examiners and their marking status; SNR or normal marker	A record of the examiners who were invited to mark in a particular session
Permanently stopped markers	Lists markers who were permanently stopped from marking particular questions	Assess examiner competences; evaluate the effectiveness of the training and standardisation exercise; assess the competence of senior markers in using the stopped marker management tool; evaluate the mark schemes
Quality of marking by related part	Number of linked questions each examiner marked; number of seeds marked and failed; number of times stopped; number of times trained	Assess examiner competences; identify normal markers who can be promoted to SNRs in future examinations; evaluate the quality of seeds set; assess SNRs' competences in setting seeds; evaluate the quality of the marking schemes
Quality of marking	Lists all examiners; quota allocated to each examiner;	Assess examiner competences; evaluate the quality of seeds set;

Report	Data displayed	Possible uses
	quota marked by each examiner; seeds marked and seeds failed; suspect/deleted seeds; average % seeds failed	evaluate the quality of the mark schemes; assess SNRs' competences in setting seeds; can be used to pay examiners
Marker Activity	Times when each examiner logged in and out of the system; the activity they were undertaking; time when a seed was marked; whether the examiner failed or passed the seed. Indicates examiners who were invited but did not turn for marking (no activity against the names)	To monitor examiner activities; assess how examiners used time during the marking session; select examiners for promotion; identify examiners who need close supervision

It was evident that the reports were generated in the administrator module and could therefore be viewed and downloaded by the SMs. The SMs were provided with a document that guided them in accessing the reports (Marking Monitoring Phase Walk Through, n.d.). There was evidence that the SMs used the reports to monitor the progress and quality of marking real-time. There was need for the ZIMSEC to design a procedure of availing the reports to the SNRs who assessed other examiners. SM4 noted that the component part completion report sometimes indicate that marking had been completed, there were no more portions to mark when marking was not actually complete. SM4 had this to say:

There is a time when the report says no more parts to mark. Examiners leave the marking venue and the next day the report indicates that there are portions to mark. In some instances, the portions continue to increase every hour. One or two examiners who live near the marking venue or Head Office have always been recalled to mark the pop up parts.

My experience as a relief SM for Biology (5008) in 2013 and 2016 also showed that all quality of marking reports were in the administrator module. I used some of the reports and the audit trail to monitor the progress and quality of marking real-time and downloaded some for grading and other purposes relating to the assessment of the subject. I also experienced the inconvenience caused to the examiners and the SMs by the pop-up scripts. The pop-up scripts were possibly a result of communication breakdown between the ZIMSEC and the SWP, especially the third part who was based in India, who had no direct link with the ZIMSEC but was tasked with the key activity of segmenting scripts into individual items. The ZIMSEC incurred unforeseen expenses by hiring examiners to clear the pop-up scripts. All the same, the real-time monitoring of the progress and quality of marking by a variety of automatically generated reports could enhance the quality of marking.

The quality of marking was also influenced by the way in which question papers and mark schemes were designed.

5.6 Test design issues

Sub-question 3: How do O Level Biology questions and mark schemes influence quality control in the OSM environment?

The findings on this research sub-question were presented in two headings, i.e. the question paper structure and questions and the marks schemes.

5.6.1 Question paper structure

Documents provided evidence that ZIMSEC designed both constrained (where answer spaces were provided on the question papers) and unconstrained (where separate answer sheets were provided) question papers for PBM. A review of the Lessons Learnt Report for the June 2013 examination and communication between the SWP and ZIMSEC (Emails, 8 April 2013; 29 April-12 May 2013) indicated that the two paper structures presented some challenges in the OSM environment. Table 5.7 presents the challenges posed by each paper structure and the solution that mitigated the challenge.

Table 5.7: Challenges and solutions: Constrained and unconstrained papers

Paper Structure	Challenge for OSM	Solution
Constrained	The numbers of pages were multiples of four, to allow them to be folded into booklets at printing. When the last printed page was on odd number, blank pages were added to come up with a multiple of four. Candidates could write responses on the blank pages. The examiners would not see and mark the responses written on the blank pages	SWP designed inline additional pages (ILAP) to replace the blank pages for the June 2013 examinations; some candidates wrote additional answers on the ILAP; system identified the ILAP and queued them up for review by senior markers.
	The ILAP presented challenges: increased workload for SNRs; extended marking period; increased costs	Eliminated blank pages; printed pages in multiples of two; glue together instead of folding pages for November 2014 and beyond (Biology 5008/4), was marked on screen for the first time in 2013).
Unconstrained	In PBM candidates could use any available paper, presenting challenges for scanning.	A 12 page answer booklet designed for the June 2014 examination and beyond.
	All OSM scripts were segmented into individual items. The segmentation process for the unconstrained	All question papers were constrained before they could be marked on screen to eliminate the unconstrained scripts and attendant costs.

Paper Structure	Challenge for OSM	Solution
	scripts was laborious and labour intensive. The process was slow and inaccurate; it ran concurrently with marking, reducing the number of computers available to markers; extended the marking period.	
	Additional manpower for segmentation sourced from the UK, through the SWP, at an additional cost of £4000.	

The decision to design the constrained papers only for OSM brought major changes to the question papers structure, creating an undesirable platform for the violation of the assessment framework prescribed by the Biology (5008) syllabus. This led to the design of the specimen papers and circulars that informed examination centres of the changes (5008/2 question papers: specimen, November 2015, June 2016, November 2016; Examination Circular Number 8 of 2015). The bouquet of subjects marked on screen in November 2013 was constrained, even in the PBM, and the circular that was sent to schools had no statements about the structure of the papers (Examination Circulars No. 41 of 2013; No. 42 of 2013). That is the time when 5008/4 was introduced to OSM. A review of the 5008/4 question papers confirmed that the 2013 paper was similar to the papers marked on paper in 2012 and backwards. The 5008/2 paper had two sections, Section A which was constrained and Section B that was unconstrained as indicated in the Biology (5008) syllabus. Section B was however constrained to avoid the segmentation process which proved laborious and expensive as presented in Table 5.7. A circular was sent to examination centres to communicate the new design (Examination Circular No. 8 of 2015). A statement in the circular read thus:

With effect from November 2015...Biology 5008/2...will be electronically marked. This new arrangement will have a bearing on the design of the question papers to make them e-marker compliant. For that reason, the question papers will be structured such that candidates will write all their answers on the question papers. The content of the subjects/components remains the same.

The circular provided evidence that question papers were indeed redesigned to suit the demands of the OSM technology. It was apparent that the papers were redesigned to enable the software to identify them at scanning so that correct script portions could be presented to examiners for marking. The document review indicated that the ZIMSEC had to provide the SWP with some blank question papers and item designation summaries that would allow the later to set up the script identifying and validation parameters and conduct a trial run before scanning. It was evident that successful scanning depended on the ability of the scanners to identify the script whose parameters were set by the SWP (OSM Work Schedule, November 2014; Lessons Learnt Report, November 2013). The SWP worked closely with me in my capacity as the QPDM to access the documents that were prepared and checked by SMs for accuracy as indicated in Table 5.4. Any scripts that did not suit the defined parameters were either rejected by scanners or were rejected by the e-marker even if they were successfully scanned. Such scripts were subsequently returned to the bureau for subsequent rescanning or killing as indicated in Section 5.5.6.

There was evidence that page numbers were also used to identify the scripts during scanning. However, the page numbers for Biology 5008/4 (which was a constrained paper in PBM) were crowded by text in the November 2013 examinations, delaying scanning. The script validation icon was turned off on some scanners to allow smooth scanning (Lessons Learnt Report, November 2013). This is evidence that the page numbers and the blanks were not adequate script identifiers; hence the need to switch off the script validation icon, allowing any script to be scanned and saved onto the script marker but rejected by the e-marker, increasing rescan requests presented earlier. SM1 was therefore justified to suspect that most problem scripts emanated from the scanning bureau.

As presented in Section 5.3.3, the SWP recommended bar-coding as a way of identifying and validating scripts, a suggestion which ZIMSEC never implemented. As presented in Section 5.3.3, the ZIMSEC leaders dragged their feet when it came to making key decisions about the OSM technology. The barcodes could have greatly reduced scanning challenges and enhanced the quality of marking.

Scanning challenges could also have emanated from the examination centres and candidates as evidenced by elaborate instructions to the same. The circular (Examination Circular Number 41 of 2013) that communicated the introduction of 5008/4 to the OSM emphasised the need for special handling of scripts to avoid problems during scanning, with a statement reading thus:

E-marking is a system where the examiner marks candidates' scripts online. The scripts are initially fed into the computer via a scanner. To avoid problems during scanning the scripts need special handling and care.

The circular went on to enumerate the special handling as follows:

- Answer scripts should not be perforated or defaced in any way and the use of strings and staples should be discouraged.
- In cases where the candidate has used additional pages to answer questions, centres/invigilators must ensure that a neat hole is perforated on the left hand corner of the answer script using a paper punch and a string inserted on the hole to secure the candidate's script.
- Answer booklets should not be separated for any reason during the course of the examination and during packing.
- Invigilators should complete attendance registers diligently.
- All scripts should be dispatched to ZIMSEC within 24 hours (for urban centres) and 48 hours (for rural centres) of the completion of the examination.

A closer look at the first and second bullets indicates that ZIMSEC issued conflicting instructions. The first instruction discourages perforation of scripts and the next instruction

allows it. This could have led to some scanning challenges emanating from defaced scripts, leading to problem scripts that were escalated and subsequently killed. Another circular relating to the marking of 5008/4 (Examination Circular Number 42 of 2013) insisted that the candidates must use black pens and H.B pencils only and should write their candidate numbers in the boxes provided on every page of the question paper. A review of question papers indicated that the instruction to use black pens and H.B pencils was on some question papers and not on others. In the absence of the instruction, some candidates might have used blue or some other colour, probably leading to the blurred scripts mentioned by SM1.

The candidates were given the instruction to write their details (name, candidate number and centre number) on every page, and to check if there were missing or duplicate pages and ask for replacement of booklets to replace the defective ones (5008/2 question paper, November 2015; June 2016, November 2016). It was apparent that some scripts were separated either at the examination centre or at the scanning bureau, as evidenced by composite scripts mentioned by SM1. In the event that some candidates did not write their details on all pages, some scripts would never be matched to the right candidates, risking remark requests. There was evidence that some candidates did not ask for replacement booklets when there were duplicate or missing pages, resulting in problem scripts that were killed and marked on paper as indicated by SM1.

There was evidence that some examination centres violated the scripts submission deadlines. The Scanning Bureau supervisor reported that some scripts for an examination sat on the 31st of October 2014 had not been delivered for scanning by the 17th of November 2014 (report, 17 November 2014). Violation of these instructions could have led to scanning challenges that compromised the quality of marking. Late scripts were probably scanned late, leading to the reduction of the marking period as was done in November 2014. In the worst case scenario, the scripts might not have been scanned, increasing PBM marking and its associated challenges. Some of the instructions were reiterated on the question papers that were reviewed.

The provision of answer spaces on all the question papers could have reduced writing space for some candidates, compromising the quality of marking. Some candidates used additional answer sheets that could not be scanned for marking on screen (Lessons Learnt Report, June 2013),

confirming that the answer spaces provided on the question papers were not enough. A review of the question papers for 5008/2 and 5008/4 indicated that there were no clear criteria for determining the amount of space provided in the question papers. There were variations in the amount of the space provided for the same type of answers worth the same number of marks as exemplified in some examination sessions. In the 5008/2 June 2016 examination, the following questions had varying amounts of answer spaces:

6(a)(i) Describe how the structure of an artery is related to its function: 8 lines for 5 marks.

6(a)(ii) Describe how the structure of a vein is related to its function: 6 lines for 5 marks.

7(a) Describe how the structure of a red blood cell is adapted to its function: 7 lines for 5 marks.

In the 5008/4 November 2013 examination, the following questions had variations in the spaces provided:

1(b)(iii) Give an explanation for the difference between P and Q.

4 lines for 2 marks.

1(b)(v) Explain why there was no reaction in test tube R.

2 lines for 2 marks.

All the reviewed question papers had no consistency in the spaces provided for the same type of questions with the same number of marks. The candidates were likely to either write answers on inappropriate places or use additional pages when the answer spaces were inadequate. The examiners would not see and mark answers written on inappropriate spaces, compromising the quality of marking. Additional pages increased the number of scripts that were marked on paper, compromising the quality of marking.

Commenting on the structure of Biology (5008) papers marked on screen, NM11 had this to say:

The structure of the paper is good. Space for responses is enough. However, experience shows that the space is not enough for those learners who can't be precise. In the end they will write all over making it difficult to read especially with e-marking...Check how the paper fits into the software...some edges are cut losing some text which is not written on the provided space.

The examiner seemed to suggest that the highly structured format of the question papers compromised quality of examinations when some candidates' responses were cut off or were not seen by examiners. This is confirmation that constrained papers reduced answer spaces for some candidates, prejudicing them. The restructuring of 5008/2 had implications for the type of questions and mark schemes.

The next section presents the findings on the type of questions and mark schemes for Biology (5008) examinations.

5.6.2 Type of questions and mark schemes

A review of the question papers and mark schemes for the Biology (5008) examinations marked onscreen indicated that there were no changes to the number of questions, marks and exam duration for both 5008/2 and 5008/4 (Appendix I and J). So, some aspects of the assessment scheme in the syllabus were preserved. The types of questions were, however, at risk of being changed, violating the weighting of the assessment objectives. A review of the question papers and mark schemes provided evidence that 5008/4 (which was a constrained paper even before OSM) had no apparent changes to the questions and mark schemes, only blank pages were eliminated from the question paper. Changes to questions and mark schemes were noticed in the Section B of 5008/2 examinations.

The questions in Section A elicited shorter responses than in Section B. Following the challenges that bedeviled the OSM of the 5008/2 November 2015 examination, the PMS of the paper called for Section B to be adapted to e-marking (PMS report, January 2016). This call greatly influenced the type of questions in future examinations (5008/2 June and November 2016; June and November 2017). There was significant reduction in the number of marks per question. For

example, the longest question in Section B of the 5008/2 November 2015 was worth 12 marks, while the longest question in the same section for 5008/2 June 2016 and November 2016 was worth five and six marks respectively. The assessment scheme for the paper was clearly violated in a bid to make the examination e-marker compliant. The responses from the participants confirmed that the restructuring of the question papers influenced the type of questions for the OSM examinations.

In the face-to-face interviews, the SMs were asked to comment on the structure of the Biology (5008) examinations that were marked on screen between 2013 and 2017. Their responses provided evidence that the type of examination questions changed when question papers were restructured to suit the technology. SM1 responded thus:

The papers had to be restructured. I think the questions should be designed to elicit a maximum of three marks. Questions with four marks and above would give challenges during marking. There were allegations that examinations were watered down when they were highly structured. If we do not resume e-marking it is better we go back to the essay type questions in Section B.

This comment points to a violation of the assessment scheme provided in the Biology (5008) syllabus, especially the skills weighting. The allegation of watering down the examinations could erode the credibility of examinations marked on screen. SM2 and SM4 responded thus:

The examinations were made of objective questions where answers were not debatable, with a maximum of ten marks and identifiable marking points. Seeds would then be easy to select. Long answers would need to be marked by percentage double marking, for example essay (SM2).

Long questions need to be marked by percentage double marking. Questions for seeds should have a maximum of 4 marks (SM4).

The comment by SM4 implies that seeds were not appropriate for quality control in Section B of 5008/2. The responses by all SMs provide confirmation that 5008/2 examinations were designed to suite the demands of the OSM technology, in violation of the assessment scheme prescribed in the syllabus. The question papers were thus designed to improve the efficiency of marking, at the expense of the validity of the examinations. Probed on the allegation raised by SM1 that 5008/2 examinations were watered down by highly structuring Section B, SM4 answered thus:

The questions were made more accessible to candidates. Even before e-marking, Biology (5008) subject panelists proposed that the questions in Section B need to be focused. It was not watering down but focusing the questions.

The response concurred with the finding that the questions in Section B of 5008/2 were longer in 2015 but became shorter from June 2016 as presented earlier, risking the violation of the skills weighting prescribed for the paper by the Biology (5008) syllabus. I tend to agree with SM1 that the restructuring watered down the examinations. I do not think that focusing a question means shortening it and converting it from a free response to a highly structured format. Even an essay question can still be focused to elicit the intended responses from the candidates. The proposal by the panelists, who sit and brainstorm all science examinations (personal experience), to focus the questions by shortening them could result in the watering down of all science examinations at the ZIMSEC, unless SMs of the calibre of SM1 stand their ground to protect the validity of the examinations.

In the face-to-face interview, SNR10 was asked to share his experiences with the length of answers in the paper that he marked. He responded thus:

Section A of Paper 2 was easy to mark. Even in 2015 there were no problems in Section A. In Section B there were too many marking points for one question. One question was worth 8 marks but had 15 marking points. There was too much reading and scrolling for such questions.

The response confirms that Section B questions were shortened in subsequent examinations to make them easier to mark, compromising the validity of the examinations. A review of the November 2015 5008/2 mark-scheme confirmed that a few questions had more marking points than the marks. For example, question 6(b) which was worth 12 marks had 16 marking points. The science panelists could either devise a consistent way of awarding marks to candidates, in the event that the marking points exceed the number of marks, or they could focus the questions to elicit responses that are commensurate to the marks. The latter option might be difficult to achieve with free response questions. The challenge of scrolling and too much reading could lead to fatigue and frustration of examiners, and hence inconsistent marking. SNR10 was further asked to give his opinion on the type of questions that should be set for Biology examinations marked on screen and he responded thus:

Questions should be specific, with fewer marking points. Questions like describe and explain would result in long mark schemes. The examiners were required to pin the mark scheme on the screen but they could not do so because of its length, so they relied on the hard copy of the mark scheme. The 2015 paper 2 was designed for manual, not for e-marking. The 2016 paper had been adapted to e-marking and was easier to mark.

The examiner concurs with SMs that short questions are more suitable for OSM than long questions. SNR10, however, provided some more evidence that the 5008/2 examinations were redesigned to suit the demands of the OSM technology, increasing the risk of violating the assessment framework of the O level Biology (5008) course. As indicated in the assessment scheme, 5008/2 is the core of the examination. The shortening of the questions in Section B compromised the validity of the whole examination by watering it down as indicated by SM1.

On the WhatsApp platform, the examiners were asked to share their experiences with the paper structure. Senior markers responded thus:

Too much scrolling was a major challenge (SNR5).

Section A was far much simpler than B where more time was required to read the answer (SNR9).

The responses confirm SNR10's concerns that reading and scrolling were real challenges in the OSM of 5008/2 examinations. As mentioned earlier, too much scrolling and reading could lead to examiner fatigue and frustration, leading to inconsistent marking.

The examiners were asked if any paper could be marked onscreen without compromising the quality of the examinations. SNR9 wrote, *yes, provided that quality seeds are set without rushing*. NM12 wrote, *in my opinion any paper can be e-marked though questions requiring descriptions and explanations are difficult to mark using the e-platform; easy for questions that require candidates to name, state, list, etc*. NM25 responded, *structured questions are more suitable*.

The normal markers concurred with the SMs that short answer questions should be set for examinations that are marked on screen. It is therefore evident from all data sources that 5008/2 examinations were redesigned to shorten the questions so that they could be easy to mark on screen, enhancing the efficiency of marking. However, the assessment framework of the course was violated, compromising the validity of the Biology (5008) examinations.

The examiners were asked if practical examinations could be marked on screen without prejudicing the candidates. An interesting conversation ensued among the examiners, with some saying it is possible while others said it is not. The conversation went thus:

It's quite difficult because centres have conditions which are unique so sometimes you need to consult the reports from each centre (SNR22).

Very possible, other exam boards are doing it, e.g. Cambridge (NM28).

SNR6 responded to NM28, *it will be difficult because of a variety of answers. We need to have chemicals from the same source.* On the other hand, NM28 responded to SNR6, *we need to adjust the questions chete (only) to have a universal key.*

NM28 reacted to SNR6's suggestion of having chemicals from the same source, *even for Cambridge the source of chemicals is not the same.* NM26 came in with an exemplar practical question, *imagine a question on inheritance were candidates picked beads independently. The variation of answers can't be standardised.* Also, SNR7 weighed in, *I think it is possible. Questions should be considered to limit variations in responses. Responses should just be within range.* Then SNR5 joined in as well, *very difficult since the correct answers are very varied. Seeding is a challenge.* At this turn of responses, SNR7 changed her mind, *very difficult due to lack of standard materials, labs etc, between urban and rural schools.*

The conversation seemed to suggest that practical examinations are difficult to standardise, but some examination boards have somehow managed to mark them on screen. The difficulty to mark practical examinations on screen was also confirmed by the SMs as indicated in the Section 5.3.1.2 because the scripts were marked against the supervisors' reports. A review of the 5008/3 November 2013 question paper and mark schemes confirmed that the practical examination conformed to the criteria for OSM examinations described by both examiners and SMs. The 5008/2 paper was short (two questions worth 20 marks each) and constrained even in PBM. The maximum marks for an item were six. Each examination centre was required to submit a supervisor's report about the conduct of the examination. A sample of the supervisors' reports from different centres was reviewed.

The report provided evidence that it was compiled by the teacher who supervised the practical examination. The teacher would have received instructions from ZIMSEC about the material requirements of the examination and would then open the question paper two days before the examination and run the practical exam to try out the materials. The teacher would then answer the questions based on the trial run (Question paper 5008/3, November 2014, November 2013; Supervisor's reports, November 2013). There were indeed variations in the responses to the

questions asked on the report form. There were two main questions that the supervisor from each examination centre were asked to answer. Below are the questions and sample responses.

1. Was there any difficulty experienced in providing any necessary material? Some of the responses from the supervisors read thus:

Exam Centre 1: *The reagents DCPIP and ascorbic acid required for Question 1 were not in the instruction sheet, fortunately the school managed to get these reagents on time.*

Exam Centre 2: *The school does not have a science laboratory. It is not capable of running such an examination.*

Exam Centre 3: *Ascorbic acid and DCPIP were difficult to get. We prepared 1% DCPIP.*

2. Did the candidates experience any difficulty during the course of the examination? If so, give particulars briefly. Reference should be made to:

- (a) Difficulty arising from faulty specimen

Exam Centre 1: *No.*

Exam Centre 2: *N/A.*

Exam Centre 3: *In place of ascorbic acid, 1% citric acid was used for question 1 and grapes replaced grapefruit.*

- (b) Accidents to apparatus or materials

Exam Centre 1: *No.*

Exam Centre 2: *N/A.*

Exam centre 3: *No.*

- (c) Any information that is likely to assist the examiner, especially if this cannot be discovered from the scripts.

Exam centre 1: *Question number 1 required use of 0.1% ascorbic acid and DCPIP. The percentage of DCPIP was not given. We, however prepared 1% DCPIP.*

Exam Centre 2: *N/A.*

Exam Centre 3: *In place of ascorbic acid, 1% citric acid was used for question 1 and grapes replaced grapefruit.*

It is apparent from the responses that there was no instruction about the concentration of DCPIP on the instruction sheet, giving rise to possible variations in the candidates' responses; the supervisors were at liberty to substitute materials they could not find. The candidates' responses actually varied depending on the materials provided by their centre, hence the requirement of the supervisor's report for each centre. As indicated earlier, all OSM examinations were marked at item level, making it difficult to link them to the supervisor's report. The document review confirmed the concerns raised by some examiners and SMs that 5008/3 could not be marked on screen without prejudicing the candidates. The ZIMSEC could, however, benchmark with examination boards that mark practical examinations to learn how the varied answers are standardised in the OSM environment. The responses from Exam Centre 2 were however surprising. Two questions arose from the responses: Why were the candidates registered for the practical examination and not the alternative to practical? How did the ZIMSEC deal with the scripts from the centre? I never followed up on the issue since it was outside the focus of this study.

I followed up on the challenges that were encountered in the marking of 5008/2 in the November 2015 examination, as indicated in the PMS report, D-grade forms and the senior marker who participated in the face-to-face interviews. On the WhatsApp platform I asked the question, "I heard there were challenges when Paper 2 was first e-marked in the November 2015 session. What were the challenges?" SNR10 responded thus:

There was inadequate orientation/training of those who were to set the seeds. As a result poor seeds were set. The poor seeds took too long to be removed. Coupled with a high number of qualification seeds, most examiners failed to qualify and wasted 2-3 days

marking seeds. The questions in Section B of the paper were not ideal for e-marking as they had too many marking points.

The SNR reiterated his face-to-face interview responses. On the other hand, SNR24 responded thus:

I think the training was not enough mainly to the senior markers who were asked to prepare the seeds; the training was done hurriedly and also the marking period was not enough.

NM21 came in with a suggestion, *adequate time is required for seeding and senior markers should mark for a day, identifying and removing poor seeds before others start marking.*

The responses from all data sources point to inadequate training of the SMs, the SNRs and the normal markers, causing challenges in the marking of November 2015 5008/2 examinations. SNR10 earlier pointed that the SM for Biology (5008) was not readily available to assist the examiners with suspect seeds in the same examination. There was a high turnover of SMs for Biology between 2013 and 2016. When the SM left the organisation in June 2013, I was asked to babysit the subject while recruitment procedures were being done. I handed over the subject to the new SM in January 2014. The SM then left the organisation end of January 2016. He was preparing to join another institution by the end of the month while marking started at the beginning of the same month. Apparently, there was no supervision of the examiners at the time when the paper was marked for the first time. For some reason, I was only called to babysit the subject in February 2016 after the SM had left the ZIMSEC. I handed over the subject to another SM in October 2016. I had to endure heavy workloads as I juggled with two full time jobs as QPDM and SM, so I did not do any justice to marking of the examinations. Coupled with inadequate training, the high turnover of SMs evidently compromised the quality of the marking of Biology (5008) examinations.

I solicited the comments of the examiners on the allegation that e-marking watered down 5008/2 examinations. Two senior markers responded thus:

Chances of passing paper two have been increased by having shorter questions; each part question carries less than 9 marks; this is also marker friendly... manual or e-marking (SNR5).

I remember Paper 2 used to have a free response section which allowed candidates' expression, so I guess removing that section kind of did that (SNR7).

The comments suggest that shorter questions were easier to mark but candidates no longer had the liberty to fully express themselves. Shorter questions also made the paper more accessible to the candidates. SNR5's response shows that shortening the questions made marking easier on screen and on paper, confirming that short questions enhanced the efficiency of marking. The ZIMSEC therefore shortened Section B questions to enhance the efficiency of marking in the OSM environment. This was contrary to the assessment scheme provided in the syllabus, which prescribed free response questions in Section B of the paper. The examiners were asked about the opportunity and challenges of the OSM technology to enhance the quality of marking Biology (5008) examinations.

The next section presents the findings on the fourth research sub-question.

5.7 Opportunities and challenges of OSM to enhance quality marking

Sub-question 4: What are the challenges that impede quality control in onscreen marking of O Level Biology?

5.7.1 Opportunities

The face-to-face interview participants were asked to evaluate the opportunities of the OSM technology to promote the marking of Biology (5008) examinations. SM1 indicated that e-marking had the opportunity to enhance quality control in several ways, *deviating markers are stopped; it is faster than manual marking; no transcription and addition errors*. SM2 listed ways in which OSM enhances quality of marking, saying this:

Addition of marks is automatic, unlike in manual marking where some candidates are awarded wrong marks. It saves time as long as there are no challenges. No paper work, the examiners do not have to carry bags of paper. Progress reports are generated for each examiner on an hourly basis. The SM can monitor examiner activities.

SNR10 in the face-to-face interview concurred with SMs that the OSM technology can enhance quality marking, responding thus:

E-marking is fair to all candidates because the scripts are anonymous. It is also fair to examiners because they are paid according to what they mark, a good marker marks more. There is no more tallying of marks because statistics are automatically captured.

The examiners on the WhatsApp platform were also asked to list the advantages of e-marking over paper-based marking. They indicated that OSM is faster and more efficient than PBM, writing thus:

No tallying of marks; fast (SNR5).

Faster and less strenuous (NM30).

Fast; enough sleeping time (rest); less labour; no addition of marks; slow markers not stressed (trying to catch up with others in the group stresses) (NM21).

The normal marker NM21's response that OSM does not stress slow markers is contrary to SNRs' responses that express worry about depletion of scripts by fast markers, as will be presented later in this chapter. NM19 compared OSM and PBM, *it is very efficient, as a verifier I know it reduces time.*

When OSM was suspended in 2018, Biology (5008) examinations were marked on paper for the last time in June 2018. NM19 assumed the role of verifying marks awarded to candidates, a role

that was not necessary in the OSM environment, hence the assertion that OSM is very fast. OSM eliminated some human errors that are associated with PBM. The examiners wrote thus:

Addition errors eliminated (NM25).

It is fast; reduces error on counting marks; no question is awarded more marks than it deserves (SNR10).

The OSM technology creates less pressure on examiners when compared to PBM as indicated by some examiners, who wrote thus:

No need to enter marks at the end of the session (NM11).

Less pressure (NM29).

Also independent of each marker, no need to wait for another person to complete a pile you need to mark (NM19).

No recording of marks; no tallying of marks; fast quality (SNR7).

NM11 thought that the OSM technology reduces malpractice by examiners and explained, *by not seeing candidates' names or centres, it reduces malpractice.*

SNR6 thought that OSM improves quality of marking, *examiners are more cautious when using e-marking than manual. Quality of marking is also improved.*

The responses provide evidence that the OSM technology enhanced efficiency of marking by automation of some marking activities that influenced the quality of marking. The examiners were relieved of the burdens associated with handling paper. However, these opportunities would only be beneficial in the marking of valid examinations. If the ZIMSEC decide to resume OSM, there is need to reconsider setting free response questions and using double percentage marking

as the quality control approach to similar components. The science panel should be trained to interpret and use the syllabus correctly, and to set questions that elicit responses that demonstrate the relevant assessment objectives in order to conform to the assessment framework. Otherwise the Council will chase after efficiency at the expense of validity, and hence quality of the examinations.

The next section presents findings on challenges of quality control in the OSM environment.

5.7.2 Challenges

The participants were asked to describe the challenges that could compromise the quality of marking. Most of the challenges raised by the participants related to the context in which the Biology (5008) examinations were marked on screen, with a few of them relating to the software.

SM1 thought that the centralised approach to OSM exerted pressure on the examiners and the administrators. To justify his assertion, SM1 had this to say:

I had remained at the marking venue with a colleague and a few senior markers to mop up scripts that were popping up. Someone from office phoned us and said ‘what are you still doing there?’ The same person phoned our hotel and instructed them to check us out that very day. We got to the hotel at around 1600hours, only to find that they could not accommodate us anymore. We had to move to another facility so that we could finish off the marking.

The comment reiterates the inconvenience of pop-up portions described by SM4. SM1 however suggested that quality of marking could be promoted when examiners mark from home, acknowledging that the technological infrastructure is a challenge though. SM2 thought that power cuts compromised the quality of marking and responded thus:

There is need for automatic generators as power back up. The breaks given to examiners improve quality of marking. The examiners are given three breaks per day, although a

few would want to continue marking. There is an overnight break that allows the system to save, export and write reports, at this time examiners and administrators are not allowed to log onto the system.

SM2 indicated that the IT personnel at host institutions was not readily available to support the OSM:

In the event of a power cut the generator would automatically switch on but internet would still be switched off. The IT technicians for the host institution were not readily available to switch on the internet, especially on weekends.

The examiners on the WhatsApp platform were asked if there were any challenges that could compromise quality of marking. The normal markers gave an example of the challenges encountered in one marking session:

Unavailability of electricity and internet. A case in Gweru 2017, more often servers would be down and markers had to rush at the last hour to beat the time allocated for staying at MSU; marking late into the night, with fatigue obviously quality is compromised (NM11).

Network problems delay marking and put the examiner under pressure. I remember the first year we had to do shifts throughout the night (NM28).

These normal markers concurred with SM2 that power cuts disrupted marking; the servers were sometimes down; and examiners had to work at night to beat the set marking deadlines. SNR5 concurred with the normal markers and wrote thus:

Server may be down, delaying marking. Average and slow markers under too much pressure, this may lead to a lot of casualties – average and slow markers losing their apportionment to fast markers, yet they are capable of completing it within the stipulated time.

In addition to servers that went down, the SNR raised concern over the pressure created for average and slow markers by fast markers. I asked if there was no limit to what a fast marker could mark and these were the responses:

The limits were above the actual total (NM30).

When one completes his/her allocation they were allowed to pick any question from slow markers (SNR22).

I further probed the examiners and asked, “So fast examiners could continue to mark as long as portions were available?” The examiners responded in the positive:

Yes (NM30).

Usually fast markers will target those portions that were easy to mark and leave the tough ones to the slow markers (SNR22).

I followed up on the issues with SM4 who confirmed that examiners’ allocations were sometimes set at limits way above what was actually available. SM4 explained thus:

The SM is supposed to calculate the exact number of scripts that individual examiners should mark. That number depends on the number of examiners and the number of candidates registered for the component. The SM would then monitor marking progress and reallocate scripts to fast markers to avoid extending the marking period beyond the set dates. To avoid the hustle of reallocating scripts, some SMs allocate a default number of scripts to all examiners, and that number is usually larger than what arises from calculation. Fast markers can therefore mark as many scripts as they can, depleting scripts for slow markers.

SNR33 also raised concern that fast markers wanted the money only and were not concerned about the quality of marking, writing thus:

High speed marking in a bid to attain high volume of scripts for better revenue; Quantity of work instead of quality

I probed SNR33 and asked if this implied that e-marking improves marking efficiency and not quality of marking and the response was *I guess volume can't be used in the formula for remuneration. Quality has to be emphasised.* The SNR's response showed that the examiners were worried about the differences in remuneration between fast and slow markers. I further probed SNR33 and asked him to suggest how quality could be used for remuneration. Two other senior markers responded thus:

Fewer errors higher remuneration (SNR5).

Quality could be independently awarded using a different formula (SNR7).

It was the SNRs who were mainly concerned with remuneration differences. The document review indicated that SNRs marked much fewer scripts than normal markers. The quality of marking reports (5008/2, November 2015; 5008/4, November 2015) revealed the statistics shown in Table 5.8.

Table 5.8: Number of scripts marked by examiners

Paper	Examiner Type	Maximum	Minimum
5008/2	NM	19653	9540
5008/4	SNR	10085	1700

Asked why the SNRs were allocated fewer scripts, SM3 responded thus:

Senior markers need time to resolve problems and manage stopped markers, so they cannot have the same number of scripts as normal markers. We have a formula for

allocating scripts to senior markers and a formula for paying them for their responsibilities.

I asked if the SNRs were paid more or less than the normal markers when the responsibility allowances were factored in. SM3 responded thus:

In belt marking the senior markers would get almost the same amount as normal markers, or slightly more. In e-marking the amount depends on the marking speed of the senior marker. The scripts loaded onto the system belong to nobody. It's like they are in a basket. Fast markers are never stopped by seeds, so they are depleting the actual scripts from the basket. Poor and slow markers are taking more seeds and fewer scripts. If a senior marker is slow, the fast, normal markers will deplete the scripts from the basket, and they might be paid far much more than senior markers. We mark to meet deadlines and to save money, so we keep allocating scripts to fast markers because they are accurate. Quality is never compromised.

This response shows that senior markers worked under pressure to mark more so that their remuneration could be comparable to normal markers. This could be the reason why some of them did not check the stopped marker window and waited for stopped markers to approach them for discussions as explained by SNR7. This could be the reason why some stopped markers would opt to abandon questions instead of approaching senior markers for discussions. Some senior markers might, therefore, not have had the time to effectively monitor the quality of marking because they concentrated on marking their own allocation. If the ZIMSEC decides to resume OSM there is need to revisit script allocation so as to avoid the streak of disgruntlement among the senior markers. Otherwise the quality of marking would be compromised when the SNRs concentrate on marking rather than quality control activities.

The examiners cited more challenges that could compromise quality of marking. NM11 cited discomfort during the marking process, explaining thus:

The immovable nature of desktops creates a lot of discomfort. Laptops can enable a lot of posture flexibility. Ndakarwara mutsipi nomusana after the session. (I had backache and sore neck after the session).

NM12 cited challenges that emanated from scanning, writing thus:

Scanning: when a question takes a small portion of a page the whole page was scanned making it difficult to read. Zooming in and out will be required for such questions, marking becomes slow. When a question takes a small portion of the page, only that portion should be scanned. Scanning the whole page would greatly reduce the picture quality.

SNR7 raised concern that ZIMSEC had no total control of the software, delaying resolution of challenges, explaining thus:

Maybe not having total control of the whole system presents not so much of a challenge but issues can quickly be dealt with had it been otherwise. Delays due to a difference in time with the UK guys and us possibly could not have been an issue.

The sentiments of SNR7 were echoed by the MAM earlier in 2013 (email, 18 May 2014) as indicated in Section 5.3.3.

The challenges that were enumerated by the examiners and SMs were echoed in the document review. Most of the challenges were a result of time constraints aimed at saving on costs. It was apparent from the data that ZIMSEC wanted to enhance efficiency of marking at minimum cost. Tight marking deadlines were set; the SMs pressured the examiners to mark and meet the deadlines; fast markers were allowed to mark as much as they could; SNRs responded by ignoring their quality control duties and concentrating on marking; and SMs might have adjusted quality control parameters to reduce chances of examiner stoppages by seeds. The scanning challenges, power cuts and erratic internet worsened the situation. The relationship between the ZIMSEC and the SWP did not help the situation.

Document review suggested that scanning challenges resulted in some missing scripts, and hence marks. It was reported that 179 candidates did not receive their results on time for the subjects marked on screen in the November 2013 examination (Memo, 29 April 2014). The ZIMSEC normally published O level examination results in February, but by the date of the memo the Council was still trying to locate scripts for the 179 candidates. Among the 179 candidates there were 11 who had not received results for 5008/3 and 5008/4. The author of the memo made comments about the missing script for each candidate. A comment about 5008/4 scripts read ‘...externally marked; no supplementary mark sheet; script not in crate....’ A comment for 5008/3 read, ‘...wrote paper 4....’ and another read ‘...not found; not queried....’ The memo provides evidence of chaos in the scanning bureau where scripts were managed. It was evident from the memo that most of the missing scripts either had additional answer sheets or had been killed, marked on paper (externally marked), and were never found. The Assistant Director TDR&E wrote on the memo, in long hand, instructing the MAM to organize aegrotation (a procedure for estimating marks) for the 179 candidates. Aegrotation could compromise the quality of the examinations marked on screen, leading to loss of credibility of the same.

The challenges in the bureau seemed to have been compounded by disunity among the ZIMSEC divisions. Document review indicated that some sections heads granted leave to personnel in their section at critical times (July 2013 E-marking Information Sheet). A statement on the sheet read thus:

...the resourcing of e-marking processes have been erratic and met with apathy from some sections. Where personnel from other sections have been integrated into the e-marking processes, some heads of sections have decided to unilaterally withdraw their charges midstream, creating gaps in the process. In some cases some sections go on leave at that critical point.....Using temporary staff has always been the last resort but the risk is that we end up farming out skills that should otherwise be invested in the system.

This statement insinuates that the OSM technology was not embraced by some key people in the institution and some of them might have decided to sabotage its use. This explains why some critical requirements, like use barcodes, were never made. Evidence abound that the operatives were keen to adopt and use the technology, especially the ER, the TDR&E and the IT departments. Their zeal was however killed by some heads of sections that were not keen about the technology. The SMs and IT personnel suggested the establishment of computer centres by the Council, as early as 2013, but the senior personnel were not keen about that as well (personal experience). On the whole, the ZIMSEC staff did what was possible in the context to enhance both the efficiency and quality of marking.

5.8 Summary of the findings

The practice of quality control in the onscreen marking of Ordinary Level Biology in Zimbabwe

The findings indicate that the quality of marking was largely influenced by the context in which Biology (5008) examinations were marked. The context was provided by an assessment framework, a technological infrastructure and human resources. The Biology (5008) syllabus provided the framework for the assessment of the course, spelling out the aims, objectives, the assessment scheme and the content. Two papers, 5008/2 and 5008/4 were marked onscreen for a period long enough to study the practice of quality control. The practical examination, 5008/3 was marked on screen once in the November 2013 examination and returned to PBM with SMs and examiners arguing that it was not suitable for OSM due to the use of supervisors' reports. The Biology (5008) examinations were therefore supposed to be designed and marked as guided the syllabus. The examinations conformed to the assessment scheme in terms of number of number of marks and the duration of the examinations. The skills weighting was, however, violated when some examinations were re-designed to suit the demands of the OSM technology, compromising the quality of the examinations.

The technological infrastructure included a scan centre; computers linked to the internet and electrified marking venues. ZIMSEC had established a scan bureau at the head office. The SWP loaned servers to the ZIMSEC in 2013 as the Council promised to buy their own servers. The

OSM technology was later delivered on cloud. ZIMSEC had no computers of their own, so they hired the machines from tertiary institutions. The computers were not adequate, so the subjects were grouped into marking sessions that would extend to January of the following year. The technological context was, therefore, inadequate to promote the practice of effective quality control.

The SWP made initiatives to train ZIMSEC staff on OSM in 2012. The scan bureau staff were trained to run the scanning process. The bureau was manned by temporary staff who were supervised by permanent officers. The SMs were trained to train examiners and as administrators. ZIMSEC staff requested to run the OSM of the June 2015 examination without on-site support from the SWP. There was evidence of planning for each OSM project with meetings being held to commission marking exercises. The marking exercises were thoroughly reviewed and lessons learnt reports compiled. There was evidence that the ZIMSEC did not follow the work schedules and plans for the OSM resulting in challenges that could compromise the quality of marking. There was also evidence that the SWP exercised too much control over the software, resulting in the ZIMSEC not able to resolve issues onsite. The ZIMSEC had to phone or email the SWP for resolution of problems that arose during scanning and marking. The quality of marking was, therefore, determined by the context. The practice of quality control was, therefore, influenced by the context provided by the assessment framework, the technological infrastructure and the human resources capacity.

Efforts were made to build the capacity of examiners in the OSM environment. There was evidence that examiners responded to newspaper adverts and applied to be selected for marking. The prospective examiners set for a test that was meant to gauge the subject expertise. The applicants were then trained to mark in general before sitting for a marking test. Successful applicants were recruited as examiners. The examiners were trained to mark onscreen at the marking venue where they went through marking in the training mode. The training was meant to familiarise the examiners with the icons and the questions. Some participants said the training was adequate and others said it was rushed and inadequate.

Senior markers attended pre-standardisation meetings where they discussed the marking schemes and marked dummy scripts that set the marking standards. All examiners attended standardisation (coordination) meetings where they discussed the marking schemes and practiced marking on dummy scripts. The examiners were required to mark and pass qualification seeds before they could mark live scripts. Quality of marking was monitored by seeds.

Seeds were used to monitor the quality of marking Biology (5008) examinations. It was established that the SMs set seed parameters by inputting the number of examiners, the number of candidates and the number of marking days. The SMs allocated scripts portions to examiners, with SNRs apportioned fewer scripts than normal markers to allow the former to supervise the latter. The SNRs set the seeds soon after pre-live marking. As the marking progressed, SNRs managed the examiners who were stopped by seeds. They could monitor the stopped marker window so as to identify stopped markers and invite them for discussions, or the stopped markers would approach them for discussions. It was however evident that it was usually the stopped markers who approached SNRs for discussions. The SNRs seemed too busy with their own marking load. Wrong seeds were identified and escalated to SMs as suspects. The SMs monitored the seed bank and would retire the suspect seeds. Retiring the seeds effectively deleted the marks awarded by SNRs and returned the scripts to the marking queue. The SMs could delete script portions marked by poor examiners, sending the scripts back to the marking queue.

There was evidence that SNRs could permanently stop examiners from marking questions if they were struggling to mark the questions. Some SMs insisted that the icon for permanently stopping examiners was clicked by error because bad examiners would not pass the qualification seeds. Another SM said the option could be used on examiners who partially attended the standardisation meeting and failed to master the mark scheme. Another SM said he would just delete the scripts marked by the examiner instead of permanently stopping them from marking a question. The SMs could monitor the progress and quality of marking through automatically generated reports and audit trails.

The examiners could escalate problem scripts to SNRs who would resolve or further escalate them to the administrator module where the SMs could view and resolve them. There was evidence that some examiners escalated scripts that they could mark using tools on their screens. The senior markers would return the scripts to the examiners with some comments or would just mark the scripts. Problems escalated by senior markers to SMs appeared as rescan request in the administrator module. The SMs would view and assess the scripts to make appropriate decisions. Genuine problem scripts were returned to base (to the scan centre) where the scripts were either rescanned or killed and marked on paper. The quality of marking was, therefore, monitored by seeds, automatically generated reports and escalation of problem scripts.

Test design issues greatly influenced the quality of marking Biology (5008) examinations. There was evidence that the quality of marking was influenced by the nature of questions and mark schemes. The Biology 5008/2 and 5008/4 question papers were redesigned to remove blank pages where candidates could possibly write their responses and risk losing marks when examiners did not see and mark the responses. Section B of 5008/2 had free response questions and candidates were provided with separate answer sheets in PBM. Answer spaces were provided on the question papers when the paper was introduced to OSM in 2015, bringing significant changes to the nature of questions. There was evidence that the questions became shorter in the June and November 2016 examinations. There were allegations that the OSM watered down the Biology 5008/2 examinations, with a SM calling for the return to the old system where candidates were provided with separate answer sheets. Participants agreed that only objective short answer questions should be marked on screen and that practical examinations could not be marked on screen without prejudicing the candidates.

The OSM technology, by its nature, could enhance the quality of marking Biology (5008) examinations. The participants enumerated several opportunities of the OSM to enhance quality of marking. Marks were automatically captured, eliminating addition errors; the system generated progress and quality of marking reports as well as audit trails that were used to monitor marking; the OSM was fast and efficient since there was no paper work; SMs could delete marks awarded by poor markers; the anonymous script portions reduced examiner bias

and eliminated examiner malpractice; the SMs could set and change quality control parameters; poor markers were identified and stopped by the seeds.

The participants identified the challenges that could compromise the quality of marking Biology (5008) examinations. The following challenges were cited: Power cuts; erratic internet; seeds frequently popping up until they could be recognised by examiners; differences in examiner payments; poor script images; control of the technology by SWP, resulting in ZIMSEC failing to solve problems on-site; answer spaces were not adequate for some candidates who then wrote their responses on inappropriate spaces; examiners were not able to see and mark responses on inappropriate spaces; additional answers were difficult to add onto mark scheme once marking started; and limited marking time. Despite the challenges, the data suggests that ZIMSEC staff did what was possible in the context to enhance the quality of marking Biology (5008) examinations.

5.9 Conclusion

The findings showed that the ZIMSEC adopted the OSM technology to enhance the efficiency and quality of marking at minimum cost. The quality of marking was influenced by the context in which the examinations were marked; the human capacity; the examiner monitoring mechanisms; and the nature of questions and mark schemes. The OSM technology had inherent opportunities to enhance the quality of marking by automation of processes such as mark capturing and addition, generation of reports and script management. Several challenges threatened to compromise the quality of marking. Major among the challenges were the time constraints created by the need to save money, high turnover of Biology SMs; the hesitation by ZIMSEC leaders to make milestone decisions and the subtle antagonism between the ZIMSEC and the SWP.

In the event that the ZIMSEC decides to resume OSM, they could establish computer centres in the provincial towns to gradually eliminate the cost of hiring computers and marking venues, reduce scanning challenges by using barcodes to identify scripts and train SMs to interpret and use the syllabus so as to preserve the validity of O Level Biology examinations. The SWP and the ZIMSEC could make effort to cultivate a mutual relationship that allows the latter to use the

OSM technology to enhance the efficiency and quality of marking at low cost while the former makes reasonable profit. On the whole, the OSM technology had the potential to enhance the quality of O Level Biology examinations in the Zimbabwean context.

The next chapter discusses the findings in relation to the conceptual framework that guided the study, related literature and personal experiences.

Chapter 6

Discussion of findings and proposed framework

6.1 Introduction

The previous chapter presented the analyses and findings of this study using narratives and tables. This chapter discusses the major finding of the study in relation to the themes that emerged from the research sub-questions. The scholarly literature which was reviewed in the study is used in a manner of cross-checking the findings against it. A sense is made of the conceptual framework which was used. Rival and competing explanations and interpretations are thoroughly interrogated. A model that can guide the practice of quality control in the OSM of Biology examinations is designed ultimately as the contribution of the study.

6.2. The context of OSM in Zimbabwe

The findings of this study indicate that the context in which Biology (5008) examinations were marked was provided by an assessment framework, a technological infrastructure and human resource. The next sub-section discusses the findings relating to the assessment framework.

6.2.1 Assessment framework

The Biology (5008) syllabus provided the framework for the assessment of the course, spelling out the aims, objectives, assessment scheme and content. The assessment framework was mainly deduced from a thorough review of the Biology (5008) syllabus and interviews with subject managers (SMs). Examiners' views were not gathered on the framework due to time constraints and logistical challenges. However, the data gathered were sufficient to conclude that there was a framework that guided the assessment of the Biology (5008) course.

The presence of an assessment framework for the Biology (5008) course is supported by several scholars. Taras (2009:58), advocating for an assessment framework, wrote that “a judgment cannot be made in a vacuum, and therefore points of comparison (i.e. criteria and/or standards) are necessary and in constant interplay”. The Biology syllabus (5008) provided clear criteria for

the examinations. The examinations could be judged against the assessment objectives, weighting and the assessment scheme as supported by Khan (2012:579), who concurs with Kapukaya (2013) when he posits that there is a need to develop clear criteria when analysing assessment information, and emphasises the need for the comprehensive criteria for setting and marking learners' work. The O level Biology examinations were therefore supposed to be set and marked according to the syllabus. The examinations were however redesigned and shaped by the demands of the technology as will be discussed later in this chapter.

The syllabus outlined three assessment objectives as presented in Chapter 5, Section 3.1. Objectives 1 and 2 were assessed in papers 5008/1 and 5008/2. This implied that some of the questions in the two papers were short while others were long, giving rise to the different types of mark schemes described in literature (Child et al, 2015; Ahmed & Pollit, 2013; Meadows & Billington, 2013). The influence of the different mark schemes are discussed later in this chapter. However, there seemed to be a wrong assumption about the practical skills (Objective 3) in the assessment framework. The candidates were required to register and sit for either 5008/3 or 5008/4. Therefore, there was an assumption that the two papers could equally assess practical skills. That could be contrary to Ghaicha's (2016:201) concept of educational assessment as a part of education where the learner achievement is appraised by collecting, measuring, analysing, synthesising and interpreting relevant information about a particular object of interest under controlled conditions. It was very unlikely that non-practical examinations could provide any information about the object of interest: the candidates' ability to follow instructions for practical work; plan, organise and carry out experimental investigations; select appropriate apparatus and materials for experimental work; use apparatus and materials effectively and safely. This was evidenced by one supervisor who wrote that *the school does not have a science laboratory. It is not capable of running such an examination.*

The syllabus designers must have realised that there were some schools that could not afford to buy equipment for practical examinations and equated the papers to avoid criticism that surrounds standardised tests. The assumption that the two papers measure the same skills could justify the concerns raised by Kandur (2017:<https://www.dailysabah.com>), who argues that standardised tests are unavoidably biased by social-class, ethnic, regional and other cultural

differences, illustrating the bias by the cartoon in Figure 2.1 in Chapter 2 Section 2.5, where different animals are instructed to climb a tree when the bird and the monkey are naturally adapted to climb trees while the fish and the penguin will never accomplish the task no matter how hard they try. Kandur further argues that standardised tests give unfair advantage to candidates who can afford test preparation. In this case, the tests were biased by social class, and only the well-to-do schools could afford to prepare learners for practical examinations. It could be possible that by assessing practical skills with non-practical examinations, ZIMSEC was asking different animals to climb different trees. That could explain why the alternative to practical paper was dropped out of the new curriculum where all the candidates are now expected to register and sit for practical examinations (Biology [4025] syllabus, 2015:41).

The existence of the alternative to practical paper (5008/4) shows that ZIMSEC could not administer practical examinations in all schools. This could be evidence that Zimbabwe, like other African countries, is still to recover from the impact of colonisation in spite of educational reforms that were meant to undo the racial disparities in the colonial education systems (Shizha & Kariwo, 2011; Bayat et al, 2014). The new assessment framework (Biology syllabus [4025]: 2015-2022) that abolished the alternative to practical option is still biased by social class because as the economy continues to nose-dive, 80% of families are living in poverty and they have cut down on educational spending, as reported by the Zimbabwe Council of Churches (ZCC) (Langa, 2019:4). Some schools might still not be able to afford equipment for practical examinations, bringing in adverse consequences for the fairness, and hence quality of marking the examinations on paper and onscreen.

Document review, face-to-face interviews and WhatsApp group discussions indicated that the two papers, 5008/2 and 5008/4 were marked on screen between 2013 and 2017. The use of the OSM technology to mark examinations that were delivered and written on paper confirm Pellegrino's and Quellmalz's (2010:120) assertion that the new technologies in public examinations have focused on logistical efficiency and cost reduction rather than improving the design of the same. This finding also confirms that ZIMSEC's examinations were nowhere near the first generation e-assessments predicted by Bennet (2015; 1998), where tests would be delivered to candidates via internet to computers, taken and marked on computers. As discussed

in Chapter 3 Section 3.3, the ZIMSEC has adopted technologies that automate existing processes, such as candidate registration, mark capturing and item authoring, without re-conceptualising them. These could be tentative steps towards Bennett's first generation assessment, as posited by Isaacs' et al (2013:42). On the whole, this study confirms Winkley's (2010:4) assertion that high-stakes CBT in the public sector education remains more the exception than the norm.

All data sources indicate that the practical examinations (5008/3) were marked onscreen once in November 2013 and returned to PBM. This finding indicates that there are some examinations that cannot be marked onscreen without compromising their quality. This is supported by Ofqual (2013:12), who indicates that UK examination boards used the OSM technology to some varying degree. By 2013 one examination authority had 88% OSM and 12% PBM while another had 13% OSM and 87% PBM. As discussed in Chapter 3, Section 3.5.2, the Hong Kong examination authority had a vision to replace PBM with OSM. However, the statement by Coniam and Yan (2016:1151), that "onscreen marking has been used for the majority of Hong Kong examinations since 2012" implies that there were a few exceptional subjects that were still marked on paper in Hong Kong. This finding, therefore, answers the question posed by Fowles (2011:5), who asked: *"can all existing paper-based examinations be marked on screen, or are there features which cannot be accommodated or which are too costly to accommodate?"* The findings of this study established that the supervisor's reports requirement for the 5008/3 could not be accommodated by the OSM technology. Some examiners, however, indicated that the practical examinations could be marked on screen, insisting that some examination boards were doing so without elaborating. However, ZIMSEC could consider marking practical examinations as whole scripts as indicated by scholarly literature.

Ofqual (2014a:4) cites a UK examination authority that marked an examination onscreen using the whole script approach, later marked at item level and resorted back to whole script marking. Although the reasons for shifting from one mode to another were not specified it is possible that there were quality concerns. Some scholars argue that whole script marking brings in examiner bias that could compromise the quality of marking (Ofqual, 2014a; Tisi et al, 2013; Chinamasa & Munetsi, 2012). The concerns about examiner bias existed in Zimbabwe as established by

Mashoko et al (2013:468). The OSM technology presents the opportunity to reduce examiner bias. Odendahl (2011:141) enumerates some strategies for improving fairness and reducing bias at scoring of scripts, which are: removing names or the types of identity information from responses; distributing responses randomly to examiners; specifying the scoring criteria in written guidelines; illustrating the scoring criteria with sample responses; training examiners; and monitoring examiners.

Removing names or identities from 5008/3 scripts poses a challenge that threatens the quality of marking. These strategies can only be employed when papers are marked at item level. With particular reference to OSM, Pinot de Moira (2013:13) states that no empirical evidence has been presented to justify the claim that item level marking eliminates examiner bias. However, research studies conducted later than 2013 provided empirical evidence to that effect (Ofqual, 2014a:2). The quality concerns in the 5008/3 examinations seemed to have outweighed the bias concerns, hence the return to PBM. Whole script marking allows the examiners to see the candidates' details and still be able to use the supervisors' reports in the OSM environment. Literature shows that OSM offers some logistical benefits such as automatic addition of marks as well as frequent and flexible monitoring of examiners (Pinot de Moira, 2013; Ofqual, 2013; Ramakrishna et al, 2012). The ZIMSEC, therefore, could have opted to mark whole scripts on screen instead of returning the practical examinations to PBM.

A review of the Biology (5008) syllabus indicated that there were some contradictions in weighting of the assessment objectives. The contradictions were not carried over to the new curriculum as alleged by one SM. It was however noticed that the new Biology (4025) syllabus put even more emphasis on the most basic skills of knowledge and comprehension. Where the skill was weighted at 55% in papers 1 and 2 in the old curriculum (5008), it is now at 60% in the same papers in the new curriculum (4025). It was also noticed that the practical examination is still made of two questions worth 40 marks but has been given more time. The paper was one hour long in the old curriculum, but it is now one and half hours long. This could unnecessarily dilute the examinations in the new curriculum as confirmed by research. In a research on marker effects on marking reliability, Baird et al (2013:14) established that a UK examination authority increased the duration of a Geography examination by 25% leading to candidates scoring higher

marks. To maintain standards, grade boundaries were adjusted accordingly. There is need to investigate how the Biology (4025) syllabus compares with the 5008 syllabus to establish if there was a real curriculum review or a superficial one that could lead to lowering of standards in biology education.

The context was also provided by the technological infrastructure.

6.2.2 Technological infrastructure

The findings established that the technological infrastructure included a scan centre, computers linked to the internet and electrified marking venues. ZIMSEC had established a scan bureau at the head office. The SWP loaned servers to the ZIMSEC in 2013 as the Council promised to buy their own servers but the OSM technology was later delivered on cloud. The scripts were scanned and saved onto the script marker and then exported to the e-marker where examiners marked them on computer screens. ZIMSEC had no computers of their own, so they hired the machines from tertiary institutions. The computers were not adequate, so the subjects were grouped into marking sessions that would extend to January of the following year. The marking was interrupted by power cuts and erratic internet.

It was reported that ZIMSEC was the first examination authority to use OSM in Africa (Coniam 2016; DRS 2013; Kachere, 2012; Karombo, 2012) and had evidently not prepared the appropriate infrastructure for the technology. As discussed in Chapter 1, Section 2.3, ZIMSEC adopted the OSM technology at a time when there were limited fiscal resources committed by government to support ICT access and use; inadequate communication infrastructure with patchy internet; inadequate commercial power, with a significant population resorting to alternative power sources that were expensive; and low digital literacy (Shafika, 2007; National ICT Policy 2015). This meant that even the government could not support ZIMSEC to establish the infrastructure for the OSM technology because of limited capital investment in ICT. Mr. John Maramba, the OSM project manager, even appealed to government to help the ZIMSEC acquire a minimum of 1000 computers to enable the expansion of the OSM project (Bulawayo24 News, 2012: <https://bulawayo24.com/index-id-news-sc-education-byo>). The poor technological infrastructure compromised the quality of marking as indicated by findings from all data sources.

This study established that the Zimbabwe context is different from others where the OSM technology had been used as confirmed in literature. The Hong Kong government sponsored the development of ICT infrastructure by giving the HKEAA US\$25 million in 2005, resulting in the establishment of three OSM centres that were ready for use in 2012 (Coniam, 2011a:1044). According to Coniam and Yan (2016:1154), the number of assessment centres increased later on because these authors claim that by the time of their study there were ten marking centres where markers could access examination scripts via intranet and mark them on computer screens. Examiners would travel to the centres and book three-hour marking periods. Some participants had problems with the travelling that limited marking time while others had no problems with the arrangement (Coniam, 2011a:1040). In the UK, examiners marked from their homes and were concerned about broadband bills rather than availability of the same (Raikes et al, 20014:18-29). In the UK and Hong Kong contexts the technological infrastructure was likely to enhance quality of marking.

The literature shows that quality and cost efficiency were the main reasons why the examination authorities in the UK adopted the OSM technology (Ofqual, 2013; Pinot de Moira, 2013; Haggie, 2008). However, the quality has to be enhanced at reasonable costs. There was evidence that the ZIMSEC incurred unexpected costs in the marking of unconstrained examinations and resorted to constraining all examinations marked on screen as established in this study. The hiring of computers could have possibly escalated the cost of running the OSM technology for the ZIMSEC. There was no evidence that the ZIMSEC conducted a cost benefit analysis before adopting the OSM technology, as was done by the Assessment Qualifications Alliance (AQA)'s finance department before adopting the technology (Fowles, 2011:2).

ZIMSEC could have, however, reduced the cost of OSM by establishing their own marking centres in the ten provinces where they have offices (www.zimsec.co.zw) to eliminate the cost of hiring computers. The centres could have been established gradually, as done in Hong Kong. The examiners would then mark from their home provinces with the ZIMSEC paying for accommodation and meals. Some examiners who stay in the provincial towns would probably opt to stay at their homes, reducing accommodation costs for the Council. The challenges that

emanated from the context influenced the training and standardisation processes as well as the actual marking as discussed under the appropriate headings.

The human resource capacity at ZIMSEC also contributed to the context in which Biology (5008) examinations were marked on screen.

6.2.3 Human resource capacity

There was evidence that the SWP made initiatives to train ZIMSEC staff on OSM in 2012. The scan bureau personnel were trained to run the scanning process. The SMs were trained to train examiners and as administrators. The training is supported by Hill (2013:20), who argues that in order to assure the quality of standardised tests, modern examination authorities should pay attention to the training of examinations personnel. However, the recruitment and training of temporary personnel in every examination session for the scanning bureau could have compromised the quality of marking as evidenced by rescan requests escalated by SNRs and SMs.

The findings of this study established that ZIMSEC put in place some quality control measures presented in Chapter 5 Table 5.4, taking cognisance of the high stakes nature of the examinations they offered and put in place quality control procedures as supported by scholarly literature. The Government of South Australia (2016:3) posits that script marking is an important aspect in examinations and has three major purposes, which are to ensure the consistent application and interpretation of assessment performance standards in the subject; that scores awarded to candidates across examination centres are fair and comparable; and that results are valid and reliable. This is supported by Hill (2013:19), who posits that because examination results are the main, if not the only basis for making high-stakes decisions, the results should always be accurate and error free. The Zimbabwean Examinations Authority and the SWP therefore made some effort to ensure that the purposes of marking are accomplished in the OSM environment. Hill (2013:20) also argues that assuring quality in examinations guarantees public confidence and goes on to enumerate quality assurance strategies that modern examination authorities should adopt—creating a culture of assuming responsibility for improving quality; establishing the effective system internal controls; and automating the examination processes to eliminate

human error. Internal controls were in the form of plans, review meetings, marking commissioning meetings and reports for the lessons learnt. The quality control procedures put in place by ZIMSEC therefore had plenty of opportunities to enhance quality of marking in the OSM environment.

However, evidence abound that there was no culture of assuming responsibility for the quality control activities as shown by failure to stick to the plans and work schedules. As discussed in Chapter 3 Section 5.4, the ZIMSEC directorate expected that the OSM technology would quicken the delivery of results so much that November results would be published in January (The Herald, 2015; DRS, 2014; Bulawayo24, 2012; The Chronicle, 2012; Kachere, 2012; Karombo, 2012). The findings show that OSM extended to January because of the failure to stick to work schedules, coupled with inadequate computers and marking venues. This means that the internal controls advocated for by Hill (2013:20) were not effective, compromising the quality of marking.

There was evidence that the SWP exercised too much control of the OSM technology. This could be attributed to three reasons: ZIMSEC seemed reluctant to take responsibility for quality control processes, so the SWP took the responsibility; they failed to appreciate the uniqueness of the ZIMSEC context; and they could lose business if they lost cost control of the technology. It was evident from the data that the SWP always had to push ZIMSEC when it came to planning, implementation of the plans and decision making. There was a subtle antagonism between ZIMSEC and the SWP that could be discerned through the findings, with major setbacks on the quality activities. The ZIMSEC staff suspected that the SWP wanted to make money at the expense of efficiency and quality of marking, hence the requested to run the OSM of the June 2015 examination without on-site support from the SWP. The SWP provider and ZIMSEC needed to engage each other in a way that maximised the benefits of the OSM technology as argued by Ramakrishna et al (2012:15) who posit that without careful engagement of appropriate stakeholders the benefits remain theoretical.

It can be concluded that the quality of marking Biology (5008) examinations was influenced by the context provided by the assessment framework, the technological infrastructure and the human resource capacity at ZIMSEC.

6.3 Sub-question 1: Capacity building of examiners

The first research sub-question was *“How does training of examiners and standardisation activities influence the quality of marking O Level Biology in the onscreen marking environment?”*

This section discusses the findings in relation to this sub-question, the conceptual framework and scholarly literature, starting with the recruitment and training of examiners.

6.3.1 Recruitment and training of examiners

There was evidence that examiners responded to newspaper adverts and applied to be selected for marking. The prospective examiners set for a test that was meant to gauge the subject expertise. The applicants were then trained to mark in general before sitting for a marking test. Successful applicants were recruited as examiners. The examiners were trained to mark onscreen at the marking venue where they went through marking in the training mode. The training was meant to familiarise the examiners with the icons and the questions. The OSM training was however hurried, leading to some challenges that compromised the activities used to monitor the quality of marking.

The recruitment and training of examiner teams for the two Biology (5008) components marked on screen between 2013 and 2017, well before marking, were in line with international standards as supported by the conceptual framework and scholarly literature. The Biology (5008) examiners were practicing teachers, in line with recruitment requirements of some international examination authorities (Ofqual, 2014c:6). Examination boards around the world train examiners on two occasions, well ahead of marking to build a pool of examiners and just before marking to ensure consistent application of the mark schemes (Ofqual, 2014e; 2013). The examiners who participated in this study were part-time staff drawn from the pools built for 5008/2 and 5008/4 examinations. This is also supported by scholarly literature which shows that it is standard

practice for the examination boards to recruit part-time staff to mark the examination scripts and examination boards need to provide comprehensive training programmes for the examiners (Hill, 2013; Ofqual, 2014c). The recruitment and training of the examiners could enhance the quality of marking Biology (5008) examinations.

The findings, however, showed that the OSM training was conducted concurrently with live marking of examinations. Grossman and Salas (2011:16) argue that the transfer of training relates to trainee characteristics such as cognitive ability, motivation and perceived usefulness of the training, training design, and the work environment. The scheduling of training during live marking could have compromised its quality and perceived usefulness by the examiners and SMs, leading to the later compiling guiding notes from practice. This is contrary to training practices in Hong Kong where examiners were trained to work in the OSM environment at the time of selection and recruitment (Tze Ho & Chong Sze, 2013:4). Training the examiners in the OSM environment to build examiner teams could enhance the quality of marking Biology (5008) examinations.

In a research study described in detail in Chapter 3 Section 3.4.1, Falvey and Coniam (2010:1) conducted a qualitative study to gauge the responses of the English language raters to the OSM and PBM in Hong Kong. All of the 17 participants indicated that they possessed the right technical skills to work in the OSM environment. This is contrary to Zimbabwe where the National ICT Policy (2015:15) confirms that there was low digital literacy. The low digital literacy influenced the marking of Biology (5008) examinations as indicated by some participants. One examiner said that the SNRs set wrong seeds because they clicked wrong icons. The SM for the subject said they reduced the teaching experience requirement from five to three years so as to capture younger markers who were technologically apt. Another SM said that the SNRs for subjects other than Biology were demoted because of technological incompetence. It can therefore be concluded that the timing of the OSM training coupled with the low digital literacy of some examiners reduced the transfer of training and hence the quality of marking.

This study established that the Biology (5008) examiners were not categorised according to skills as indicated in the conceptual framework. The conceptual framework of this study indicates that

quality of OSM is enhanced by recruitment and training of a pool of examiners made up of different examiner types: clerical, graduate and expert markers. Clerical markers are trained and standardised markers who have little or no knowledge of the subject and are proficient in the language in which the scripts are written and should possess the adequate IT skills. Graduate markers are trained and standardised markers who are recent graduates or post graduate learners in the subject. Assistant examiners/expert markers are trained and standardised markers who have experience in the performance of candidates at the level of the examination in addition to language proficiency, ICT skills and subject expertise (Ofqual, 2013; Raikes et al, 2004:).

The research study conducted in the UK by Suto and Nádas (2008:9) discussed in detail in Chapter 2 Section 2.7, justified the categorisation of examiners in the OSM environment. The study established that questions that demanded simple cognitive marking strategies were marked with high accuracy by all examiner categories. The questions demanding complex cognitive marking strategies were marked with less accuracy by all examiner categories, with experts being the most accurate and non-graduates being the least accurate. A similar research conducted by Meadows and Billington (2013:9) established that all marker groups generally marked accurately. There were some undergraduates who marked as well as the best GCSE examiners. These studies provide evidence that categorisation of examiners does not compromise quality of marking examinations but saves time and subsequently reduces the marking period and attendant costs.

Document review in this study indicated the OSM technology could automatically mark some items in the examination. This is supported by Baired, Hayes, Johnson et al (2013:12) who revealed that a UK examination authority marked a Sociology paper where Section A was computer marked, and Sections B and C were shared between expert and general markers, depending on the need for subject expertise required for each question or part question. There was evidence that all Biology (5008) examiners marked all types of questions, with some participants complaining that fast markers targeted easy questions, leaving the difficult ones to slow markers. This is contrary to the conceptual framework and scholarly literature. The Council did not select items for automatic marking. The conceptual framework for this study illustrates key concepts involved in marking by humans and therefore does not include automatic marking.

It can be argued that automatic marking reduces the marking load for human markers in the OSM environment. The categorisation of examiners coupled with automatic marking could have greatly reduced the marking period, saving time and money. It can be concluded that ZIMSEC neither adapted recruitment procedures to the OSM environment nor exploited automatic marking, resulting in the Council's failure to fully exploit the affordances of the OSM technology to their benefit.

Examinations are marked to meet specific timelines (Ofqual 2013; DRS 2013), exerting pressure on both examiners and their supervisors to mark and meet the deadline. The findings of this study confirmed that Biology (5008) examinations were marked to meet tight deadlines, resulting in some hurried training and standardisation meetings; challenges such as erratic internet and power cuts disrupted marking; marking extended to January of the following year, threatening to delay the release of examination results. The conceptual framework on Figure 2.3 indicates that quality of marking is influenced by activities such as training, pre-standardisation, standardisation meetings, the marking of practice live scripts, approval of markers and the re-training of errant markers by senior markers. These activities were evidently rushed through in ZIMSEC context.

The findings showed that there were hierarchies within the examiner teams, with the principal marking supervisor (PMS), several deputy principal marking supervisors (DPMS) and belt marking supervisors (BMS), collectively named SNRs. The SNRs set marking standards and supervised NMs. The hierarchies were existent in other examination authorities as indicated by scholarly literature (Ofqual, 2014e; 2013; Tze Ho & Chong Sze, 2013). This is also inline with the conceptual framework which shows that the SNRs ensure the quality of marking by monitoring and retraining examiners during marking. The hierarchies in the examiner teams for 5008/2 and 5008/4 had the opportunity to enhance the quality of marking in the OSM environment.

However, the PMS played a distinct role in the actual OSM of examinations in the UK, whereas the PMS in this study had the same roles as DPMSs and BMSs and they were collectively known as SNRs. This enabled me to protect the privacy of the participants to some degree. To avoid

deductive disclosure, Saber and Ben-Yehoshua (2017:413) advise that the researcher should avoid detailed description of the participants. Detailed description of the PMS in the data could have led to deductive disclosure. All SNRs had the same role in the OSM environment, eliminating the need to describe any particular supervisor. The readers of this study are unlikely to identify and assign responses to the PMS, DPMS or BMS because they are all called SNRs. The quality of marking was also enhanced by standardisation meetings.

6.3.2 Standardisation meetings

The findings showed that SNRs attended pre-standardisation meetings where they discussed the marking schemes and marked dummy scripts that set the marking standards. The standardisation meetings held are supported by the conceptual frame, which indicates that the pre-standardisation meeting is always conducted face-to-face while the standardisation meeting can be either face-to-face or online. Ofqual (2014e:5) insists that the pre-standardisation meeting is always conducted face-to-face probably to maximise the interaction among the examiners. The face-to-face meetings were the best for the ZIMSEC context where internet was patchy and the digital literacy was low.

Baired et al (2013:10) state that the pre-standardisation meeting confirms the marks pre-awarded by principal examiners to a sample of candidates' responses for use in monitoring marker performance throughout the marking period, emphasising that the pre-allocated marks are known within UK bodies as 'true scores'. This means that seeds were set at the pre-standardisation meeting, discussed and agreed upon by the principal examiners. There was evidence that seeds were not set at the pre-standardisation meetings for Biology (5008) examinations, neither were they set at the standardisation meeting. The seeds were set soon after the two standardisation meetings. It is therefore apparent that the seeds were not discussed during standardisation, leading to a myriad of challenges in the seed system as discussed later in this chapter. Instead of marking dummies only, the SNRs should have set the seeds and discussed them as well. It can be argued that the Biology (5008) SNRs set marking standards that were not relevant to the practice of quality control in the OSM environment when they attended the two standardisation meetings, compromising the quality of marking.

The Biology (5008) standardisation meetings were held soon after the pre-standardisation meetings and were attended by all examiners. The activities of the standardisation meetings were supported by Ofqual (2014e:7), who describes similar activities for face-to-face standardisation meetings for major UK boards, emphasising that examiners mark some practice scripts and are assessed by the team leaders before they can mark live scripts. Baired et al (2013:11) confirm that UK examination authorities used either online or face-to-face standardisation methods, with Ofqual (2014b:4) arguing that the method of standardisation does not influence marking reliability although examiners prefer the face-to-face method, purporting that it builds a community of practice. The Zimbabwean context, discussed in Chapters 1 and 4, confirmed by the findings of this study, would not allow online standardisation, leaving face-to-face standardisation as the only option. As shown in the conceptual framework, the team leaders have to approve the marking of examiners before allowing them to mark live scripts. The findings of this study indicated that the practice scripts were replaced by qualification seeds, meaning that the examiners were automatically trained and approved.

The participants in this study concurred that the standardisation meetings ensured the consistent application of the mark schemes and could enhance the quality of marking as supported by scholarly literature. Baired et al, (2013:10) concur with Ofqual (2014b:59), that standardisation ensures that examiners are fully competent in applying the marking scheme consistently before they begin marking and has several purposes, which include providing a context within which the marking process takes place; defining the task to be performed by the examiners; marking and discussing sample responses; discussion of the marking scheme; and confirmation of the marking scheme. The standardisation meetings conducted by ZIMSEC could therefore enhance the quality of marking in the OSM environment. These results confirm that standardisation influences the quality of marking as established by some studies reviewed by Tisi et al (2013:27), and refute the findings in the same review, that examiners who went through standardisation marked as accurately as those who did not, concluding that examiners marked accurately when they were provided with the mark scheme only.

Some participants indicated that the standardisation time for Biology (5008) was adequate while others indicated that it was inadequate, arguing that the discussions did not exhaust all possible

answers. Similar results were reported by Falvey and Coniam (2010:14) in a study that was discussed in detail in Chapter 3, Section 3.4.1. Some of the English raters who participated in the study said they had received satisfactory training at the standardisation meeting, while others said that the training was rushed with SNRs, who were under pressure, dictating marks and comments instead of allowing examiners to work through the practice scripts. There were also concerns that the qualifying scripts were too few (only four), with some examiners arguing that they went through the standardisation without mastering the marking standards (Falvey & Coniam 2010:14). Despite the differences in the Hong Kong and Zimbabwean contexts, the similar results indicate that training and standardisation meetings were rushed in order to meet marking deadlines. Some examiners still mastered the marking standards in this study and in Falvey's and Coniam's (201:14) study.

These results can be explained by Grossman's and Salas's (2011:16) argument that the transfer of training relates to trainee characteristics such as cognitive ability, motivation and perceived usefulness of the training; training design; and the work environment. The participants who thought that the standardisation time was adequate probably had the cognitive abilities to grasp the demands of the marking standards, as opposed to those who thought time was inadequate. It could also be that the working environment compromised the transfer of training for the examiners who thought the standardisation time was not adequate. The standardisation time could have been reduced to create time for OSM training that ran concurrently with live marking, exerting pressure on some examiners.

However, in this study the examiners who thought time was inadequate raised an important issue. They said that some correct answers were missed at standardisation meetings and could not be added onto the mark scheme once marking started. This could be true given that the SNRs set seeds at a time when they could not discuss and approve the correctness of the seeds. I followed up with the SMs but could not get evidence that such answers were added to the mark schemes. The quality of marking could be compromised if the standardisation meetings do not set seeds and the OSM technology renders editing of the mark schemes impossible.

From this discussion, it can be concluded that ZIMSEC put in place recruitment, training and standardisation procedures that could enhance the quality of marking in the OSM environment. However, the OSM training was conducted concurrently with live marking, exerting pressure on the SMs and the examiners and compromising the quality of marking. There is a need for ZIMSEC to align the activities of standardisation meetings to quality control in the OSM environment and to explore the editing of mark schemes to include correct answers that may still be missed at standardisation. The findings sufficiently answered the first research sub-question. The next section discusses finding on the second research question.

6.4 Sub-question 2: Monitoring the quality of marking

The second research sub-question was “*How is the quality of marking O Level Biology monitored in the OSM environment?*” The research findings indicated that there were several ways of monitoring the quality of marking in the OSM environment, with qualification and seeds as the main approach to quality control.

6.4.1 Qualification and seeds

This study established that the quality of marking was monitored by SNRs as supported by the conceptual framework. In PBM, the quality of marking was also monitored by the SNRs who moderated the scripts marked by examiners. A team leader was required to sample and moderate at least ten percent of scripts in an envelope to ensure adherence to the mark scheme and, hence consistency and accuracy of marking (Mashoko et al, 2013; Bukenya, 2006). The SNRs therefore monitor quality of marking in PBM and OSM. This confirms that OSM is only meant to automate existing assessment processes without reconceptualising them as supported by literature (Ramakrishna et al, 2012; Roan, 2009).

According to the conceptual framework on Figure 2.3, there are three mechanisms of moderating scripts in the OSM environment; seeding, double percentage marking and backreading. This study established that ZIMSEC used percentage double marking for some examinations and seeds for Biology (5008) and other examinations. The three approaches have been used by some examination authorities, percentage double marking (UK) or double marking (Hong Kong), seeds and back-reading (DRS, 2015; Ofqual, 2013; Pinot de Moira, 2013; Tze Ho & Chong Sze,

2013; Hudson, 2009). There was no evidence that ZIMSEC was aware of double marking and back reading. The training materials identified and reviewed in this study were on percentage double marking, seeds and the S-process, which was never used by ZIMSEC. The seeds were used as the quality control approach to marking Biology (5008) examinations. This is supported by the literature which shows that the seed approach to quality control is appropriate for constrained examinations, where candidates respond to short answer questions on spaces provided on the question paper (DRS, 2015; 2013; Hudson, 2009). The document review, however, showed that Section B of 5008/2 was originally unconstrained, with candidates responding to long questions on the separate answer sheets, but was constrained in 2015 when the paper was first marked on screen. It can be therefore concluded that seeds were not appropriate for Section B of 5008.

The Biology (5008) examiners marked qualification seeds before they could mark live scripts. The document review and face-to-face interviews indicated that the qualification replaced the first ten live scripts that were used to assess accuracy of marking before the examiners could be allowed to mark live scripts. The first ten live scripts marked in PBM are supported by scholarly literature which shows that the examiners will only be allowed to begin the actual marking when the team leaders are satisfied with the accuracy of marking the sample scripts (Ofqual, 2013; Tze Ho & Chong Sze, 2013), the stage referred to as approval on the conceptual framework. The qualification seeds are the approval scripts in the OSM environment.

It was however established that all 5008/2 examiners, including the SNRs, failed qualification seeds in the November 2015 examinations, with the PMS for the Paper 2 calling for more pre-standardisation time. As discussed earlier, SNRs did not set and discuss the seeds at any of the standardisation meetings, implying that no standards had been set for the approval process. Increasing pre-standardisation time was unlikely to improve the quality of qualification seeds used for approval. It can be argued that unless seeds are set, discussed and approved at the pre-standardisation meeting, the approval process in the OSM environment would not serve its intended purpose.

This study established that the SMs set seed parameters by inputting the number of examiners, the number of candidates and the number of marking days, and allocated scripts to examiners. This is supported by Ramakrishna et al (2012:16), who list the functionalities of the administrator module and illustrated in Figure 3.2. The setting of the parameters in the OSM system eased the logistical challenges associated with script moderation in PBM. The challenges include the following: it is the examiner who selects scripts for the team leader at fixed intervals, resulting in possibilities of examiners being thorough when marking scripts for moderation; the examiners continue to mark in the intervening periods without feedback from team leaders; and there are logistical challenges of moving scripts between examiners and team leaders, limiting the frequency of script moderation (Johnson & Black 2012; Hudson 2009). In the OSM environment the SNRs assumed the role of identifying stopped markers and discussing marking points with them. The OSM technology therefore had the opportunity to enhance the efficiency and quality of monitoring the accuracy of marking.

It was however surprising to note that only 5% of the scripts were used as seeds compared to 10% moderated in PBM. Some Biology (5008) examiners in this study suspected that there were times when there were no seeds in the system. In a mathematical simulation, Pinot de Moira (2013:1) established that a quality control system that includes any element of sampling cannot directly influence the marking reliability, but can only identify the deviating markers for retraining, and that percentage double marking has higher chances of picking the deviating markers than seeding. In view of these conclusions, it can be argued that the seed parameters set at 5% for Biology (5008) had very low chances of identifying poor markers, hence the examiner's suspicion that there were times when there were no seeds.

There was evidence that the SMs could adjust the seed parameters, as done with the qualification for the November 2015 5008/2 examinations, but there was no evidence of any policy that regulated thresholds for the parameters. In the absence of such a policy, the SMs could adjust the parameters to even lower limits, compromising the quality of marking. The same seed percentage of 5% was used to mark Liberal Studies in Hong Kong. The examiners who participated in Coniam's (2011a:1049) study raised concern that the seeds were too few, but they commended the seeds saying that they kept markers on track during marking. Baired et al

(2013:12-13) reveal that a UK examination authority marked Geography and Psychology at a seeds rate lower than 0.2% for three years, presenting two seeds to a marker for every 100 actual script portions. The Ofqual (2013:17) confirms that seeds are usually set at 5% but cites example of two boards that set seeds at 10%. This is evidence that seed rates can be adjusted to levels so low that they cease to serve the purpose of identifying poor markers.

Some participants in this study thought that a seed rate higher than 5% would be better for effective quality control. The document review however indicated that increasing the seed percentage would increase the marking load for examiners since more seeds would be presented to them for marking, extending their stay at the marking venue and increasing the cost of hiring computers, accommodation and subsistence allowances. Therefore, a seed percentage higher than five would have increased the cost of marking Biology (5008) examinations in the onscreen environment. This dilemma was acknowledged by Pinot de Moira (2013:6), who argues that existing systems seem to represent the best compromise between two conflicting imperatives - statistical robustness and practical viability - where statistical robustness is the lesser partner.

However, ZIMSEC could save time by utilising the ability of the OSM technology to automatically mark some items and distribute others to clerical and expert markers. A marking rate analysis indicated that OSM is 15% faster than PBM (Fowles, 2011:7). The time so saved could then be used for marking the load increased by a higher percentage of seeds.

The findings of this study established that some seeds were identified and marked in such a way that the examiners for Biology (5008) would not be stopped even if the seeds were wrong. This is contrary to scholarly literature that seeds should be presented to examiners at a pre-determined rate and examiners do not know when they are marking seeds (Baired et al, 2013; Pinot de Moira, 2013). The participants said that the wrong seeds were identified when SNRs discussed them with stopped markers, flagged as suspects and deleted by the SMs. The deletion of suspect seeds was an effective way of dealing with wrong seeds. Now that the examiners had found a way of cheating the quality control system, the chances of identifying wrong seeds were reduced. Some examiners could therefore get away with poor marking. Given that the senior marker's mark is considered as the 'true mark' (Baired et al, 2013; Pinot de Moira, 2013), it follows that

suspect seeds prejudiced candidates whose scripts were set as seeds. This supports my argument that SNRs should have set, discussed and approved qualification and seeds at the pre-standardisation meeting to reduce suspect seeds.

The results of Pinot de Moira's (2013:1) study indicate that the effectiveness of quality control by seeds depends on the action taken on the deviating examiners who have been stopped from marking. It was worrisome to note from the findings of this study that some SNRs did not train stopped markers as required for effective quality control, and that there was no mechanism of enforcing discussions between stopped markers and the SNRs. The findings are similar to Johnson's and Black's (2012:400) results of a study on monitoring of marking. The study established that as team leaders in the research monitored examiners, some gave feedback that guided examiners throughout the marking period, while others gave meaningless feedback that left examiners confused. The other team leaders did not give any feedback to examiners throughout the marking process, resulting in examiners feeling isolated. The conclusions made by the researchers implied that there was no automatic feedback provided by the OSM system, so the SNRs were expected to provide it. However, the OSM system in Hong Kong seemed to generate statistics that related to the quality of marking. In a study described in Chapter 2, Section 2.9.3 (Coniam, 2011a:1045), an examiner noted that they were not adequately trained to access and interpret their marking statistics. There was no consensus about the feedback provided by the OSM system to Biology (5008) examiners in this study. Some SMs and examiners said the system did not provide any feedback about the quality of marking, while one subject manager and an examiner indicated that the system provided feedback to examiner. A conclusion could not be made about the issue.

In a research on monitoring of marking quality in the OSM environment, Johnson and Black (2012:400) established that discrepancies between examiners' marks and definitive marks on seeds emanated from (i) the examiner awarding a wrong mark, (ii) grey areas in the mark schemes, where the answers were open to examiner interpretation and (iii) wrong marks on seeded scripts. The researchers conducted their study for a UK examination authority that set, discuss and approve seeds at pre-standardisation meetings, but SNRs still set wrong seeds. The situation could be worse in the ZIMSEC context where seeds for Biology (5008) examinations

were set at a time when SNRs could not discuss and approve them. The setting of wrong seeds by SNRs indicates that they had not fully mastered the mark scheme and so had set wrong standards for all examiners. The wrong marks awarded to scripts by stopped examiners indicated that there were grey areas in the mark scheme. It can be concluded that the seed approach to quality control offered the opportunity to enhance quality but was not properly used in the marking of Biology (5008) examinations.

In addition to the seeds, there were other mechanisms of controlling the quality of marking Biology (5008) examinations.

6.4.2 Other mechanisms of monitoring the quality of marking

Besides seeds, the quality of marking Biology (5008) was also monitored by automatically generated reports and audit trails that could only be viewed in the administrator module. This is contrary to Ofqual (2013:17), who posits that one of the benefits of OSM is that it enables continuous, real-time monitoring, with the SNRs being able to view examiners' marking on screen so as to monitor the speed at which examiners mark. The SNRs for Biology (5008) worked in the senior marker module and therefore could not access reports in the administrator module. However, there was evidence that the SNRs could permanently stop examiners from marking questions if they were struggling to mark the questions. The SNRs could possibly identify such markers from the frequency at which they were stopped by the seeds set for the particular question.

It was however established that only one marker was stopped in the November 2015 5008/2 examinations that were characterised by quality control challenges. Some SMs insisted that the icon for permanently stopping examiners was clicked by error because bad examiners would not pass the qualification seeds. As discussed earlier, the Biology (5008) SNRs seemed to be busy with their own marking load and probably did not count the number of times individual markers were stopped by seeds of the same question. It can be concluded that the Biology (5008) SNRs did not fully utilise the 'permanently stopped' functionality to control the quality of marking probably due to low digital literacy, coupled with inadequate training.

This study established that the quality of marking Biology (5008) examinations could also be monitored by escalation of problem scripts. The problem scripts that could not be marked on screen were killed and marked on paper. The killed scripts could easily be misplaced, leading to the missing marks reported in the November 2013 examinations. There was evidence that some examiners, both SNRs and NMs, escalated the scripts that they could mark using tools on their screens. This could be attributed to either inadequate training or low digital literacy that prevailed in Zimbabwe at the time. None of the literature reviewed in this study reported on escalation of problem scripts as a way of monitoring the quality of marking. This study, therefore, contributes new knowledge to the practice of quality control in the OSM environment which is accounted for at the end of this chapter.

6.5 Sub-question 3: influence of question papers and mark schemes

The third sub-question was *“How do O Level Biology mark schemes influence quality control in the OSM environment?”*

As indicated on the conceptual framework, the nature of examination questions and mark schemes influence the quality of marking. The findings of this study established that the questions in 5008/4 and Section A of 5008/2 were mainly short and elicited short responses. The examiners, therefore, mostly used System 1 strategies of marking described by Child et al (2015:8), involving scanning for simple items and is almost effortless. This is supported by the literature which shows that highly structured questions are marked more reliably than open ended questions (Ofqual, 2013; Meadows & Billington, 2005). Tisi et al (2013:21) posit that marking reliability can be increased by increasing item constraint; highly specified mark schemes; and lower maximum marks, among other factors. Therefore, 5008/4 and Section A of 5008/2 were marked accurately, enhancing the quality of marking.

However, according to the assessment framework, questions in Section B of 5008/2 were supposed to elicit long responses that required examiners to use System 2 marking strategies described by Child et al (2015:8), which involve scanning for complex items, evaluating and scrutinising, and requires some effort. Shortening the questions increased the accuracy, hence the reliability of marking. The literature, however, shows that validity and reliability are inversely

related; increasing one reduces the other (Tisi et al, 2013; Ofqual, 2013; Meadows and Billington, 2005). By shortening questions in Section B of 5008/2, ZIMSEC increased the reliability of marking the paper and reduced its validity.

The findings of this study indicate that mark schemes for 5008/2 and 5008/4 were a combination of objective and point based or semi-constrained types. The objective mark schemes arise when there are short responses and the answers are unambiguously correct; and point-based or semi constrained arise when there are identifiable words, statements or ideas were listed, as confirmed by scholarly literature (Child et al, 2015; Tisi et al, 2013; Ahmed & Pollit, 2011 Bramley, 2008). The objective mark schemes therefore enhanced quality of marking. The examiners who participated in this study indicated that there was too much scrolling that delayed the marking pace in the November 2015 5008/2 examinations. The challenges encountered with Paper 2 in 2015 emanated from long responses elicited by questions in Section B, resulting in calls to adapt the questions to e-marking.

This is supported by Child et al (2015:8), who posit that when using the point-based mark scheme, the examiner reads the whole response to determine the sections that relate to the mark scheme. The finding is also supported by Johnson et al (2012a; 2012b) who established that scrolling and between marker variability were associated with long responses, arguing that the latter can be mitigated by reinforcing procedures such as standardisation. The scrolling, the slow marking pace and frequent stoppages by suspect seeds could have frustrated the eight examiners who abandoned the November 2015 marking exercise. It can therefore be concluded that questions in Section B of 5008/2 were marked less accurately in the November 2015 examinations, hence reducing the quality of marking.

The findings of this study established that there were variations in the amount of space provided for questions with the same number of marks. This implies that there were also variations in the cognitive strategies used by the examiners from question to question. This is supported by Bramley (2008:8) who posits that the amount of space available for candidates to write their answers is related to the amount of writing required and to the maximum mark and might have an influence on the marker agreement as well. The author emphasised that the larger the area to

view and locate the correct responses, the greater the opportunity for a cognitive process error. Section B of 5008/2, where responses were longer might have reduced accuracy of marking due to the large amounts of space the examiners had to scan for correct answers. The accuracy of marking might have been enhanced on questions where smaller spaces were provided for the candidates' responses. It can therefore be concluded that some questions were marked with high accuracy while others were marked with low accuracy.

Closely related to the amount of space is the amount of writing required. Bramley (2008:8) argues that a large amount of writing space gives candidates a greater opportunity to express themselves correctly or incorrectly. It is therefore expected that questions that required a lot of writing demanded System 2 skills and were marked with low accuracy and less marker agreement. It can therefore be argued that 5008/2 examinations were marked with less accuracy and marker agreement. Bramley (2008:8) envisaged that the points to mark ratio in the mark scheme influences marking accuracy and marker agreement, arguing that where the examiner has a wide range of responses against which to compare the actual responses, the marking task is more cognitively demanding, and less marker agreement is expected. Biology 5008/2 and 5008/4 mark schemes had some questions where the number of marking points equaled the number of marks and others where the marking points were more than the number of marks.

Some examiners complained that too many marking points made the 5008/2 mark scheme too long and difficult to pin on the screen during marking. This finding is related to the findings of the study conducted by Johnson et al (2012b: 58) who established that navigation is more difficult on screen than on paper for essays. This study established that navigation is equally difficult with Biology (5008) examinations that were not as long as the essays investigated by Johnson et al (2012b: 58). The difficulty to navigate on screen could frustrate the examiners, leading to poor marking.

The Biology 5008/2 and 5008/4 mark schemes allowed some variations in the answers credited by examiners, such as '*AW (Alternative wording)*'. Bramley (2008:8) posits that the effect of restrictions, qualifications and variants on marker agreement could not be predicted, adding that these features could enhance or reduce marker agreement depending on their nature. In the

context of this study, it can be argued that 'A/W' could open a wide range of interpretations by examiners. Where others would credit the 'A/W' others would penalise it, resulting in reduced marker agreement, and hence quality of marking.

The Biology 5008/2 and 5008/4 mark schemes sometimes specified answers that were not acceptable with '*R (reject)*' and acceptable answers with '*A (accept)*'. Bramley (2008:8) predicted that mark schemes that specify wrong answers could enhance marking accuracy and marker agreement. These specifications in the (5008) mark schemes were likely to increase marking accuracy and marker agreement. Bramley (2008:8), however warns that examiners who mainly use System 1 strategies of marking may erroneously credit or penalise answers specified as unacceptable or acceptable respectively. There were no reasonable grounds to conclude that some (5008) examiners could have credited or penalised candidates where there were specified correct or wrong answers respectively.

The findings of this study indicate that the OSM technology supported the marking of 5008/4 and Section A of 5008/2. It did not support the marking of 5008/2 Section B and 5008/3, which was a practical examination, without violating the assessment framework and, hence compromising the validity of the assessments. This is supported by Ramakrishna et al (2012:15) and Roan (2009:7), who concur that the advantages of an onscreen marking technology would be valuable if it supports assessment practices and principles of the examination body. Describing the challenges that face marking systems, the Ofqual (2013:20) emphasised that as with any measurement tool, any assessment is likely to have some element of unreliability in its results, calling on examination authorities to ensure that marking is as good as it can be in the context of the examination system. The Ofqual (2013:21) further argues that while tightly constrained, short answer questions will result in higher reliability of an exam, they are not always a valid means of assessing certain knowledge and skills, writing thus:

In some subjects the use of high mark questions with complex, extended responses is an important aspect of validity. Here, an education system may accept lower levels of reliability where we believe the question type to be essential in assessing certain knowledge and skills. However, if levels of reliability become too low, results are not a consistent measure of candidates' performance and the assessment becomes meaningless.

It is therefore important for ZIMSEC to accept some element of unreliability in order to preserve the validity of the examinations. When the Council decides to resume OSM, they could consider marking whole scripts and methods of quality control other than seeds for extended response questions and practical examinations in order to take advantage of the opportunities offered by the technology to enhance the efficiency and quality of marking.

This study established that the nature of questions and mark schemes in the examination influenced quality control in the OSM environment as supported by the conceptual framework that guided it. The findings sufficiently answered the research sub-question on the influence of questions and mark schemes on quality control in the OSM environment. The findings of this study also answered one of the questions posed by Fowles (2011:2). The Ofqual (2013:20) argues that achieving validity is the single most important aim of an assessment. In a literature review on effects of e-marking on assessment, Fowles (2011:2) raised questions that related to the validity of examinations. The first question was, *“Is there evidence of assessment schemes being developed and shaped by what technology can offer, at the expense of the validity of the assessment?”* The findings of this study provide evidence that one of the two papers marked on screen, 5008/2, was redesigned when it was introduced to OSM. It can therefore be concluded that the OSM technology shaped the design of the question papers and mark schemes for 5008/2, threatening the validity of the examinations.

6.6 Sub-question 4: Opportunities and challenges of quality control in OSM

The fourth research sub-question was *“What are the opportunities and challenges of control in onscreen marking of O Level Biology?”*

6.6.1 The opportunities of OSM to enhance the quality of marking

The findings of this study established that the OSM had inherent opportunities to enhance the quality and efficiency of marking. The participants enumerated several opportunities of the OSM to enhance quality of marking as follows: Marks were automatically captured, eliminating addition errors; the system generated progress and quality of marking reports as well as audit trails that were used to monitor marking; the OSM was fast and efficient since there was no

paper work; the SMs could delete marks awarded by poor markers; the anonymous script portions reduced examiner bias and eliminated examiner malpractice; the SMs could set and change quality control parameters; poor markers were identified and stopped by the seeds.

Some of these opportunities of the OSM technology to enhance the efficiency and quality of marking are supported by the literature as discussed in Chapter 3 Section 6, which includes these and more, such as enhanced process control; improved access to wider examiner expertise; enhanced communication and support across the evaluator teams; increased marking speed (15% faster than PBM); richer data on examiner, candidate, item, component and paper performance; raised marking quality/consistency; and reduced administrative error (DRS, 2015; Pinot de Moira, 2013; Ramakrishna et al, 2012; Fowles, 2011; Roan, 2009). Some of these opportunities were confirmed in research studies (Coniam & Yan, 2016; Fowles, 2011). It can be concluded that some of the benefits of the OSM technology were realised in the marking of Biology (5008) examinations, enhancing the quality of marking in the same.

As discussed in Chapter 3, Section 5.4, ZIMSEC management adopted the OSM technology expecting to benefit from its advantages and more so, to save money. However, ZIMSEC did not fully take advantage of some of the opportunities offered by the OSM technology. As discussed earlier, ZIMSEC did not categorise examiners according to the demands of the test items, neither did they use automatic marking. As discussed in Chapter 2 Section 7, researchers have established the benefits of using non-expert examiners (Suto & Nádas, 2008; Meadows & Billington, 2013). It can be argued that the ability of the OSM technology to speed up marking is a result of its ability to distribute items to different categories of examiners, coupled with automatic marking. By distributing items to expert and non-expert examiners and identifying items for automatic marking, ZIMSEC could have eliminated the marking schedules that extended to January of each year, saving on time and money.

6.6.2 Challenges that could compromise quality of marking

All data sources indicated that the quality of marking was influenced by several challenges which included power cuts; erratic internet; seeds frequently popping up until they could be recognised by examiners; differences in examiner payments; poor script images; control of the

technology by SWP, resulting in ZIMSEC failing to solve problems on-site; lack of commitment and support from some senior managers; answer spaces were not adequate for some candidates who then wrote their responses on inappropriate spaces; examiners were not able to see and mark responses on inappropriate spaces; additional answers were difficult to add onto mark schemes once marking started; missing marks; and limited marking time. As discussed in Chapter 1 Section 2.4, the OSM technology was introduced at a time when there was limited use of ICT in general and in education in particular; the internet coverage was patchy; power supply was erratic; and the economy was ailing. These challenges disrupted quality control activities in the OSM environment as established by the findings of this study.

The AQA faced challenges with the marking of unconstrained papers, leading to 3353 candidates from 1335 examination centres receiving wrong marks for 48 subject components. The Ofqual (2011:5) enumerates the factors that contributed to the system failure as follows: the process for dealing with the variety of ways in which candidates recorded their answers; the process for fixing the segmented images of the candidate's responses before they are released to examiners for marking; the role and training of examiners in the onscreen marking process; the selection of components for onscreen marking of unconstrained answers in separate answer booklets; limitations of the pilot exercises carried out in 2009 and January 2010; inadequate user acceptance testing; and the absence of appropriate project and risk management arrangements.

The same factors influenced the quality of marking for ZIMSEC. Whereas the AQA marked unconstrained examinations, ZIMSEC shied away from them after marking one. In order to avoid the challenges of fixing candidates' responses for unconstrained examinations, ZIMSEC constrained all examinations before marking them on screen, compromising their validity as established in this study. The examiners, especially SNRs, were not adequately trained to perform their roles in the OSM environment, leading to poor monitoring of the quality of marking. Just like the AQA, ZIMSEC conducted one pilot study but its results were apparently not disseminated within the organisation (personal experience). As was the case with AQA, there was inadequate acceptance testing, resulting in passive resistance of the technology by some ZIMSEC senior managers and inefficient use of human and material resources within the organisation.

There was evidence that there was no commitment by the senior management to solve the challenges that bedeviled the use of the OSM technology. As explained in Chapter 4, Section 4.4.3, ZIMSEC suspended OSM in 2018 after a massive leakage of the November 2017 English Paper 2 (marked on screen) that had been printed outside the Council. The ZIMSEC state-of-the-art printing press was commissioned at its Norton premises on the 23rd of August 2019 by the Minister of Primary and Secondary Education, Professor Mavhima, who stood in for His Excellency President Emerson Munangagwa (personal experience; Tshili, 2019; www.zimsec.co.zw). Tshili (2019: www.chronicle.co.zw) quoted the minister saying that the printing press enabled the in-house printing of all examinations from Grade 7 to A level: "...We have never had problems with what we have done internally for ourselves and we assume that with the printing press in-house we will increase security and address the issue of examination leakages..." If ZIMSEC decides not to resume the OSM despite the in-house printing, then its suspension could be a result of more than what the eye can meet, possibly management apathy. Zimbabwe had been celebrated as the pioneer of the OSM in Africa as indicated in Chapter 1, Section 2.4. Examination authorities in Southern Africa trooped to Zimbabwe to learn about the OSM (personal experience). Abandoning the technology would be a betrayal of not only Zimbabwe but the African continent.

If ZIMSEC fails to resume the OSM all efforts made so far to automate assessment processes would go to waste. As discussed in Chapter 3 Section 3.3, ZIMSEC adopted the OSM technology for the June 2012 examinations. In the same year, the Council designed and implemented an e-registration programme, where candidates' details are electronically captured at the examination centre and passed on to the Council for processing as well as software for authoring of examinations in 2016. As discussed earlier, the Council also commissioned a printing press in 2018. It would therefore be retrogressive to electronically author items and digitally print them, electronically register candidates and then mark the scripts on paper.

The findings of this study were summarised into a framework that could guide the practice of quality control in the OSM environment.

6.7 The framework for quality control in the OSM environment

On the whole, the findings of this study were summarised into a framework that could guide the practice of quality control in the OSM environment in Zimbabwe as shown in Figure 6.1, to answer the question *‘How can quality control in the OSM of O Level Biology examinations be framed to provide guidelines for its practice?’*

Quality control activities were carried out before, during and after the actual marking. Figure 6.1 summarises the activities into a framework that could guide the quality of marking Biology examinations in the ZIMSEC context.

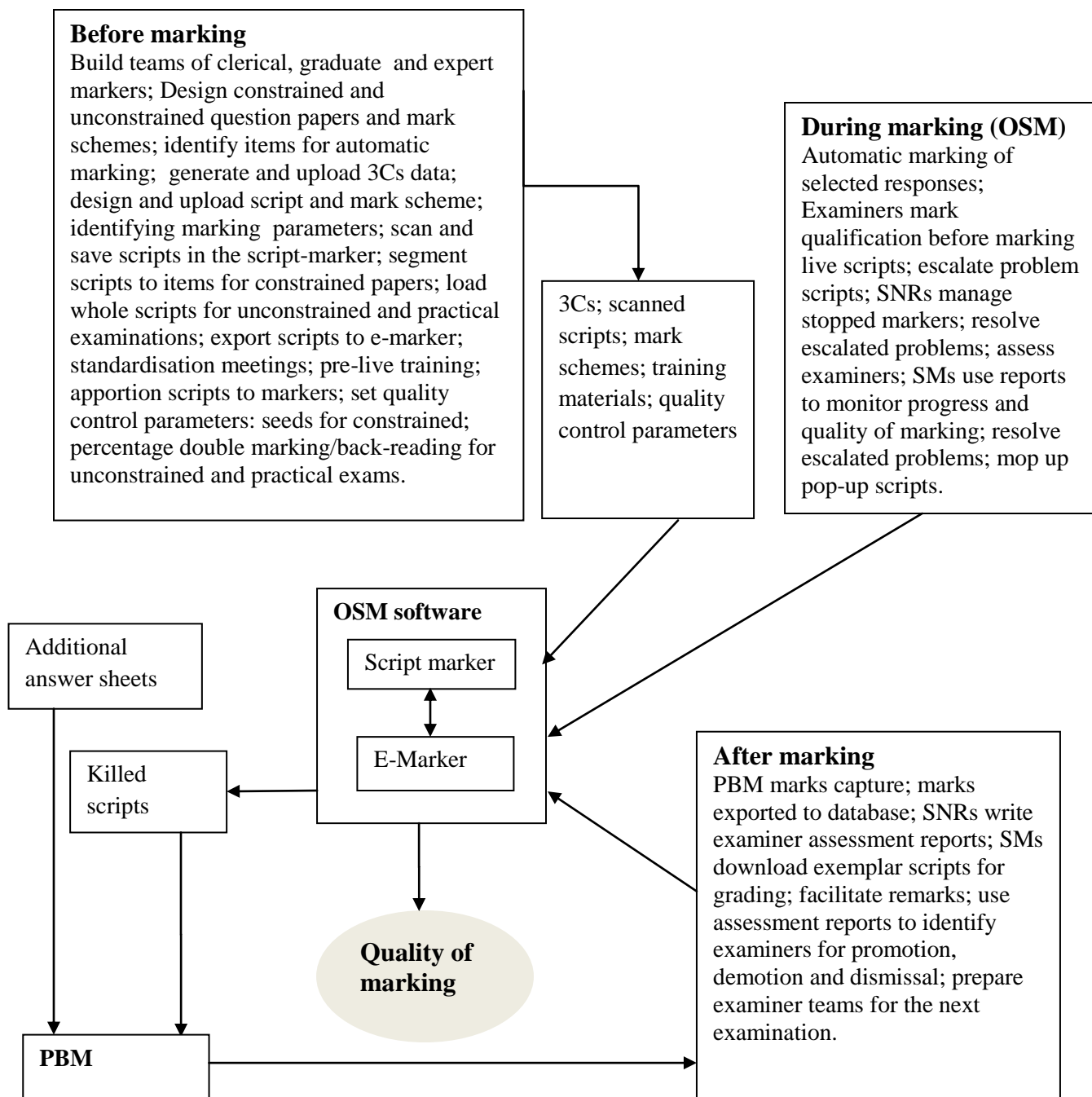


Figure 6.1: A framework for quality control in the ZIMSEC context

This framework takes cognisance of the fact that not all paper-based examinations can be marked on screen. The examination authority should set the clear criteria for selecting examinations for the OSM to preserve the validity of the same as prescribed by the assessment framework and policies that determine the thresholds for quality control parameters. The ability of the OSM

technology to enhance quality of marking is largely influenced by the context in which it is used and in the ZIMSEC context, the assessment framework, the technological infrastructure and the human capacity mattered. The commitment of the management is also crucial for the practice of quality control in the OSM environment. This is the major contribution of the study. Furthermore, the study responded to some pertinent questions which were not answered in the literature regarding quality control in the OSM of Biology examinations.

Ofqual (2014c:5) reviewed literature on quality of marking internationally and accessed data from New South Wales, Canada, China, Hong Kong, New Zealand, Korea and the United States. Some of the research questions in the literature review were answered in the Zimbabwean context, with the example of O Level Biology (5008) examinations. Although the current study established that Biology (5008) examinations were marked at item level, there was no evidence of how this marking approach influenced the quality of marking. Given that ZIMSEC was the first to use the OSM technology in Africa; this study contributes valuable literature on the quality control in a context that may be similar to many countries in the continent.

The current study also contributes to the literature that answers the questions raised by Raikes et al (2004:13-14) about quality control in the OSM environment. Some of the questions included the criteria to be used to select the answers for seeds; the roles of senior examiners in the OSM quality control system; the feedback that should be provided to the examiners, when and how it should be communicated. Pinot de Moira (2013:2) lamented that the evaluation of operational quality assurance in the OSM environment is largely absent in literature and that there is no evidence that the results of pilot studies replicate in live examinations. The results of this study add to the few ones on quality control in the OSM environment, highlighting the opportunities and challenges using live examinations. The study also confirmed or disputed the results of other studies conducted in other contexts.

6.8 Conclusion

The findings suggested that the economic conditions that prevailed in Zimbabwe influenced the technological context in which Biology examinations were marked as indicated by inadequate technological infrastructure characterised by hired computers, patchy internet and erratic power;

and low digital literacy. The technological context influenced the capacity of the ZIMSEC staff to operate in the OSM environment. The recruitment, training and standardisation procedures were not aligned to the OSM environment, leading to the failure of the Council to fully utilise the affordances offered by the technology. The training and standardisation were therefore not transferred to marking in the OSM environment.

The quality of marking was monitored by seeds, escalation of problem scripts and automatically generated reports and audit trails. There was the risk to set seeds at low levels that could compromise quality of marking, given that the SMs in this study adjusted the seed parameters during marking. The assessment framework provided by the syllabus determined the type of questions and mark scheme as well as the question paper structure. There exists the risk of designing examinations to suit the demands of the OSM technology and compromising the assessment framework, and hence the validity of the assessment.

The assessment framework provided by the syllabus determined the type of questions and mark scheme as well as the question paper structure. Some papers could not be marked on screen. However, some papers were redesigned to meet the demands of the OSM technology. The quality of marking was influenced by the nature of examination questions and mark schemes. The OSM technology had the opportunity to enhance quality control. Some of these opportunities were however reduced by challenges that emanated from the context. The quality control activities were summarised into a framework that could guide the practice of quality control in the OSM environment.

The next chapter summarises the study, presents conclusions and makes recommendations about the practice of quality control in the OSM environment

Chapter 7

Conclusions and recommendations

7.1 Introduction

The purpose of this study was to explore the influence of the practice of quality control on the marking of Ordinary Level Biology in the OSM environment in Zimbabwe in order to propose a framework which can help to improve the practice. This chapter summarises the study by articulating the major conclusions that were drawn from the findings. The limitations of the study that are discussed before the recommendations are made for practice and further research. The major themes that emerged from the study are presented by research sub-question.

This section summarises the study by articulating the major conclusions that were drawn from the findings. The limitations of the study are discussed before the recommendations are made for practice and further research. The major themes that emerged from the study are presented.

7.2 Influence of examiner training and standardisation

The quality of marking largely depended on the competence of the SNRs who set the marking standards by editing the mark schemes, selecting and marking dummies, set the seeds and pass on the marking standards to the examiners through standardisation meetings. The need to mark examinations within set deadlines exerted pressure on the SMs and the examiners who had to rush through the training and standardisation. The skills and knowledge gained by the examiners from these crucial quality control processes were therefore not effectively transferred to live marking for several reasons that include: inadequate skills transfer from SWP to the ZIMSEC; recruitment, training and standardisation processes that were out of sync with OSM; variations in the digital literacy of the examiners; and time constraints. The limited transfer of training and standardisation apparently had a bearing on the capacity of SNRs to set marking standards and to monitor the quality of marking as evidenced by the setting of wrong seeds; and the failure to use the icon that permanently stopped errant markers.

7.3 Mechanisms of monitoring quality of marking

7.3.1 Seeds approach to quality control

Quality of marking was monitored by seeds. In this study the seed parameters were set by SMs and the seeds were set by the SNRs. There were clear criteria for selecting the scripts that were set as seeds. Some factors seemed to enhance the robustness of the seeds system. First, the examiners were required to accurately mark qualification seeds before the system allowed them to mark live scripts. Poor markers continued to mark seeds, resulting in them marking more seeds than live scripts; second, the SNRs could permanently stop errant markers from marking questions that they struggled with and request SMs to delete marks from the scripts marked by such examiners; third, SNRs could retrain stopped markers before allowing them to resume marking; wrong seeds could be identified and flagged as suspects; Fourth, the SMs could monitor the seedbank and delete the suspect seeds. This study revealed that some of these opportunities were utilised while others were not fully utilised.

I also found that there was the risk to set seeds at low levels that could compromise quality of marking, given that the SMs in this study adjusted the seeds parameters during marking, and one examiner suspected that there were times when there were no seeds in the system. Some participants thought that higher seed percentages would enhance the quality of marking while document review showed that increasing the seed percentage would also increase the marking load and, hence, the marking period and attendant costs. The finding demonstrates the dilemma faced by examination authorities in the choice of mechanisms of monitoring marking at reasonable costs and timeframes. However, ZIMSEC could have saved time and money by utilising the ability of the OSM technology to automatically mark some items and distribute others to clerical and expert markers.

7.3.2 Reports and audit trails

The quality of marking is also monitored using real-time reports and audit trails. I found that the reports were in the administrator module, used by managers in real-time and not by senior markers as purported in literature. Document review and face-to-face interviews however indicated that the SNRs used the reports to assess examiners at the end of each marking period.

This study was focused on the mechanisms of monitoring the quality of marking and therefore did not collect data on the procedures of retrieving the reports from the administrator module and availing them to the SNRs. Such procedures can be interrogated in future research studies.

7.3.3 Escalating problem scripts

The quality of marking was also monitored by allowing examiners to escalate problem scripts as rescan requests. Document review and face-to-face interviews with SMs indicated that some scripts failed to scan because they had inadequate identification features; and that some problem scripts were killed and marked on paper, together with additional answer sheets used by candidates. The nature of the software also increased killed scripts by not allowing scan operators to reverse commands. This means that considerable PBM is done for components marked on screen. None of the literature reviewed in this study reported on escalation of problem scripts as a way of monitoring the quality of marking. This study therefore contributes new knowledge to the practice of quality control in the OSM environment.

7.4 Test design issues

7.4.1 The syllabus

The assessment framework provided by the syllabus determined the type of questions and mark scheme as well as the question paper structure, and the examinations could be judged against the assessment objectives, weighting and the assessment scheme. Document review indicated that constrained and unconstrained papers could be marked on screen using seeds and double marking respectively.

The findings suggested that the structure of question papers informed the approach to quality control. In this study the papers that were constrained as required by the assessment framework were most suitable for OSM, with seeds as the mechanism of script moderation. I however established that not all such papers can be marked onscreen. The practical examinations were an exception to this rule because they could not be marked onscreen without prejudicing the candidates. Of necessity, examination authorities need to carefully consider the syllabi for subjects and set out clear criteria to select examinations for OSM without compromising their validity. This finding has implications for the marking of Biology practical examinations in the

new curriculum (O Level Biology [4025]) where all candidates are required to sit for the practical examinations. The Council could however consider the option of marking practical examinations at whole script level to, at least, eliminate challenges associated with PBM.

The findings showed that there exists the risk of designing examinations to suit the demands of the OSM technology and compromising the assessment framework, and hence the validity of the assessments. This is evidenced by the constraining of Section B of 5008/2 for OSM and the subsequent shortening of the questions. The tendency to redesign questions shows a disregard of the assessment validity in favour of its reliability, leading to the violation of the assessment framework. This practice led to the allegations of watering down the examinations, which could erode the public confidence in the examination system.

I also found that mark scheme features influenced the quality of marking. Objective mark schemes demanded low cognitive abilities from the examiners than point-based mark schemes that required higher cognitive abilities. Objective mark schemes therefore facilitated accurate marking, while point-based mark schemes open some examiner disagreements. The wording of the mark scheme and the ratio of marks to marking points also influenced the quality of marking. The test designers are therefore faced with the temptation to set short questions with objective answers, especially when answer spaces are provided on the question paper, violating the requirements of the syllabus.

7.4.2 The dilemma of dealing with standardised tests

Another theme that emerged from my analysis is the dilemma of dealing with standardised tests. This is evidenced by the existence of alternative to practical paper for centres that could not administer practical examinations; and the supervisors' report that was meant to cater for laboratory conditions that could not be standardised, suggesting that ZIMSEC faced the dilemma of dealing with standardised tests that are alleged to be inherently biased. This dilemma might militate against the adoption of e-assessments in the ZIMSEC context, in the wake of increasing calls to focus on computer-based assessments where examinations are designed, written and marked on computers. Examination authorities might not be able adopt such technologies

without attracting the attention of criticism related to standardised tests. Such criticism erodes the credibility of public examinations.

7.4.3 The influence of answer spaces

Another factor that seemed to influence the quality of marking was the amount of space provided for candidate's responses and the amount of writing by the candidates. Document review shows that ZIMSEC constrained all papers marked on screen to avoid challenges and costs related to unconstrained examinations that offer the candidates unlimited space to write their responses. Limited spaces will force the candidates to write on inappropriate spaces where examiners might not see and mark the responses, leading to remark requests that also erode public confidence in the examination system. Candidates might not fully express themselves on the limited spaces. Technology and assessment experts could work together to come up with a technology that can ease the marking of unconstrained examinations at minimum costs.

7.5 Opportunities and challenges of quality control

The OSM technology had inherent opportunities to enhance the quality of marking by automation of processes such as mark capturing and addition, generation of reports and script management. Several challenges threatened to compromise the quality of marking. Major among the challenges were the poor technological infrastructure, low digital literacy, time constraints created by the need to save money, high turnover of Biology SMs; the hesitation by ZIMSEC leaders to make milestone decisions and the subtle antagonism between the ZIMSEC and the SWP.

On the whole, the findings of this study suggest that quality control activities in the OSM of Biology (5008) examinations were carried out before, during and after marking. The activities were summarised into a framework (Chapter 6, Figure 6.1) that could guide the practice of quality control in the OSM environment. This is the major contribution of this study.

7.6 Limitations of the study

This study was limited by the inability to observe the OSM software. There was no way of verifying most of the information gathered about the software itself. A conclusion could not be

made about the ability of the software to provide the examiners with feedback on the quality of their marking. It could also have been useful to observe, for example, the position of the icon for permanently stopping markers in relation to others, to establish why it was accidentally pressed by senior markers. Screen shorts of the different modules could have been taken to illustrate the findings about the software.

The amount of data collected was limited by the economic conditions that prevailed in the country, where fuel prices continued to increase to unsustainable levels, electricity became available for a maximum of six hours a day at night. At the same time this study was conducted to meet specific deadlines. The examiners could have been given more time to respond to the questions posted on the WhatsApp group. The data was collected in three days only to allow me to write the report in the few hours when electricity was available.

The sampling bias is inherent in the WhatsApp group discussions where examiners who were not on WhatsApp or who had no smart phones were left out of the sample. This limited the diversity of the responses.

7.7 Conclusion

The quality of marking Biology examinations was influenced by the context in which they were marked. The technological infrastructure, digital literacy and the skills transfer from the SWP influenced the capacity of the ZIMSEC staff to operate in the OSM environment. Although there were quality control activities such as examiner recruitment, training and standardisation, the procedures for these activities were, however, not aligned to the OSM environment, leading to the failure of the Council to fully utilise the affordances offered by the technology.

Quality of marking was monitored by seeds that were set by SNRs at thresholds set by SMs. The findings of this study demonstrate the dilemma faced by examination authorities in the choice of mechanisms of monitoring marking at reasonable costs and timeframes. There was the risk of redesigning papers to suit the OSM technology, violating the assessment scheme and the validity of the assessments. The findings demonstrated the dilemma faced by examination authorities in

dealing with standardised tests. The OSM technology could inherently enhance the quality of marking, however, it encountered challenges that came from the context.

7.8 Recommendations

7.8.1 Policy and action

I propose recommendations that could improve the practice of quality control in the marking of Biology examination. The recommendations are mainly directed to the ZIMSEC and the parent ministry (MoPSE).

1. Recruitment, training and the standardisation procedures need to be aligned to the OSM environment. By distributing items to clerical, graduate and expert examiners and identifying items for automatic marking, ZIMSEC could have eliminated the marking schedules that extended to January of each year, saving on time and money.
2. Benchmarking recruitment, training and standardisation with international examination authorities that mark examinations on screen could help.
3. Instead of marking dummies only, the SNRs could set, discuss and approve the seeds at the pre-standardisation meetings to reduce suspect seeds;
4. There might be need to put in place policies and procedures that guide the thresholds for quality control parameters to enhance the efficiency of the technology at reasonable costs; criteria for selecting examinations for OSM; and procedures for retrieving automatically generated reports and audit trails from the administrator module and availing them to SNRs for effective monitoring of examiners during and after marking; alternatively, some SNR roles could be transferred to the SMs, especially after marking.
5. The Council could consider setting up computer centres in the ten provinces of the country to eliminate the cost of hiring computers and marking venues. It may be costly at first but it will turn out to be cost effective in the long run.
6. Being the torch bearer of the OSM technology in Africa, ZIMSEC is encouraged to resume the use of the technology and continue to lead the adoption of assessment technologies in the continent.

7.8.2 Further research

This study has raised some research questions that need to be interrogated.

1. Other researchers are welcome to review the framework and make recommendations to refine it.
2. Although the SMs in this study indicated that they had reduced teaching experience for prospective examiners from five to three years, this research did not collect data on demographic characteristics such as age, gender and teaching experiences. Further research work could focus on the influence of demographic characteristics on the quality of marking.
3. The influence of other factors such as the scanning process and the PBM for e-marked components can also be investigated.
4. Some questions on problem scripts need to be answered. What happens if the examiner chooses not to escalate a script they have failed to mark? Does the system allow the examiner to choose another script when they have not escalated a problem script?

7.8 Autobiographical reflection

I conducted this research study in the shortest time possible, two years, mainly because I was inspired by my supervisor who showed genuine interest in my work without exerting unnecessary influence on my line of thinking. Every time I submitted a chapter he would acknowledge receipt and promise to come back to me in two weeks. He would, however, always provide feedback in less than a week. His comments were thought provoking and encouraged me to read widely. As I worked with Prof. Gumbo, I reflected on my practice as a teacher/tutor and wondered if I motivated my students in the manner he did. Despite the economic challenges that militated against progress, I would press on because I did not want to betray or disappoint my supervisor who was supporting me so much as an emerging researcher.

The process of conducting this research study taught me valuable lessons that will shape my future as a researcher and an assessment practitioner. As a researcher I learnt that research plans do not always obtain on the ground. The data collection methods were influenced by the suspension of OSM in 2018; the fact that examiners were part time employees of ZIMSEC and were scattered all over the country; and the economic climate that continued to deteriorate.

I was excited to learn that the social media networks can offer easier access to the research participants that are otherwise not accessible. I accessed the examiners who were on WhatsApp groups, where consent issues were not easy to address. It was not possible to get consent from everyone on the group, so, I had to assume that those who did not respond to my questions did not want to participate in the study. The participants responded to the questions or commented on responses from their colleagues at different times by simply highlighting and typing a response or a comment. The responses to one issue were therefore scattered all over the charts. They could use emojis to express their feelings, e.g. laughing. The highlighted questions and emojis disappeared and were replaced by □□ when I copied and pasted the charts on Word, distorting the original conversations. All usernames disappeared except mine, forcing me to write up the responses straight from my phone. This involved a lot of scrolling up and down the charts since I could not code responses to the same issue on the phone. I was faced with the dilemma of data storage because I could not keep the charts on my phone for too long, lest I lose the phone and

the data. After writing up the responses from the phone I deleted the charts and secured the distorted Word document.

I also learnt that research is sometimes frustrating and cumbersome, and yet exciting and rewarding. I thought that the insider advantage always worked in the researcher's favour until I failed to access some documents that I knew about. I knew the diversity of the data would be limited but I just had to do with the available documents. I also learnt that research requires persistence and perseverance, otherwise one can easily quit. I was frustrated by the economic situation that continued to decline, creating material resource and time constraints for me. I could not buy the audio recorders that I had intended to use for face-to-face interviews, so, I had to write notes on paper, write up the interviews and seek confirmation from the participants. As the power cuts extended, I had to work during the night when electricity was available, yet I had to go to work. It was easy for me to quit but I persevered, spurred on by my supervisor, family and friends.

This study provided me with useful insights that made me reflect on my practice in the OSM environment as SNR trainer, SM and the QPDM. If ZIMSEC were to resume e-marking today, I would not do business the same way. I have gained insights that will shape my conduct as an assessment practitioner, not only in OSM but in other areas such as test design and syllabus interpretation. I intend to explore further the practice of quality control in the assessment processes where technologies have been adopted by ZIMSEC, such as the e-registration of candidates, electronic item authoring and banking software and the mark capturing system to establish how the systems can be harmonised to enhance the efficiency and quality of examinations.

List of references

- Adams, C. 2011. Assessment: Past, present and the future. Paper presented at the IBEAM Conference. The Hague, Netherlands.
- Adewo, A. 2012. The positive and negative impacts of ICT. Retrieved from <https://ajahana.wordpress.com>. Accessed 6 November 2018.
- Ahmed, J.U. 2010. Documentary research: New Dimensions. *Indus Journal of Management & Social Sciences*, 4(1): 1-4.
- Ahmed, A. & Pollit, A. 2011. Improving quality of marking through a taxonomy of mark schemes. *Assessment in Education: Principles and Practice*, 18(3): 259-278.
- Asiamah, N., Mensah, H.K. & Oteng-Abayie, E.F. 2017. General, target and accessible population: Demystifying the concepts of effective sampling. *The Qualitative Report*, vol 22 Number 6, p1609-1622.
- Atieno, O.P. 2009. An analysis of strengths and limitations of qualitative and quantitative research paradigms. *Problems of Education in the 21st Century*, vol 13, 2009, p13-18.
- Automatically generated reports: 5008/2, November 2015.
- Automatically generated reports: 5008/4 November 2015.
- Baired, J.A., Hays, M., Johnson, R., Johnson, S. & Lamprianou, I. 2013. Marker effects and marking reliability: A comparative exploration from the perspectives of generalisability theory, Rasch modeling and multilevel modeling. Coventry: Ofqual Retrieved from www.ofqual.gov.uk. Accessed 8 January 2014.
- Baxter, P. & Jack, S. 2008. Qualitative case study methodology: Study design and implementation for novice researchers. *The Qualitative Report*, 13(4): 544-559.
- Bayat, A., Louw, W. & Rena, R. 2014. The impact of socio-economic factors on the performance of selected high school learners in the Western Cape Province, South Africa. *J Hum Ecol*, 45(3): 183-196.
- Bennett, R.E. 1998. Reinventing assessment: Speculation on the future of large scale educational assessment. Princeton: Educational Testing Services.
- Bennett, R.E. 2002. Inexorable and inevitable: The continuing story of technology and assessment. *The journal of Technology, Learning and Assessment*, 1(1). Retrieved from www.jtla.org. Accessed 09 November 2018.

- Bennett, R.E. 2015. The Changing nature of Educational Assessment. *Review of Research in Education*, 39: 370-407.
- Bhawani, S.S. 2004. Emerging trends of research on transfer of learning. *International Education Journal*, 5(4): 591-599.
- Block, E.S. & Erskine, L.E. 2012. Interview by telephone: Specific considerations, opportunities and challenges. *International Journal of Qualitative Methods*, 11(4): 428-445.
- Boeren, E. 2017. The methodological underdog: A review of quantitative research in the key adult education journals. *Adult Education Quarterly*, 68(1): 63-79.
- Bolderston, A. 2012. Conducting a research interview. *Journal of Medical Imaging and Radiation Sciences*, vol43, 2012, p66-76.
- Boughton, L. 2019. 5 reasons you should use WhatsApp in your online qualitative research. Retrieved from <https://info.angelfishfieldwork.com>. Accessed 25 September 2019.
- Bowen, G. 2009. Document analysis as a qualitative research method. *Qualitative Research Journal*, 9(2): 27-40.
- Boyce, C. 2006. Conducting in-depth interviews: A guide for designing and conducting in-depth interviews. Watertown: Pathfinder International.
- Boyle, A. 2010. Some forecasts of the diffusion of e-assessment using a model. *The Public Sector Innovation Journal*, 15(1): 1-29.
- Bramley, T. 2008. Mark scheme features associated with different levels of marker agreement. Cambridge: Cambridge Assessment.
- Braun, H., Kanjee, A., Bettinger, E. & Kremer, M. 2006. Improving education through assessment, innovation, and evaluation. Cambridge: American Academy of Arts and Sciences.
- Brown, P.A. 2008. A review of the literature on case study research. *Canadian Journal of New Scholars in Education*, 1(1): 1-13.
- Bukenya, M. 2006. Comparing reliability of the conveyor belt marking system with the traditional marking system. Paper Presented at the 32nd International Association of Educational Association. Singapore, Singapore.
- Bulawayo24 News. 2018. High court to sit over exam re-sit. Retrieved from www.bulawayo24.com. Accessed 22 February 2019.

- Charmaz, K. 2017. The power of constructivist grounded theory for critical inquiry. *Qualitative Inquiry*, 23(1): 34-45.
- Child, S., Munro, J. & Benton, T. 2015. An experimental investigation of the effects of mark scheme features on marking reliability. Cambridge: Cambridge Assessment.
- Chinamasa, E. & Munetsi, C. 2012. Examinations question specialized marking: A quantitative analysis of inter-marker reliability mode at Chinhoyi University of Technology. *Zimbabwe Journal of Educational Research*, 24(2): 76-191.
- Chindaro, S. 2013. Call for dedicated policy on ICT in education for Zimbabwe. *Newsday*, 14 September 2013. Retrieved from www.newsday.co.zw. Accessed 10 August 2017.
- Cohen, L., Manion, L. & Morrison, K. 2007. Research methods in education 6th Edition. London: Routledge.
- Coniam, D. 2009. Examining negative attitudes towards onscreen marking in Hong Kong. *Educational Journal*, 37(1-2): 71-87.
- Coniam, D. 2010. Examining negative attitudes towards onscreen marking in Hong Kong. *Educational Journal*, 37(1-2): 71-87.
- Coniam, D. 2011a. A qualitative study of the attitudes of liberal studies markers towards onscreen marking in Hong Kong. *British Journal of Educational Technology*, 42(6): 1042-1051.
- Coniam, D. 2011b. The double marking of Liberal Studies in the Hong Kong public examinations system. *New Horizons in Education*, 59(2): 1-12.
- Coniam, D. 2013. The Increasing acceptance of onscreen marking: The 'tablet computer effect'. *Educational Technology and Society*, 16(3): 119-129.
- Coniam, D. & Yan, Z. 2016. A comparative of the ease of use and acceptance of onscreen marking by markers across subject areas. *British Journal of Educational Technology*, 47(6): 1151-1167.
- Coniam, D. & Yeung, S.A. 2010. Markers' perceptions of Liberal Studies in the Hong Kong public examination system. *Asia Pacific Journal of Education*, 30(3): 249-271.
- Coniam, D. & Yeung, S.A. 2011. Markers' perceptions regarding the onscreen marking of liberal studies in Hong Kong public examination system. *Asia Pacific Journal of Education*, 30(3): 249-271.

- Creswell, J.W.2009. Research design: Qualitative, quantitative and mixed method approaches, 3rdEdition. Los Angeles: Sage.
- Creswell, J.W.2014. Research design: Qualitative, quantitative and mixed method approaches, 4thEdition. Los Angeles: Sage.
- Creswell, J.W. 2007. Qualitative inquiry and research design: Choosing among five approaches. California: Sage.
- Darling-Hammond, L., Herman, J., Pellegrino, J. et al. 2013. Criteria for high quality assessment. Chicago: Stanford University.
- Dawson, C. 2007. A practical guide to Research methods: A user friendly manual for mastering research techniques and projects 3rd Edition. Oxford: How-To-Content.
- Dodd, L. 2014. Quality of marking in General Qualifications - survey of teachers. Coventry: Ofqual.
- Dongre, A.R. & Sankaran, R.2015. Ethical issues in qualitative research: Challenges and options. *International Journal of Medical Science and Public Health*, 5(6): 1187-1193.
- DRS. 2013. e-Marker® Administrator training manual version 2.6. Buckinghamshire: DRS.
- DRS. 2014. First pilot of electronic marking in Africa sees growing take-up, as Zimbabwe School Examinations Council increases speed of delivery and accuracy using DRS eMarker® technology. Retrieved from www.drs.co.uk. Accessed 10 January 2015.
- DRS e-Marker® Brochure. 2015. On-screen marking: Transforming the future of examinations. Buckinghamshire: DRS. Retrieved from www.drs.co.uk. Accessed 14January 2016.
- Dube,V. 2014. Technology vexes exam markers. *The Sunday News*, 12January 2014.
- E-marking vexes ZIMSEC. *NewsdzeZimbabwe*, Sunday, January 12,2014. Retrieved from www.newsдзеzimbabwe.co.uk. Accessed 21 January 2014.
- Erstad, O. 2008. Changing assessment practices and the role of IT. In Voogt, J. & Knezek, G. (Eds.). *International handbook of information technology in primary and secondary education*, pp. 81–194. New York: Springer.
- Etikan, I., Musa, S.A. & Alkassim, R.S. 2016. Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics*, 5(1): 1-4.
- Fadel, C. 2008. 21st Century skills: How can you prepare students for the global economy? Paris: Partners for 21st Century Skills.

- Falvey, P. & Coniam, D. 2010. A qualitative study of the response of raters towards onscreen and paper-based marking. *Melbourne Papers in Language Testing*, 15(1) p1-26
- Farooq, M.B. & de Villiers, C. 2017. Telephonic qualitative interviews, when to consider them and how to do them. *Meditari accounting research*, Retrieved from <https://repository.up.ac.za>. Accessed 10 January 2019.
- Fletcher, D. 2009. Standardised testing. Retrieved from <http://content.time.com/time/nation/article/0,8599,1947019,00.html> Accessed 5 January 2018.
- Flyvbjerg, B. 2006. Five misunderstandings about case study research. *Qualitative Inquiry*, 12(2):219-245.
- Fowles, D. 2011. Literature review on effects on assessment of e-marking. Retrieved from www.cerp.org.uk. Accessed 8 January 2014.
- Gaya, D.H. & Smith, E.E. 2016. Developing a qualitative single case study in the strategic management realm: An Appropriate research design? *International Journal of Business Management and Economic Research*, 7(2): 529-538.
- Geisha, K. 2012. CXC marking to go electronic. *The Trinidad & Tobago Guardian Online*. Retrieved from www.guardianmedia.org. Accessed 14 January 2014.
- Gentles, S.J., Charles, C., Ploeg, J. & McKibbin, K.A. 2015. Sampling in qualitative Research: Insights from an overview of the methods literature. *The Qualitative Report*, 20(11): 772-1789.
- Geranpayeh, A. 2011. The impact of online marking on examiners' behavior. *Cambridge English: Research Notes*, 43: 15-21.
- Ghaicha, A. 2016. Theoretical framework for educational assessment: A synoptic review. *Journal of Education and Practice*, 7(24): 212-231.
- Gill, P., Stewart, K. & Treasure, E. et al 2008. Methods of data collection in qualitative research: Interviews and focus groups. *British Dental Journal*, 204(6): 291-295.
- The Government of South Australia. 2017. Examination marking (Stage 2): Procedures and guidelines. Retrieved from www.sace.sa.edu/documents. Accessed 19 January 2018.
- Grossman, R. & Salas, E. 2011. The transfer of training: what really matters. *International Journal of Training and Development*, 15(2): 103-120.

- Guba, E.G. & Lincoln, Y.S. 1994. Competing paradigms in qualitative research. In Denzin, L.K. & Lincoln, Y.S (Eds.). *Handbook of qualitative research*, pp. 105-117. Thousand Oaks: Sage.
- Guetterman, T.C. 2015. Descriptions of sampling practices within five approaches to qualitative research in education and the health sciences. *Forum: Qualitative Social Research*, 16(2). Retrieved from www.qualitative-research.net/. Accessed 18 September 2018.
- Haggie, D. 2008. The strategic use of marking technologies to support innovation and diversity in assessment. Retrieved from: www.rm.com/_RMVirtual/.../IAEA_paper_Sept_2008_David_Haggie.pdf. Accessed 8 January 2014.
- Harrison, H., Birks, M. & Franklin, R. et al. 2017. Case study research: Foundations and methodological orientations. Retrieved from www.qualitative-research.net. Accessed 6 August 2017.
- Hays, J. 2013. Chinese imperial exams. Retrieved from <http://factsanddetails.com/china/cat2/4sub9/entry-5385.html>. Accessed 5 January 2018.
- Hill, P. 2013. Asia Pacific secondary education system review series No. 1: Examination systems. Bangkok: UNESCO.
- Houghton, C., Casey, D., Shaw, D. & Murphy, K. 2013. Rigour in qualitative case study research. *Nurse Researcher*, 20(4): 12-17.
- Hudson, G. 2009. Improving marking quality in essays: Can technology help? Paper presented at the 35th IAEA Conference. Brisbane, Australia.
- Huddleston, A.P. & Rockwell, E.C. 2015. Assessment for the masses: A historical critique of high-stakes testing in reading. *Texas Journal of Literacy Education*, 3: 38-49.
- Isaacs, T., Zara, C., Hebert, G., Coombs, S.J. & Smith, C. 2013. Key concepts in educational assessment. New Delhi: Sage.
- Johansson, R. 2003. Case study methodology. A keynote speech at the International Conference on Methodologies in Housing Research. Stockholm, Sweden.
- Johnson, R.B. & Onwuegbuzie, A.J. 2004. Mixed methods research: A paradigm whose time has come. *American Educational Research Association*, 33(7): 14-26.

- Johnson, M. & Black, B. 2012. Feedback as scaffolding: Senior examiner monitoring processes and their effects on examiner marking. *Research in Post-Compulsory Education*, 17(4): 391-407.
- Johnson, M., Hopkin, R. & Shiel, H. 2012a. Marking extended essays onscreen: Exploring the link between marking process and comprehension. *E-Learning and Digital Media*, 9(1): 50-68.
- Johnson, M., Hopkin, R. & Shiel, H. 2012b. Extended essay marking on screen: Is examiner marking accuracy influenced by marking mode? *Educational Research and Evaluation*, 18(2): 107-124.
- Johnson, D. & Johnson, B. 2009. High Stakes Testing. Retrieved from <http://www.education.com>. Accessed 9 October 2014.
- Johnson, M. & Nádas, R. 2009. Marginalised behaviour: Digital annotations, spatial encoding and the implications for reading comprehension. *Learning, Media Technology*, 34(4): 323-336.
- Johnson, M., Nádas, R. & Bell, J.F. 2010. Marking essays on screen: An investigation into the reliability of marking extended subjective texts. *British Journal of Educational Technology*, 41(5):814–826.
- Kachere, P. 2012. ZIMSEC scores a first in Africa. *The Sunday Mail Zimbabwe*, 15 July 2012.
- Karombo, T. 2012. Electronic marking for Zimbabwe exams. *ITWeb Africa*, 16 July 2012.
- Karombo, T. 2017. Nearly half of all internet traffic in Zimbabwe goes to WhatsApp. QuartzAfrica, 23 October 2017. Retrieved from <https://qz.com/africa>. Accessed 25 September 2019.
- Kandur, J.L. 2017. Testing times: How tests standardised learners. Retrieved from www.dailysabah.com/feature/2017/09/23/testing-times-how-tests-standardizedlearners. Accessed 9 January 2018.
- Kanyongo, G.Y. 2002. Zimbabwe's education system reforms: Successes and challenges. *International Education Journal*, 6(1): 65-74.
- Kapukaya, K. 2013. Assessment: A help or hinderance to educational purposes. *International Journal of Humanities and Social science*, 3(6): 84-92.
- Kaukab, J.S. & Mehrunnisa, S. 2016. History and evolution of standardised testing: A literature review. *International Journal of Research*, 4(5): 126-132.

- Kellaghan, T. 2004. Examinations, National and international assessments and educational policy. Dublin: Education Research Centre.
- Kellaghan, T. & Graeney, V. 2004. Assessing learner learning in Africa. Washington: The World Bank.
- Khan, B. 2012. Relationship between assessment and learner learning. *International Journal of Social Sciences and Education*, 2(1):576-588.
- Khlifi, Y. & El-Sabagh, H.A. 2017. A novel authentication scheme for e-assessments based on student behaviour over e-learning platforms. *International Journal of Emerging Technologies in Learning*, 12(4): 62-89.
- Kingdon, M. (Ed). 2014. The development of e assessment 2004-2014. East Sussex: Exams on Demand Group.
- Kivunja, C. & Kuyini, A.B. 2017. Understanding and applying research paradigms in educational contexts. *International Journal of Higher Education*, 6(5): 26-41.
- Korstjens, I. & Moser, A. 2018. Practical guidance to qualitative research, Part 4: Trustworthiness and publishing. *European Journal of General Practice*, 24(1): 120-124.
- Kurshan, B. 2017. Teaching 21st century skill for 21st century success requires an ecosystem approach. Retrieved from www.forbes.com/sites/barbrakurshan. Accessed 5 August 2018
- Kuwaza, K., Kwinjo, K. & Gonditii M. 2019. Power blackouts force Zim firms to shut down. The independent, 19 July 2019. Retrieved from www.theindependent.co.zw. Accessed 25 September 2019.
- Lal, S., Suto, M. & Unger, M. 2012. Examining the potential of combining the methods of grounded theory and narrative inquiry: A comparative qualitative analysis. *Qualitative Report*, 12: 1-22.
- Langa, E. 2019. Churches decry economic hardships. *Newsday*, 1 February 2019, p. 4.
- Mack, L. 2010. The philosophical underpinnings of educational research. *Polyglossia*, 19: 5-11.
- MAM. 2017. Final IDS-November 2017 e-marked components. Memo to QPDM, 14 September 2017.
- MAM. 2013. Training needs – Urgent. Email forwarded to QPDM, 6 December 2013.
- MAM. 2013. ILAP pages and unconstrained booklets. Memo to TDR&E, 11 April 2013.
- MAM. 2013. December 2013 marking project. Memo to TDR&E, 10 September 2013.
- MAM. 2013. July 2013 e-marking information sheet. Unpublished.

- MAM. 2014. 2013 marking post-mortem programme. Unpublished.
- MAM. 2014. 2013 marking post mortem. Email to SWP, 18 May 2014.
- MAM. 2013. Training Blocs – Urgent. Email forwarded to QPDM, 8 December 2013.
- MAM. 2014. June 2014 information. Email exchanges with SWP, 8-16 April 2014.
- MAM. 2014. E-marking programme. Memo to Assistant Director TDR&E, 23 October 2014.
- Maravanyika, G. & Sguazzin, A. 2019. Zimbabwe teachers call off strike...to return to work Monday. Retrieved from www.bloomberg.com. Accessed 14 February 2019.
- Manayiti, O. 2019. Apex Council backs off strike threat. *Newsday*, 1 February 2019, p. 2.
- Manayiti, O. & Munyeke, T. 2018. Govt orders O Level English paper re-sit. *Newsday*, 9 February 2018. Retrieved from www.newsday.co.zw. Accessed 22 February 2019.
- Mangudhla, T. 2016. Zimbos in love with WhatsApp and Facebook— Potraz report. *The Independent*, 24 March 2016. Retrieved from www.theindependent.co.zw. Accessed 25 September 2019.
- Maramwidze-Merrison, E. 2016. Innovative methodologies in qualitative research: Social media window for accessing organisational elites for interviews. *The Electronic Journal of Business Research Methods*, 14(2): 157-167.
- Marist International Solidarity Foundation [FMSI]. 2011. Universal periodic review of the Republic of Zimbabwe. Retrieved from www.lib.ohchr.org. Accessed 5 May 2018.
- Mashanyare, I. & Chinamasa, E. 2014. School examinations leakage: Case Zimbabwe School Examinations Council. *IOSR Journal of Humanities and Social Science*, Volume 19, Issue 4, Ver. I, p47-54
- Mashoko, D., Mateveke, P., Kufakunesu, M. & Mashoko, E. 2013. Public examinations: A fair deal or a burden? The Zimbabwean experience. *International Journal of English and Education*, 2(3): 462-470.
- Matiashe, F. 2019. Govt moves to deduct striking teachers' salaries. *Newsday*, 8 February 2019. Retrieved from www.newsday.co.zw. Accessed 14 February 2019.
- Mawonde, A. 2018. Exam leaks: ZIMSEC tightens security...acquires its own printing machine. *Nehanda Radio*, 24 October 2018. Retrieved from www.nehandaradio.com. Accessed 22 February 2019.

- Meadows, M. & Billington, L. 2013. The effect of marker background and training on the quality of marking in GCSE English. Manchester: Assessment and Qualifications Alliance.
- Meadows, M. & Billington, L. 2005. A review of literature on marking reliability. Manchester: Assessment and Qualifications Alliance.
- Mhlanga, B. 2019. Government refuses to meet teacher unions. *Newsday*, 22 February 2019. Retrieved from www.newsday.co.zw. Accessed 23 February 2019.
- Mirabeau, L., Mignerat, M. & Grangé, C. 2013. The Utility of Using Social Media Networks for Data Collection in Survey Research: Research-in-Progress. Paper presented at the Thirty-Fourth International Conference on Information Systems. Milan, Italy.
- Muir, C. 2017. High-stakes testing: Timelines. Retrieved from www.ruby.fgcu/courses. Accessed 19 January 2018.
- Munoz, R. 2013. High-stakes testing pros and cons. Retrieved from www.education.com. Accessed 10 February 2014.
- Musarurwa, C. & Chimhenga, S. 2011. Credibility of school examinations in Zimbabwe: A reflective analysis. *Academic Research International*, 1(1):173-179.
- Mwanyumba, D. & Mutwiri, J.G. 2009. Challenges associated with implementation of control mechanisms in public examinations and how the Kenya National Examinations Council (KNEC) has handled some of these challenges. A Paper presented at the 27th Annual Conference of the Association for Educational Assessment in Africa (AEAA). Yaoundé, Cameroon.
- Namwandi, D. 2014. The ministry of education has implemented e-marking for some subjects for 2014 examinations of Grades 10 and 12. *The Namibian*. Retrieved from www.namibian.com. Accessed 1 July 2015.
- Ncube, X. & Langa, V. 2019. Opposition leader demands meeting with Munangagwa. *Newsday*, 1 February 2019, p. 2.
- Ndlovu, M. 2019. Government shows teachers the middle finger. *Bulawayo24 News*, 21 February 2019. Retrieved from www.bulawayo24.com. Accessed 23 February 2019.
- Newgen flyer. n.d. Onscreen marking solution: Making examination marking more accurate, secure and cost effective. Retrieved from www.newgensoft.com/solution/education/flyer. Accessed 14 July 2017.

- Newsday, 6 February 2019. 80% teachers heed strike call. Retrieved from www.newsday.co.zw. Accessed 7 February 2019.
- Ngara, R. & Ngara, R. 2013. Conveyer belt marking: Opinions of ZIMSEC markers in Chikomba district. *HOPE Journal of Research (House of Pakistani Educationists)*, 1(2): 33-42.
- Njie, B. & Asimiran, A. 2014. Case study as a choice in qualitative methodology. *Journal of Research & Method in Education*, 4(3): 35-40.
- November 2013 examinations O Level missing mark status. Memo to Assistant Director Exams, 29 April 2014.
- Nowell, L.S., Norris, J.M. & White, D.E. et al. 2017. Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, 16: 1-13.
- Nichols, S.L. & Berliner, D.C. 2008. Collateral damage: How high-stakes testing corrupts America's schools. Cambridge: Harvard Education Press.
- Nyathi, P. 2019. A full tank of fuel now costs more than a civil servant's salary and more than half a doctor's. Zimeye, 7 October 2019. Retrieved from www.zimeye.net/. Accessed 13 October 2019
- Odendahl, N.V. 2011. Testwise: Understanding educational assessment. Volume 1. Lanham: Rowman & Littlefield.
- Ofqual. 2011. Inquiry into the failure of part of AQA's GCSE, AS and A level Script-marking process in the summer 2010 Examination Series: Final Inquiry Report. Coventry: Ofqual.
- Ofqual. 2013. Review of quality of marking in exams in A levels, GCSEs and other academic qualifications: Interim Report. Retrieved from www.ofqual.gov.uk. Accessed 8 January 2014.
- Ofqual. 2014a. Review of literature on item-level marking research. Coventry: Ofqual.
- Ofqual. 2014b. Standardisation methods, mark schemes and their impact on marking reliability. Coventry: Ofqual.
- Ofqual. 2014c. Review of marking internationally. Coventry: Ofqual.
- Ofqual. 2014d. Review of double marking research. Coventry: Ofqual.
- Ofqual. 2014e. Quality of marking: The description of marking processes used in external exams in general qualifications. Coventry: Ofqual.
- Ofqual. 2014f. Research on quality of marking. Coventry: Ofqual.

- Ofqual. 2014g. Review of quality of marking in exams in A levels, GCSEs and other academic qualifications: Final Report. Retrieved from www.Ofqual.gov.uk. Accessed 6 June 2014.
- Oltmann, S.M. 2016. Qualitative interviews: A methodological discussion of the interviewer and respondent contexts. Retrieved from <https://uknowledge.uky.edu>. Accessed 10 January 2019.
- P21. 2007. Framework for 21st Century learning. Washington: P21.
- Palys, T. 2008. Purposive sampling. In Given, L.M. (Ed.). *The Sage Encyclopedia of qualitative research methods*, pp. 697-698. Los Angeles: Sage.
- Patel, Z. 2012. Critical evaluation of different research paradigms. London: University of Westminster.
- Pellegrino, J.W. & Quellmalz, E.S. 2010. Perspectives on the integration of technology and assessment. *Journal of Research on Technology in Education*, 43(2): 119-134.
- Pellegrino, J.W. 2004. The evolution of educational assessment: Considering the past and imagining the future. Paper presented at the 6th W.H. Angoff Memorial Lecture. New Jersey, USA.
- PMS' report on November 2015 marking. Unpublished.
- Pinot de Moira, A. 2013. Identifying errant markers: Quality assurance systems in an e-marking environment. Centre for Education Research and Policy AQA Education. Retrieved from https://cerp.aqa.org.uk/sites/default/.../CERP_RP_APM_01022010_0.pdf. Accessed 18 June 2015.
- Pinot de Moira, A 2011. Why item mark? The advantages and disadvantages of e-marking. The Assessment and Qualifications Alliance (AQA). Retrieved from www.cerp.org.uk. Accessed 14 January 2014.
- Pollit, A. 2011. Comparative judgment for assessment. *International Journal of Technology & Design Education*, 22: 157-170.
- Ponelis, S.R. 2015. Using interpretive qualitative case studies for exploratory research in doctoral studies: A case of information systems research in small and medium enterprises. *International Journal of Doctoral Studies*, 10:535-550.
- QCA. 2007. Regulatory principles of e-assessment. London: QCA.

Quality of marking report: 5008/2, November 2016.

Quality of marking report: 5008/4, November 2016.

Qudsia, A.D., Farooq, A. & Muhammad, R et al. 2017. Use of social media tool “Whatsapp” in medical education. *Annals* Vol 23, Issue 1, Jan. – Mar. 2017, pg 39-42

Race, P. 2009. Designing assessment to improve Physical Science learning: A Physical Science practice guide. Retrieved from www.heacademy.ac.uk/physsci. Accessed 15 August 2017.

Raikes, N., Greateorex, J. & Shaw, S. 2004. From paper to onscreen: Some issues on the way. Cambridge: University of Cambridge Local Examination Syndicate.

Ramakrishna, A., Navya, S.B., Sri Harish, P., Swarna, S. & Vasundhara, C.H. 2012. Design and implementation procedure for administration and evaluation in e-marking system. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 2(1): 15-19.

Razemba, F. 2019. Teachers end strike. *The Chronicle*, 11 Febraury 2019. Retrieved from www.chronicle.co.zw. Accessed 14 February 2019.

Roan, M. 2009. Electronic marking: How it could bring substantial benefits to quality control and help minimise cultural bias in public examinations in a multicultural society. Data Research Services. Retrieved from www.drs.co.uk. Accessed 14 January 2014.

Ridgeway, J., McCusker, S. & Pead, D. 2004. Literature review of e-assessment. Bristol: Future Lab.

Risiro, J. 2014. Traditional or conveyor belt marking: Exploring the way forward at Great Zimbabwe University, Zimbabwe. *Greener Journal of Educational Research*, 4(4):99-106.

Rust, C. 2002. Purposes and principles of assessment. Oxford Brooks University. Retrieved from www.brooks.ac.uk/services. Accessed 10 May 2017.

Saber, G. & Ben-Yehoshua, N.S. 2017. ‘I’ll sue you if you publish my wife’s interview’: Ethical dilemmas in qualitative research base on life stories. *Qualitative Research*, 17(4): 408-423.

Sanjari, M., Bahramnezhad, F. & Fomani, F.K. et al. 2014. Ethical challenges for researchers in qualitative studies: The necessity to develop a specific guideline. *Journal of Medical*

- Ethics and History of Medicine*. Retrieved from www.ncbi.nlm.nih.gov. Accessed 22 October 2018.
- Samaita, K. 2019. Most miners asked to import their own power, a move that could double their losses. Retrieved from www.businesslive.co.za. Accessed 25 September 2019.
- Shah, S.R. & Al-Bargi, A. 2013. Research paradigms: Researchers' worldviews, theoretical frameworks and study designs. *Arab World English Journal*, 4(4): 2013 252-264.
- Shahid, S. 2018. Content analysis of WhatsApp conversations: An analytical study to evaluate the effectiveness of WhatsApp application in Karachi. *International Journal of Media, Journalism and Mass Communications*, 4(1): 14-26.
- Sibanda, E. 2019. ZIMTA forces us to down tools: Teachers. *The Herald*, 1 February 2019, p. 3.
- Shizha, E. & Kariwo, M.T. 2011. Education and development in Zimbabwe: A social, political and economic analysis. Rotterdam: Sense.
- Shafika, I. 2007. ICT in Education in Zimbabwe. Survey of ICT and education in Africa: Zimbabwe country report. Retrieved from www.infodev.org. Accessed 10 August 2017.
- SM: Biology (5008). IDS. 5008/2 June 2016.
- SM: Biology (5008). IDS. 5008/4 June 2016.
- SNRs: 5008/2. 2016. D-grade reports A-H, 29 January 2016. Unpublished.
- Social Media Research Group ([SMRG] .2016. Using social media for social research: An introduction. Retrieved from www.assets.publishing.service.gov.uk. Accessed 13 October 2019.
- Southern Africa Association of Educational Assessment. 2014. A comparative report on the education landscape of the countries in the Southern African Association of Educational Assessment. Johannesburg: SAAEA.
- Stake, R. 1995. Data gathering. In *The art of doing case study research* (49-68). Thousand Oaks: Sage.
- Stake, R.E. 2005. Qualitative case studies. In Denzin, N.K. & Lincoln Y.S. (Eds.). *The Sage handbook of qualitative research* (3rd Edition), pp. 443-466. Thousand Oaks: Sage.
- Stake, R. 2008. Qualitative case study. In N.K. Denzin & Y.S. Lincoln (Eds). *Strategies of qualitative inquiry*, pp. 134-164. Thousand Oaks: Sage.

- Starman, A.B. 2013. The case study as a type of qualitative research. *Journal of Contemporary Educational Studies*,1: 28-43.
- SWP. 2010. Administration routines. Unpublished.
- SWP. n.d. Commencing marking phase. Unpublished.
- SWP. n.d. Parameter calculator user guide. Unpublished.
- SWP. n.d. e-marker +senior marker quick reference guide. Unpublished.
- SWP. n.d. Controlling marking quality walk through. Unpublished.
- SWP. 2014. Training schedule Work package. Unpublished.
- SWP. n.d. Quality control. Unpublished.
- SWP. 2014. Bureau training. Unpublished.
- SWP. 2014. Work schedule: November 2014. Unpublished.
- SWP. 2014. ZIMSEC project handover report, 13 November 2014. Unpublished.
- SWP. 2014. Some marking approach decisions. Email exchange with ZIMSEC staff, 29 April–12 May 2014.
- SWP. 2014. ZIMSEC 2014 – support planning. Email to Assistant Director TDR&E, 1 October 2014.
- SWP. 2015. ZIMSEC Nov-teleconference. Email to ZIMSEC staff, 25 February 2015.
- SWP. 2015. Mark scheme queries. Email to ZIMSEC staff, 28 October 2015.
- SWP. Planning day at ZIMSEC. Email to ZIMSEC staff, 12 September 2013.
- SWP. 2013. Script segmentation function. Email to ZIMSEC staff, 8 May 2013.
- SWP. 2013. Question paper blanks. Email to QPDM, 17 December 2013.
- SWP. 2013. Support while we're in transit. Email to ZIMSEC staff, 17 December 2013.
- SWP. 2013. ZIMSEC infrastructure for November 2013. Email to OSM Project Manager, 27 August 2013.
- SWP. 2013. ILAP pages. Email to QPDM, 8 April 2013.
- SWP. 2013. Question paper blanks. Email to QPDM, 22 October 2013.
- Suto, I. & Nádas, R. 2008. Towards a new model of marking accuracy: An investigation of IGCSE biology. Cambridge: University of Cambridge Local Examination Syndicate.
- Sweiry, E. 2013. A framework for the qualitative analysis of examinee responses to improve marking reliability and item and mark scheme validity. A paper presented

- at the 39th Annual Conference of the International Association for Educational Assessment. Tel Aviv, Israel.
- Taras, M. 2009. Summative assessment: The missing link for formative assessment. *Journal of Further and Higher Education*, 33(1): 57-69.
- Taras, M. 2012. Assessing assessment theories. *Online Educational Research Journal*, 3(12) Retrieved from <http://sure.sunderland.ac.uk/3321>. Accessed 20 February 2017.
- Taras, M. 2010. Back to basics: Definitions and processes of assessment. *Praxis Educativa*, 5: 123-130.
- Taras, M. 2005. Assessment: Summative and formative – Some theoretical reflections. *British Journal of Educational Studies*, 53(4): 466-478.
- Tarisayi, K.S & Manhivi, R. 2017. Social media tools in education: A case of WhatsApp use by Heritage Studies teachers in Zimbabwe. *Greener Journal of Social Sciences*, Vol. 7 (4), pp. 34-40.
- Tazviinga, L. 2019. ZESA faces 1.4 bn loss over power cuts. *The independent*, 23 August 2019. Retrieved from www.theindependent.co.zw. Accessed 25 September 2019.
- TDR&E. 2013. Request for blank copies of e-marked components for the November 2013 examination. Memo to Exams, 17 October 2013.
- Tisi, J., Whitehouse, G., Maughan, S. & Burdett, N. 2013. A review of literature on marking reliability. Coventry: Ofqual.
- Tracey, S.J. 2013. Qualitative research method: Collecting evidence, crafting analysis, communicating impact. West Sussex: Wiley-Blackwell.
- Triad 3. 2016. Research methodology in education: An introduction to document analysis. Retrieved from <https://lled500.trubox>. Accessed 10 January 2019.
- Tshili, N. 2019. Government commissions US\$5m ZIMSEC printing press. *The Chronicle*, 26 August 2019. Retrieved from www.chronicle.co.zw. Accessed 13 October 2019
- Tucker, B. 2009. Beyond the bubble: Technology and the future of assessment. Washington: Educational Sector
- Tze Ho, F. & Chong Sze, T. 2013. Marking and grading procedures for 2012 HKDSE Liberal Studies examinations. *Hong Kong Teachers' Centre Journal*, 12: 1-19.
- Thompson, N.A. 2016. Reliability and validity. Minnetonka: Assessment Systems Corporation.

- UNDP. 2018. WhatsApp surveying guide: Lessons learnt from two qualitative WhatsApp surveys in Lebanon. Oxford: UNDP.
- University of Cambridge Local Examinations Syndicate. 2008. How have school exams changed over the past 150 years? Retrieved from www.cambridgeassessment.org.uk/news/how-have-school-exams-changed. Accessed 5 January 2018.
- Wajdi, M., Khalil, A. & Maria, N.P.A. 2014. Training strategies, theories and types. *Journal of Accounting – Business & Management*, 21(1): 12-26.
- “WhatsApp has come in to fill the void”: In Zimbabwe, the future of news is messaging. The independent: Voice of the voiceless: Retrieved from www.theindependent.co.zw. Accessed 25 September 2019.
- Weiss, D.J. 2011. Item banking, test development and test delivery. In Keisinger, K.F (Ed.). *The APA handbook on testing and assessment*. Washington DC: American Psychological Association. *In Press*.
- Winkley, J. 2010. E-assessment and innovation. Retrieved from www.becta.org.uk. Accessed 20 December 2013.
- William, D. 2014. Formative assessment and the contingency in the regulation of learning process. Paper presented at the Meeting of American Educational Research Association. Philadelphia, USA.
- World Health Organisation (WHO). 2017. Purposeful sampling for qualitative studies. Retrieved from <http://apps.who.int/medicinedocs/en>. Accessed 5 August 2018
- Yan, Z. & Coniam, D. 2014. The effects of key demographic variables on markers’ perceived ease of use and acceptance of onscreen marking. *Assessment. Education: Principles, Policy & Practice*, 21(4): 464-480.
- Yazan, B. 2015. Three approaches to case study methods in education: Yin, Merriam and Stake. *The Qualitative Report*, 20(2):134-152.
- Yin R. 2013. Case study research: Design and methods. SAGE: Thousand Oaks
- Yin, R.K. 2011. Qualitative research: From start to finish. New York: The Guildford Press
- Yin, R. 2004. Case study methods. Revised draft p1-28. Available on <https://www.academia.edu> Accessed 5 August 2018
- Yin, R.K. 2003. Case study research design and methods. Washington: Sage.

Zhou, T. 2017. ZIMSEC still to buy printing press-Dokora. Bulawayo24 News, 12May 2017. Retrieved from www.bulawayo24.com. Accessed 22 February 2019.

Zimbabwe national policy on for information communication technology. 2015. Harare: Government of Zimbabwe.

ZIMSEC Act Number 17 of 1994. 1994. Harare: Government Printers.

ZIMSEC commissions printing press. Retrieved from www.zimsec.co.zw. Accessed 13 October 2019.

ZIMSEC. 2017. Continuous assessment framework. Harare: ZIMSEC.

ZIMSEC November 2016 O Level centralized marking: 5008/2 team structure. Unpublished

ZIMSEC entry procedure booklet: 2008-2012. Unpublished

ZIMSEC Examination Circular Number 8 of 2015. Harare: ZIMSEC.

ZIMSEC Examination Circular Number 10 of 2015. Harare: ZIMSEC.

ZIMSEC Nov-Dec 2014 Marking administration activity schedule. Unpublished.

ZIMSEC. 2013. Quality policy manual. Harare: ZIMSEC.

ZIMSEC Examination Circular Number 41 of 2013. Harare: ZIMSEC.

ZIMSEC Examination Circular Number 42 of 2013. Harare: ZIMSEC.

ZIMSEC November 2016 O Level centralized marking: 5008/4 team structure. Unpublished

ZIMSEC minutes of the marking commissioning meeting, 30 June 2017. Unpublished

ZIMSEC minutes of the commissioning of July 2015 marking programme by the Director, 3 July 2015. Unpublished.

ZIMSEC minutes of a meeting held by TDR&E, 5 February 2014.

ZIMSEC minutes of the marking commissioning meeting, 2 December 2016. Unpublished.

ZIMSEC November 2014 proposed marking sessions for e-marking. Unpublished.

ZIMSEC lessons learnt report: June/November 2017 Live marking. Unpublished.

ZIMSEC lessons learnt report: June 2015 Live marking. Unpublished.

ZIMSEC lessons learnt report: June/November 2014 Live marking. Unpublished.

ZIMSEC lessons learnt report: June 2013 Live marking. Unpublished.

ZIMSEC lessons learnt report: November 2013 Live marking. Unpublished.

ZIMSEC November 2012 E-marking training programme. Unpublished.

ZIMSEC June/November 2017 question paper development record. Unpublished.

ZIMSEC O Level Biology syllabus (5008): 2011-2020. Harare: ZIMSEC

ZIMSEC Question papers: 5008/2, November 2015; June 2016; November 2016.

ZIMSEC Question papers: 5008/3, November 2013; November 2014.

ZIMSEC Question papers: 5008/4, November 2013 – November 2016.

ZIMSEC Mark schemes: 5008/2, November 2015; June 2016; November 2016.

ZIMSEC Mark schemes: 5008/3, November 2013; November 2014.

ZIMSEC Supervisors' reports: 5008/3, November 2013.

ZIMSEC. 2003. Procedure booklet for test development. Harare: ZIMSEC.

ZIMSEC.n.d. Mark capturing user manual. Unpublished.

ZIMSEC begins exams e-marking. *Bulawayo24 News*, 14 July 2012. Retrieved from <https://bulawayo24.com/index-id-news-sc-education-byo>. Accessed 11 January 2014.

ZIMSEC selects GradeMaker to supply item banking technology.2016.Retrieved from <https://www.grademaker.com/about/news>. Accessed 17 July 2018.

ZIMSEC begins exams e-marking. *The Chronicle*, 14 July 2012. Retrieved from www.chronicle.co.zw/zimsec-begins-exams-e-marking/. Accessed 11 January 2014.

ZIMSEC sets pace in e-marking. *The Herald*, 19 August 2015. Retrieved from www.herald.co.zw/zimsec-sets-pace-in-e-marking/. Accessed 17 July 2018.

ZIMSEC. n.d. O Level Physics syllabus 5055: 2011-2020. Harare: ZIMSEC.

ZIMSEC. n.d. O Level Chemistry syllabus 5071: 2011-2020. Harare: ZIMSEC.

ZIMSEC acquires printing machine. *Pindula News*, 25 October 2018. Retrieved from www.news.pindula.co.zw. Accessed on 22 February 2019.

Zimbabwe Curriculum Development and Technical Services. 2015. Biology syllabus forms 3-4: 2015-2022. Harare: Curriculum Development and Technical Services.

Zimbabwe Ministry of Primary and Secondary Education. 2015. Curriculum framework for primary and secondary education: 2015-2022. Harare: Government of Zimbabwe.

Zimbabwe Ministry of Primary and Secondary Education. 2018. Infrastructural development needs in education. Paper presented at Presidential Indaba on infrastructure development. Borrowdale, Harare.

Appendices

Appendix A: Letter to the Director of the Zimbabwe School Examination Council

Request for permission to carry out an educational research at the ZIMSEC

Date: 29 May 2019

Title: Exploring the Practice of Quality Control in the Onscreen Marking of Ordinary Level Biology in Zimbabwe

The Director

The Zimbabwe School Examinations Council

P. O. Box CY 1464

Causeway

Harare

The Director

I, Ebba Masiri, am doing research towards a PhD under the supervision of Professor M.T. Gumbo in the College of Education at the University of South Africa. I am requesting permission to conduct the research at your organisation. The study is entitled: Exploring the Practice of Quality Control in the Onscreen Marking of Ordinary Level Biology in Zimbabwe. I am conducting the study to explore the experiences and perspectives of Examiners and Subject Managers about the influence of examiner training and standardisation, question papers and mark schemes and the mechanisms of monitoring marking on the practice of quality control in the marking of O Level Biology examinations in the old curriculum. The study focuses on the onscreen marking of O Level Biology (5008) between 2013 and 2017, before the implementation of the new curriculum.

Your examiners and subject managers have been purposefully selected because of their knowledge, expertise in the marking of O Level Biology examinations for syllabus (5008), which was last marked on screen in November 2017.

The study will encompass data collection using document review and semi-structured interviews. All participants will be requested to complete the consent form. Furthermore, all participants will be given a choice to willingly participate and are allowed to withdraw anytime without giving reasons. There will be no penalty given for withdrawal. In this study there are no risks and rewards involved. It is anticipated that the outcome of this research will provide important insights into the practice of quality control in the OSM environment as well as provide a framework that could guide the practice.

An electronic feedback of the findings will be emailed to interested participants upon request.

Yours sincerely



Ebba Masiri

Researcher

Appendix B: Participant information sheet

Date_____

Title: Exploring the Practice of Quality Control in the Onscreen Marking of Ordinary Level Biology in Zimbabwe

Dear prospective participant

My name is Ebba Masiri and I am doing research towards a PhD under the supervision of Professor M.T Gumbo, a Professor in the College of Education at the University of South Africa. I am inviting you to participate in a study entitled *Exploring the Practice of Quality Control in the Onscreen Marking of Ordinary Level Biology in Zimbabwe*.

What is the purpose of the study?

This study is expected to collect important information that could lead to a framework that could guide the practice of quality control in the OSM environment in Zimbabwe and similar contexts.

Why am I being invited to participate?

You are invited because of your experience with the OSM of O Level Biology (5008) examinations. I obtained your contact details from the Zimbabwe School Examinations Council's examiner database. Fifteen O level Biology Examiners will participate in this study.

What is the nature of my participation in this study?

The study involves semi-structured interviews where you will answer questions relating to quality control in the OSM environment. The interview will be 30 minutes long.

Can I withdraw from this study even after having agreed to participate?

Participating in this study is voluntary and you are under no obligation to consent to participation. If you do decide to take part, you will be given this information sheet to keep and be asked to sign a written consent form. You are free to withdraw at any time and without giving a reason.

Will the information that I convey to the researcher and my identity be kept confidential?

By participating in this study, you make an important contribution to knowledge on the practice of quality control in Zimbabwe, which is the first African country to mark examinations on computer screens.

Are there any negative consequences for me if I participate in the research project?

Your participation in this study will disrupt some of your normal activities when you make time to participate in the interviews.

Will the information that I convey to the researcher and my identity be kept confidential?

Your name will not be recorded anywhere and no one will be able to connect you to the answers you give. Your answers will be given a code number or a pseudonym and you will be referred to in this way in the data, the research report, any publications, or other research reporting methods such as conference proceedings. However, your answers may be reviewed by people responsible for making sure that research is done properly, including the transcriber, external coder, and members of the Research Ethics Review Committee. Otherwise, records that identify you will be available only to people working on the study, unless you give permission for other people to see the records. Your anonymous responses may be used for other purposes such as conference proceedings, research report and journal articles.

How will the researcher(s) protect the security of data?

Hard copies of your answers will be stored by the researcher for a period of five years in a locked cupboard/filing cabinet at Unisa Library for future research or academic purposes; electronic information will be stored on a password protected computer. Future use of the stored data will be subject to further Research Ethics Review and approval if applicable. After the five years, hard copies will be shredded and electronic copies will be permanently deleted from the hard drive of the computer.

Will I receive payment or any incentives for participating in this study?

You will not receive payment for participating in this study and you will not incur any expenses.

Has the study received ethics approval?

This study has received written approval from the Research Ethics Review Committee of the College of Education, Unisa. A copy of the approval letter can be obtained from the researcher if you so wish.

How will I be informed of the findings/results of the research?

If you would like to be informed of the final research findings, please contact Ebba Masiri on +236776002853/+263712504772 or email rumapox@gmail.com. The findings are accessible for three months. Should you have concerns about the way in which the research has been conducted, you may contact Professor M.T. Gumbo on +27823258353 or gumbomt@unisa.ac.za.

Thank you for taking time to read this information sheet and for participating in this study.

Thank you.



Ebba Masiri

Appendix C: Consent (Return slip)

I, _____ (participant name), confirm that the person asking my consent to take part in this research has told me about the nature, procedure, potential benefits and anticipated inconvenience of participation.

I have read (or had explained to me) and understood the study as explained in the information sheet.

I have had sufficient opportunity to ask questions and am prepared to participate in the study.

I understand that my participation is voluntary and that I am free to withdraw at any time without penalty (if applicable).

I am aware that the findings of this study will be processed into a research report, journal publications and/or conference proceedings, but that my participation will be kept confidential unless otherwise specified.

I agree to the recording of the interview.

I have received a signed copy of the informed consent agreement.

Participant Name & Surname (please print) _____

Participant Signature

Date

Researcher's Name & Surname (please print) _____

Researcher's signature

Date

Appendix D: Document analysis form

Feature	Notes
Author/Creator	
Context (place and time of document creation)	
Intended audience	
Purpose for document creation	
Type of document (pamphlet, newspaper, memo, etc)	
Main points expressed in the document	
Relevance of main points to research questions: Examiner training and standardisation	
Monitoring quality	

Feature	Notes
of marking	
Influence of questions and mark schemes on quality of marking	
Challenges of quality control in the OSM environment	
Conclusion	

Appendix E: Interview schedule for subject managers

Date.....

Name of Interviewer.....

Name of interviewee.....

Introduction

Thank you very much for sparing your precious time to answer my questions on the practice of quality control in the OSM of Biology (5008) examinations. I will be recording your responses (if it is ok with you), so that I can listen carefully to your responses. I will, however, be writing a few notes.

1. What was your role in the marking of Biology (5008) examinations?

.....
.....

2. How were you trained to perform your role in the OSM environment?

.....
.....
.....
.....
.....

3. Describe the selection and training of Examiners who marked Biology (5008) examinations in the OSM environment, highlighting how the training prepares them to mark accurately.

.....
.....
.....

.....

.....

.....

4. Describe the pre-standardisation activities that promoted accuracy of marking in the OSM environment.

.....

.....

.....

.....

.....

.....

5. Describe how seeds worked, emphasising on the following:

1.1 Who set the seeds

.....

5.2 Criteria for selecting the seeds

.....

.....

.....

.....

.....

.....

a. The criteria for determining the number of seeds

.....

.....

.....

.....

.....

.....

5.4 Dealing with wrong seeds

.....

.....

.....

.....

.....

.....

5.5 Dealing with stopped markers (Probe on mechanisms to ensure that stopped markers are trained before they are activated).

.....

.....

.....

.....

.....

.....

.....

5.6 Criteria for permanently stopping a marker

.....

.....

.....

.....

.....

.....

6. What feedback was provided examiners about the quality of their marking?

.....

.....

.....

.....

.....

.....

.....

7. What criteria were used to select Biology (5008) examinations that were marked on screen from 2013 to 2017? Probe on (i) the structure of the question paper (ii) the type of questions and the length of answers.

.....

.....

.....

.....

.....

.....

.....

.....

.....

8. Evaluate the ability of the OSM technology to promote the quality and efficiency of marking Biology examinations in the Zimbabwean context. Emphasise on opportunities and challenges.

8.1 Opportunities

.....

.....

.....

.....

.....

.....

8.2 Challenges

.....

.....

.....

.....

.....

9. Besides examiner training and standardisation, what other activities promoted quality of marking Biology (5008) examinations in the OSM environment?

.....

.....

.....

.....

.....

.....

.....

10. What else can you say about the quality of marking O level Biology (5008) examinations marked on computer screens?

.....

.....

.....

.....

Thank you very much for taking your time to participate in the interview. I will summarise your responses and contact you for verification of the summary.

Appendix F: Interview schedule for examiners

Date: 24 – 26 July 2019

Name of Interviewer: Masiri E

WhatsApp group discussions

Introduction

Greetings to you ladies and gentlemen. I am collecting data for my PhD thesis on quality control in the e-marking of O Level Biology. I am kindly asking you to share with me your experiences on the topic on this platform. I will ask questions then you respond. I am seeking your consent before I ask questions.

1. How were you selected and trained for e-marking?
2. Comment on the relevance of the training to e-marking.
3. Can you share your experiences of standardisation meetings? Focus on adequacy of time, relevance to e-marking and the extent to which the meetings enhanced quality of marking.
4. Please share your experience with seeds.
5. Senior markers were supposed to discuss the failed seeds with stopped markers. Please share your experience of these discussions.
6. What is your opinion on the type of papers that should be e-marked? Can any paper be e-marked without compromising the quality of the exam?
7. I heard there were challenges when Paper 2 was first e-marked in the November 2015 session. What were the challenges?
8. How did marking Section A compare with marking Section B?
9. What are the advantages of e-marking over manual marking?
10. What were the challenges that could compromise quality of marking?

Thank you for participating in the interview. I will remain in the group until the end of August so that I can ask follow-up questions.

Appendix G: Face-to-face Interview transcription

Date: 28 June 2019

Name of Interviewer: Masiri E

Name of interviewee: SM1

Thank you very much for taking part in the interview. I have summarised your responses to the interview questions. I kindly request you to read the summary, add some points that I might have missed and remove any points that I might have erroneously added.

1. What was your role in the marking of Biology (5008) examinations?

- Training Senior markers: quality control (seed setting)

2. How were you trained to perform your role in the OSM environment?

- There was no formal training: I learnt by discovery
- Groups of subject managers were learning as they were working
- At first we presumed that script portions were missing, until we discovered that they were never missing, but they had not been loaded onto the system.

3. Describe the selection and training of Examiners who marked Biology (5008) examinations in the OSM environment, highlighting how the training prepares them to mark accurately.

- There was no special training for examiners
- All examiners were migrated from paper based marking to e-marking with their roles
- For some subjects, it turned out that older markers had no computer skills. Younger markers had to take on senior roles because of their technological competencies
- The examiners were trained on the job
- New examiners were trained on e-marking as they came in

4. Describe the pre-standardisation activities that promoted accuracy of marking in the OSM environment.

- The meeting is face-to-face
 - Prepare for the big meeting
 - Major issues relating to the examination are discussed
 - Problems that might arise during marking are discussed (can you please give examples of the problems?)
 - Seeds were not disc
 - Mark schemes are edited and uploaded
 - Pre-live scripts were replaced by qualification seeds
 - Qualification seeds were marked at the start of every day
- Qualification seeds

Pre-live marking:

- Examiners work in the training mode
- All examiners (senior and normal markers go through pre-live marking
- The examiners mark scripts for their subject or for any other subject

Qualification:

- Qualification seeds were marked for the first two days
- Examiners were not informed about the duration of the qualification so that they exercise caution during marking
- Fast examiners would have finished marking in the first two days, qualification would therefore not be useful thereafter.

5. Describe how seeds worked, emphasising on the following:

5.1 Who set the seeds?

- Senior markers; soon after the pre-live marking
- Seeds are set in the live mode

5.2 Criteria for selecting the seeds

- not tricky
- Should be legible
- Something worth discussing
- No blanks; there must be something written
- Reasonable
- Legible

The role of qualification seeds

- Marked in the first two days for qualification to mark
 - The markers would not be told about the number of days they would mark qualification
- Probe: why would qualification seeds be marked for two days only?
- Fast examiners would have completed their portions within the first two days, so qualification would not be useful anymore

5.3 The criteria for determining the number of seeds

- Determined by the formula
- Depends on the number of examiners and candidates
- Number of marking days varied, with fast markers taking 2-3 days

5.4 Dealing with wrong seeds

- Identified in the discussions of seniors and stopped markers
- Deleted and sent back for marking

5.5 Dealing with stopped markers (Probe on mechanisms to ensure that stopped markers are trained before they are activated).

- Markers are stopped by qualification and seeds.
- Stopped marker approached senior marker
- Some stopped markers would just sit until the administrators notice that they are not marking
- That's when they start approaching the senior marker for discussions
- No mechanism for enforcing discussions between senior markers and stopped markers.
- They might even activate an examiner without discussion

5.5 Criteria for permanently stopping a marker

- The icon has only been used accidentally
- It can only be used when markers are marking from home
- Deviation had been minimised by marking a series of questions

6. What feedback was provided to examiners about the quality of their marking?

- Unless the marker requests, no feedback is given

- Quality of marking report is visible to the administrator only.

7. What criteria were used to select Biology (5008) examinations that were marked on screen from 2013 to 2017? Probe on (i) the structure of the question paper (ii) the type of questions and the length of answers.

- The papers had to be restructured
- Short answer questions
- Questions worth more than four marks would give challenges
- The examinations were allegedly watered down when the question papers were highly structured
- If we do not resume e-marking it is better we go back to essay type questions in Section B

8. Evaluate the ability of the OSM technology to promote the quality and efficiency of marking Biology examinations in the Zimbabwean context. Emphasise on opportunities and challenges.

8.1 Opportunities

- Deviating markers were stopped by seeds
- E-marking is faster than paper based marking
- No transcription and addition errors

8.2 Challenges

- Duplication of script pages; such scripts could not be marked on screen
- Blurred images
- Problem scripts had to be pulled out and externally marked
- Scanning was done overnight, the operators probably got tired
- Intended to visit the scan centre to observe the process but e-marking was suspended
- Erratic internet
- Power cuts

9. Besides examiner training and standardisation, what other activities promoted quality of marking Biology (5008) examinations in the OSM environment?

- Percentage double marking is another quality control system

Probe: Was the S-Process ever used for quality control?

- S-Process was never used for quality control; it is probably to mark essays.

10. What else can you say about the quality of marking O level Biology (5008) examinations marked on computer screens?

- Quality marking can be enhanced when examiners mark from home
- Centralised marking exerts pressure on examiners and administrators

Case: Two SMs and senior markers had remained at the venue to finish off scripts that were popping up. We received phone calls from office by someone asking “what are you still doing there’. When we got to the hotel where we were accommodated we were told that we should check out because somebody from our office had phoned to say we should check out that day. We had to move to another facility so that we could finish off marking.

- It would be better to mark from home. Technological infrastructure is a challenge though.

Thank you very much for taking your time to participate in the interview.

Appendix H: WhatsApp chats transcription

WhatsApp Platform: 24 -26 July 2019

4025/2

Question	Username	Response
Introduction	SM3	That's madam Masiri ladies and gentlemen
	NM11	Welcome Madam
	Me	Greetings to you ladies and gentlemen. I am collecting data for my PhD thesis on quality control in the e-marking of O Level Biology. I am kindly asking you to share with me your experiences on the topic on this platform.
	NM12	Experience on which aspects precisely?
	Me	I will ask questions then you respond. I am seeking your consent before I ask questions.
	NM13	Cushion first – laughs(<i>in apparent reference to airtime</i>)
	Me	How much (laughing as well) <i>I later sent ZW\$10 (R20 at that time) to everyone who responded to my questions on 24 July 2019. The participants must have informed each other that 'cushion' is being paid to those who are responding to research questions. More and more responses started coming on the 25th of July 2019.</i>
	NM13	Granted Mam
How were you	NM13	Applied after seeing a newspaper advert for markers.

Question	Username	Response
<p>selected and trained for e-marking?</p> <p>Probe: Was the test used for selection?</p>		<p>Training was mainly marking dummies, no computers involved</p> <p>We also wrote a test. O level paper</p> <p>Not sure if the test was used for selection but it was part of the training</p>
	NM11	<p>Normal marker for paper 2</p> <p>Selected for manual marking and later trained for e-marking</p> <p>Yes the test was used as well as proper marking of more than 15 dummies.</p>
	NM17	<p>Normal marker for Paper 2</p> <p>I was trained as a marker before e-marking.</p> <p>Since I was already a marker I received a letter to attend the marking session for e-marking. I had previously been trained to mark and when e-marking started I was trained to e-mark at the standardisation meeting.</p>
<p>Probe: Did you apply for marking during or before e-marking?</p>	NM12	<p>Normal marker</p> <p>I think those who can answer the question on selection are those who selected us. Training was compulsory, step by step and practical.</p> <p>I applied for marking during e-marking</p>

Question	Username	Response
	NM14	All those who were involved in e-marking was the young generation who were computer literate Yes the test was used for selection
Probe: Was the training specific to e-marking or to marking in general?	NM14	Marking in general at Belvedere but when we went to Chinhoyi we were trained on how to use the computer and marking using the ZIMSEC portal.
	NM15	We were trained for marking in general, then when e-marking was introduced we were trained for e-marking
	NM16	Marking in general, later e-marking
	SNR9	Senior marker Yes, it was necessary
Comment on the relevance of the training to e-marking.	NM13	Complex to simple
	NM11	The training was relevant because the main aspects of marking rest on adherence to the marking scheme, which was mainly done manually , then e-marking was the marker's ability to use the computer effectively
	NM17	It was difficult due to the fact that we were trained through coordination rather than being given separate time for training
Can you share your	NM12	Normal marker

Question	Username	Response
<p>experiences of standardisation meetings? Focus on adequacy of time, relevance to e-marking and the extent to which the meetings enhanced quality of marking.</p>		<p>Time was adequate considering that everyone grasped the e-marking idea within the specified time, and the software was easy to use.</p> <p>Training was relevant as it enabled new markers to familiarise with all the functions available on the software</p> <p>Use of dummies and seeds for training addressed the issue of quality and uniformity in marking.</p>
	NM11	<p>Whilst I agree that the marking and discussion of seeds do great to enhance quality marking, I had reservations on the number of dummies/seeds used for standardisation. I think more time and more seeds should be used if we are not to under mark or over mark some candidates. Experience with e-marking shows that some seeds have a pattern of showing up or are kind of recognizable during live marking. in this case the marker takes due care on such scripts, or in some cases the marker deviates from the mark scheme which is obviously not recognizable if they happen not to be stopped and in this case quality marking is compromised</p> <p>With e-marking, there is no room for inclusion of new ideas that might arise during live marking, thus there is need to take more time exploring all possible responses to</p>

Question	Username	Response
		<p>a question.</p> <p>Whilst in principle the marker can escalate issues, it is difficult to ascertain if that happens in practice. If more time is taken on standardisation, no marker should be stopped during marking, otherwise inconsistent markers can continue marking for as long as they evade seeds.</p>
	NM17	<p>Time was inadequate though everyone grasped the e-marking idea within the specified time, the software is user friendly.</p> <p>Use of dummies and seeds for training addressed the issue of quality and uniformity in marking. however, poor seeds set by senior markers made the first encounter in Chinhoyi a nightmare for many markers</p>
	NM18	In response to NM11: couldn't have said it better
Probe: Let's explore the issue of escalations. What kinds of scripts were escalated by examiners?	NM11	Where the handwriting/print is invisible and when the candidate's response seemingly correct but not on the e-marking guide and the marker is in doubt.
	NM14	<p>Normal marker, later became verifier</p> <p>Scripts with questions without answers or with blank</p>

Question	Username	Response
		questions
Probe: what action did senior markers take on escalated issues?	NM11	There are obviously rectified but worry is on those that may not be escalated and the marker is not stopped because the scripts might not be a seeds. What quality control is there for marked scripts using the e-marking?
Please share your experience with seeds.	NM12	<p>Correct points that do not appear on the marking scheme are ignored. If you try to mark them correct on a seed, you will be stopped from marking.</p> <p>Seeds also ensure adherence to the marking scheme.</p> <p>Some seeds do not allow the markers to make their own decisions especially where the marking schemes says alternative wording. What I feel is correct might appear incorrect to another person.</p>
	NM19	<p>Seeds need to be increased in number to ensure proper quality control. They impede marking speed yes, however, if more time is allocated to the marking period it will improve us as markers.</p> <p>This may help avoid strain caused by the process considering that it is a sedentary job with little movement, if more time is allocated alertness and level of concentration increase, hence increased quality of marking.</p>

Question	Username	Response
Suggest how additional answers can be added after marking has started.	NM12	That will be tricky I think. Altering the marl scheme would mean remarking all the marked scripts. To avoid that more dummies are supposed to be used rather than leaving other valid points.
Senior markers were supposed to discuss the failed seeds with stopped markers. Please share your experience of these discussions.	NM12	<p>I did e-marking once so I have insignificant experience.</p> <p>Discussion done if the marker doesn't understand why they failed the seeds. If you revisit the failed seeds and see your mistakes then there is no need for discussions, but if you don't see your mistakes then you discuss with your senior.</p> <p>In some (rare) cases, the seeds might be the ones with errors, resulting in markers being stopped.</p>
	SNR9	It's easy, destroy the bad seeds and better ones
Probe: How were wrong seeds dealt with?	NM12	Removed I think
In response to no particular question	NM20	Guys am worried about e-marking on the fact that it comes a time when even wrong answers will be accepted, like there is a loophole somewhere.
Probe: Please explain why you think wrong answers can be accepted.	NM20	There are times when seeds are not available. I believe at those times wrong answers can be marked correct. I realised that after qualification the speed increases and concentration reduces.

Question	Username	Response
What are the advantages of e-marking over manual marking?	NM20	Fast and less strenuous
	NM21	Fast; enough sleeping time(rest); less labour; no addition markers; slow markers not stressed (trying to catch up with others in the group stresses).
	NM19	It is very efficient, as a verifier I know it reduces time Also independent of each marker, no need to wait for another person to complete a pile you need to mark.
	NM11	It is fast; reduces error on counting marks; no question is awarded more marks than it deserves By not seeing candidates' names or centers, it reduces malpractice; no need to enter marks at the end of the session.
Where there any challenges that could compromise quality of marking?	NM11	Unavailability of electricity and internet. A case in Gweru 2017, more often savers would be down and markers had to rush at the last hour to beat the time allocated for staying at MSU; marking late into the night, with fatigue obviously quality is compromised. The immovable nature of desktops creates a lot of discomfort. Laptops can enable a lot of posture flexibility. Ndakarwara mutsipa nomusana after the session. (I had backache and sore neck after the session)
	NM12	Scanning: when a question takes a small portion of a page

Question	Username	Response
		<p>the whole page was scanned making it difficult to read. Zooming in and out will be required for such questions, marking becomes slow.</p> <p>When a question takes a small portion of the page, only that portion should be scanned. Scanning the whole page would greatly reduces the picture quality.</p>
Did you get any feedback about the quality of your marking, maybe by clicking an icon?	NM11	Despite that option being available, surely there wasn't enough time to constantly check for that at all.
	SNR9	No feedback could be made
	SNR10	No feedback was possible. No one knew how many good portions they had marked except the counter on how many you had marked for each part question.
What is your opinion on the type of papers that should be e-marked. Can any paper be e-marked without compromising the quality of the exam?	SNR9	Yes, provided that quality seeds are set without rushing
	NM12	In my opinion any paper can be e-marked though questions requiring descriptions and explanations are difficult to mark using the e-platform; easy for questions that require

Question	Username	Response
		candidates to name, state, list etc
Is it possible to mark the biology practical exam on computer screens without prejudicing candidates?	SNR22	It's quite difficult because centres have conditions which are unique so sometimes you need to consult the reports from each centre
Let's talk about seed setting. Senior markers please share your experiences.	SNR9	Quality seeds refers to clear marking guidelines; not to seed dubious ones
What criteria were used to select seeds?	SNR9	Not zero mark or blank; written, with clear marking points
	SNR22	Clear; legible' Each question had a minimum and maximum number of seeds
	SNR23	Seeds to be selected must be straightforward. Seeds must not attract debate. Remember seeds are there to check the alertness of examiners not for them to act like dummies.
Probe: what would happen if the seeds were below minimum?	NM20	Quality of marking could get low

Question	Username	Response
	SNR22	The senior marker was required to add some more
Probe: Did the system allow examiners to continue marking when the seedbank was below minimum?	SNR22	Yes
Where markers stopped by seeds?	SNR22	In which section?
Both A and B	SNR9	Yes, very much. Even those who seeded To James M: Bothe sections but more importantly in B
	SNR22	Too many stops in Section B
	SNR9	To James M: Yes
I heard there were challenges when Paper 2 was first e-marked in the November 2015 session. What were the challenges?	SNR24	I think the training was not enough mainly to the senior markers who were asked to prepare the seeds; the training was done hurriedly and also the marking period was not enough.
	SNR10	There was inadequate orientation of those who were to set the seeds. As a result poor seeds were set. The poor seeds took too long to be removed. Coupled with a high number of qualification seeds, most examiners failed to qualify and

Question	Username	Response
		<p>wasted 3 to 3 days marking seeds.</p> <p>The questions in Section B of the paper not ideal for e-marking as they had many marking points.</p>
	NM21	I suggest that adequate time is required for seeding and senior markers should mark for a day, identifying and removing poor seeds before others start marking.
How did marking Section A compare with marking Section B?	SNR9	Section A was far simpler than B where more time was required to read the answer
	SNR22	Section A was much easier and faster.

Appendix I: Question paper review

Question Paper: 5008/2 November 2015

Feature	Notes
Author/Creator	ZIMSEC: Test Development Division
Context (place and time of document creation)	ZIMSEC Head Office; November 2015 Exam Session
Intended audience	Candidates; Examiners
Purpose for document creation	Examination
Type of document (pamphlet, newspaper, memo, etc)	Question paper
Main points Expressed in the documents	<ul style="list-style-type: none">- It is a theory paper- 2 hours long; 100 marks- No instruction to use black ink as in 5008/4 for November 2015 and 2016.- Candidates are instructed to write their details (name candidate number and centre number) on every page.- Candidates instructed to check if there are missing or duplicate pages and ask for replacement of booklet if there are duplicate or missing pages.- Section A: answer all questions – 40 marks- 5 questions with total marks varying from 7-10- Each question had sub-questions, e.g. 1(a)(i)-(ii), 1(b), 1(c) etc

Feature	Notes
	<ul style="list-style-type: none"> - Short answer question - The majority testing objective 1: Knowledge and understanding - A few questions testing Objective 2 : Handling information and problem solving - One stray question worth 3marks addressing skill 3: Practical skills - Section B: answer any three out of 5 questions - 60 marks - Each questions worth 20 marks - Each question has sub-questions, e.g. 6(a), 6(b) etc - Some part questions demanded extended answers worth a maximum of 12 marks - Majority of questions testing skill 2 - A few questions testing skill 1 - Write answers on spaces provided - Variations in the spaces for same type of response worth the same number of marks. <p><i>8(b)(ii) Suggest measures that can be taken to reduce drug abuse in Zimbabwe: 10 lines for 6 marks</i></p> <p><i>9(b) Describe the inheritance of Down's syndrome in humans: 8 lines for 6 marks</i></p> <p><i>10(b) Explain why mothers who smoke when they are pregnant are likely to have small babies: 13 lines for 6 marks.</i></p>
<p>Relevance of main points to research questions:</p> <p>Examiner training and standardisation</p>	<ul style="list-style-type: none"> - The paper might need more time for training and standardisation than paper 4 because it is longer. - The PMS report indicated that the paper needed more time for coordination

Feature	Notes
	<ul style="list-style-type: none"> - The pre-standardisation meeting was rushed leading to poor understanding and application of the mark scheme by both normal and senior markers (PMS Report).
Monitoring quality of marking	<ul style="list-style-type: none"> - Seeding approach to quality control - Most examiners failed qualification seeds (PMS report and D-Grade reports) - Some examiners (Senior and normal) abandoned the marking exercise (PMS' report, January 2016; D-grade reports January 2016). - Hierarchical seeding approach for constrained papers (Pinot de Moira, 2013; DRS, 2013; Hudson, 2009, Roan, 2009 personal experience as subject manager). - True score determined by the senior marker (Pinot de Moira, 2013:16) - Errant markers can be identified by the system and stopped from marking; errant markers need training during marking (Pinot de Moira 2013; Hudson, 2009; DRS, 2013) - Higher chances of candidates being awarded marks within 10% of the true mark (Pinot de Moira, 2013:12) - Senior markers should train errant markers (Pinot de Moira, 2013; DRS 2013; Hudson 2009; Roan, 2009) - Small tolerance ranges for quality control, enabling greater chances of identifying errant markers (Pinot de Moira, 2016:17) - Any examiner comments about seeding approach to quality control (Raikes 2004; 19; Coniam (2011a:1045)?
Influence of questions and mark schemes on quality of marking	<ul style="list-style-type: none"> - Section A: easier to mark than section B where longer answers are required (Ahmed & Pollit, 2011; Bramley 2008). - Section B of the paper needed to be adapted to e-marking (PMS report). This call by the PMS greatly influenced the design of question papers in future examinations (5008/2 June and November 2016; June

Feature	Notes
	<p>and November 2017).</p> <ul style="list-style-type: none"> - Section B questions much longer than in the subsequent examinations. E.G J 2016 with a maximum of 5 marks. - Several questions allegedly derailed marking process because of lengthy responses (PMS report). - The questions in section A elicited shorter responses than Section B. - All responses can be marked accurately with Level 3 constrained mark scheme (Ahmed & Pollit, 2011:267) - Higher level of marker agreement during marking (Bramley 2008:2). - The type of questions in the paper can be distributed to the three types of markers: clerical, graduate and expert markers (Ofqual, 2014e; Raikes et al, 2004; Suto and Nádas, 2008:9; Meadows & Billington 2013:9)
opportunities of quality control in the OSM environment	<ul style="list-style-type: none"> - Automatic quality control enhances quality marking - Accurate marking of short questions in section A - Shorter marking period when questions are marked by three types of examiners. - Identification of errant markers by the quality control system - Opportunity for training of errant markers during marking
Challenges of quality control in the OSM environment	<ul style="list-style-type: none"> - The first paper to have a constrained section B. Possibilities of candidates writing on inappropriate spaces when they are provided with limited answer spaces (AQA challenge: Ofqual, 2011) - The PMS report requested that the paper be designed to make it compliant with e-marking. Risk of designing questions to suit the demands of technology and compromising the validity of the examinations (Roan, 2009; Pinot de Moira, 2013). - It is the paper that was abandoned by some examiners at marking

Feature	Notes
Conclusion	<p>The paper was marked on screen for the first time in this examination session. Section A is easier to mark than Section B. Automated seeding approach had the opportunity to enhance quality of marking. However, the majority of examiners had challenges with the qualification seeds, which they failed (PMS report). Section B was constrained for the first time (Examination Circular Number 8 of 2015). There is no criterion for determining the amount of answer spaces. Providing answer spaces on the question paper might limit candidates to shorter responses, prompting them to write on wrong spaces. Examiners might miss such answers, resulting in remark requests (Ofqual 2011: the AQA challenge)</p>

Appendix J: Mark scheme Review: 5008/2 November 2015

Feature	Notes
Author/Creator	ZIMSEC: Test Development Division
Context (place and time of document creation)	ZIMSEC Head Office; November 2015 Exam Session
Intended audience	Examiners
Purpose for document creation	Examination
Type of document (pamphlet, newspaper, memo, etc)	Mark scheme: 5008/2
Main points expressed in the document	<ul style="list-style-type: none"> - possible answers are listed for all questions - fewer responses and marks in Section A - more responses and marks in Section B - too many marks allocated to one concept at skill 1(knowledge and understanding) <p><i>6(b) Fig 6.2 shows the human eye. Identify and describe the function of the parts labelled A, B, C and D. [12]</i></p> <p>Mark scheme: the four parts were identified and their functions described.</p> <p>The marking points were not evenly distributed among the parts.</p> <p>Part A: 3 marking points</p> <p>Part B: 4 marking points</p> <p>Part C: 4 marking points</p>

Feature	Notes
	<p>Part D: 5 marking points</p> <ul style="list-style-type: none"> - More answers were added at pre-standardisation and standardisation meetings (Interview responses; personal experience). - When is the marking scheme closed? (interviews) - The number of marking points was equal to the marks for the majority of questions - The number of marking points was more than the marks for a few questions - The marking points were worth one mark. - Some marks were awarded for similar points <p><i>8(a) (ii) explain the term drug abuse. [6]</i></p> <p>Mark scheme:</p> <p><i>Wrong use of drug;</i></p> <p><i>For leisure/illegal use/because of peer pressure;</i></p> <p><i>Used after their expiry date;</i></p> <p><i>Prescription is not followed appropriately;</i></p> <p><i>Sharing prescribed drugs;</i></p> <p><i>Not completing the course;</i></p> <p>Wrong use could be alternative wording for illegal use, used after expiry date, sharing prescribed drugs and not completing course. The examiner might have run out of answers worth six marks.</p> <ul style="list-style-type: none"> - Candidates were allowed to use alternative wording for some answers, <p>e.g. Candidates could use their own words for ‘quick response; impulse is generated’.</p>
Relevance of main points to research	<ul style="list-style-type: none"> - Addition of some more answers at standardisation meetings could improve the quality of the mark schemes and the accuracy of marking.

Feature	Notes
<p>questions:</p> <p>Examiner training and standardisation</p>	<ul style="list-style-type: none"> - The standardisation meeting could provide guidance to the examiners on how to award marks to candidates' responses - The pre-standardisation time for the mark scheme was inadequate (PMS' Report, January 2016). - The pre-live training enhances mastery of the marking scheme
<p>Monitoring quality of marking</p>	<ul style="list-style-type: none"> - Hierarchical seeding approach for constrained papers (Pinot de Moira, 2013; DRS, 2013; Hudson, 2009, Roan, 2009; personal experience as subject manager). - True determined by the senior marker (Pinot de Moira 2013:16) - Errant markers can be identified by the system and stopped from marking; errant markers need training during marking(Pinot de Moira 2013; Hudson 2009; DRS 2013) - Higher chances of candidates being awarded marks within 10% of the true mark (Pinot de Moira 2013:12) - Senior markers should train errant markers (Pinot de Moira, 2013; DRS 2013; Hudson 2009; Roan 2009) - Small tolerance ranges for quality control, enabling greater chances of identifying errant markers (Pinot de Moira 2016:17) - Any examiner comments about seeding approach to quality control (Raikes 2004; 19; Coniam 2011a:1045)
<p>Influence of questions and mark schemes on quality of marking</p>	<ul style="list-style-type: none"> - Level 3, Semi-constrained mark scheme, where examiners have to judge the adequacy of evidence provided by the candidates (Ahmed & Pollit, 2011:267). - The longest question was basically testing Skill 1: 6(b) – 12 marks In Section B - Some Section B questions were not free response (but were highly structured) as prescribed by the syllabus: <i>6(b) Fig 2 shows the human eye. Identify and describe the functions of the parts labelled A, B, C and D. [12]</i>

Feature	Notes
	<p>The mark scheme listed specific short responses.</p> <ul style="list-style-type: none"> - The uneven distribution of marks could mislead the candidates, prejudicing them of marks they deserve. - This could be an indication that the Section B of this paper has not been properly set even before OSM. - Some questions elicited free responses as prescribed by the syllabus. <i>6(a) Describe the route taken by a nerve impulse when a person touches a hot object. [8]</i> <p>Mark scheme lists 11 marking points that the candidates can join in continuous prose to come up with a long response.</p> <ul style="list-style-type: none"> - The questions elicit short responses that are marked accurately with Level 3 semi-constrained mark scheme (Ahmed & Pollit, 2011:267) - Higher level of marker agreement during marking (Bramley, 2008:2). - The paper was marked on screen in November 2013 (Examination Circular Number 41 of 2013) without any modifications to its structure. This reduces the risk of designing tests to suit the demands of technology. - The Service provider had indicated that all marking in the OSM environment is done at question level because the quality control processes were linked to this approach (email, 8 May 2013)
Opportunities of quality control in the OSM environment	<ul style="list-style-type: none"> - Accurate marking of short questions - Shorter marking period when questions are marked by three types of examiners. - Identification of errant markers by the quality control system - Opportunity for training of errant markers during marking
Challenges of quality control in	<ul style="list-style-type: none"> - Chances of examiners missing candidates' responses written on inappropriate spaces. This could increase the chances of remark

Feature	Notes
the OSM environment	<p>requests (Ofqual, 2011: AQA challenge).</p> <ul style="list-style-type: none"> - Bunching of marks for highly structures questions could reduce the quality of the mark schemes - Setting some questions that elicit short responses for Section B, compromising the assessment scheme. - Inadequate time for the pre-standardisation meeting could compromise mastery and application of the mark scheme.
Conclusion	<p>The paper can be marked with high accuracy due to the nature of the questions and mark schemes. The seeding approach to quality control offers an opportunity to identify and train errant markers, improving the quality of marking. The paper structure was modified for OSM, increasing the risk of designing examinations to suit the demands of technology. There is a risk of inaccurate marking that may arise when examiners are not able to see and mark responses written on inappropriate spaces and where the mark allocation is not clear.</p>

Appendix K: Findings from documents

Research Question 1: Standardisation and Training

Document	Key findings
O Level Biology (5008) syllabus	<ul style="list-style-type: none"> - examiners should be trained to mark according to the syllabus specifications (Lit) - Standardisation meetings enhance the mastery of the content and assessment objectives prescribed in the syllabus. (Lit)
Quality of marking Overall: 5008/2 N2015	<ul style="list-style-type: none"> - All examiners (Normal and senior markers) failed some seeds implying that all of them did not consistently apply the mark scheme - The PMS' report for the November 2015 5008/2 indicated that the pre-standardisation meeting was allocated inadequate time. - This could be an indication that the pre-standardisation meetings were rushed. - These examiners were probably allowed to go into live marking passing the live sample scripts.
Quality of marking Overall: 5008/4 N2015	<ul style="list-style-type: none"> - All examiners (Normal and senior markers) failed some seeds implying that all of them did not consistently apply the mark scheme <p>Conceptual framework and Literature:</p> <ul style="list-style-type: none"> - This could be an indication that the pre-standardisation and standardisation meetings were not effective - These examiners were probably allowed to go into live marking without passing the live scripts. - Follow up with subject managers on how the practice live scripts are assessed in the OSM environment.
D-Grade Form: A - H	<ul style="list-style-type: none"> - First time the paper was marked on screen - The examiner A (Normal marker) had not fully grasped the marking scheme during standardisation, hence the 'fair' understanding - The examiner B (Normal Marker) had not fully grasped the marking

Document	Key findings
	<p>scheme during standardisation, hence the ‘fair’ understanding</p> <ul style="list-style-type: none"> - Poor interpretation of the mark scheme could be a result of inadequate training and standardisation. - Examiner C applied the mark scheme quite well. - The examiner had fully grasped the marking scheme during standardisation, hence the ‘quite well’ application of the mark scheme - Examiner D (Senior Marker) applied the mark scheme quite well. - The TL commented that Examiner E (Senior Marker) had a fairly high percentage of failed seeds 8.66% compared to an average of 5.07. - The examiner F (Senior Marker) exhibited good mastery of the marking scheme - The senior marker had fully grasped the marking scheme during standardisation. - Examiner G (Normal marker)’s understanding of the marking scheme was fair with a seed failure rate of 3.3% compared to an average of 5.07. - Examiner H (once a senior marker but demoted to normal marker) was slow and struggled to interpret and apply the mark scheme - The examiner had not fully grasped the marking scheme during standardisation, hence the ‘fair’ understanding
Exams Admin work schedule: November 2014 exam	Training and standardisation, loading of marking guidance and confirmation of marking guidance activities were planned within the same period (01-07/12/14). This may create pressure to rush through the training and standardisation processes to meet the timelines.
OSM work schedule: November 2014 exam	<ul style="list-style-type: none"> - Senior markers for Biology 5008/4 were given one day to hold their pre-standardisation meeting (04/01/15) - The standardisation meeting was in a day as well (05/01/15) - The mark scheme was loaded on the same day (05/01/15)

Document	Key findings
	<ul style="list-style-type: none"> - Seeds were set by senior markers on one day (06/01/15) - Normal markers trained to mark on screen on the same day – Pre-live marking (06/01/15) <p>Normal markers were not monitored by senior markers during training. The senior markers were setting seeds.</p>
Lessons learnt Report: June and November 2014 exam	<ul style="list-style-type: none"> - Training materials were loaded onto the server for the two exam sessions - Subject managers were trained to train examiners - Subject managers trained the examiners using the loaded material - There was no training on mark schemes in 2014 sessions. - The software provider recommended that ZIMSEC staff be trained to clip and load mark schemes onto the software.
Lessons learnt Report: June 2013 exam	<ul style="list-style-type: none"> - Shortage of computers prevented computer based training
OSM project handover report	<ul style="list-style-type: none"> - Scanning was delayed because the working space was not immediately available - The delay in the scanning of 5008/4 scripts could impact on the standardisation and training period.
PMS' Report: 5008/2 November 2015	<ul style="list-style-type: none"> - Senior markers needed more time for standardisation - The pre-standardisation meeting was rushed leading to poor understanding and application of the mark scheme by both normal and senior markers (PMS' report; D-grade reports). - This could be the reason why some senior markers failed the seeds (8.66% compared to an average of 5.07%) (D-grade reports).
OSM training: Dec	<ul style="list-style-type: none"> - Administrators and subject managers were trained to manage OSM

Document	Key findings
2012	<p>and to train senior markers.</p> <ul style="list-style-type: none"> - Normal markers were trained on computer appreciation by an IT specialist, and on marking by the senior markers and the subject managers, under the supervision of the software provider - Another group of administrators were trained to manage pre and post marking activities - A group of participants was invited for general observation and appreciation of the system - The trainees were advised to appreciate that the training was being conducted during a live marking session with set timelines, so the delivery of the December 2012 marking should take precedence in terms of thrust. Should there be need for an additional training session to accommodate any training gaps, the issue will be escalated to the relevant authorities. - Senior markers and normal trainers were trained before they could mark on screen. - Subject managers were trained as administrators in the OSM environment.
Marking Monitoring Reporting phase walk through	<ul style="list-style-type: none"> - Administrators were trained to monitor marking in the OSM environment. - They were trained to monitor quality of marking through routine maintenance of data and checking reports.
Parameter calculator user guide	<ul style="list-style-type: none"> - A guidance to support users who have the core data for the marking of a related part to determine sensible and appropriate seeding or double percentage marking parameters. - The OSM administrators would need to understand the significance and impact of each parameter on quality control. - Interview subject managers on the training they received and use of

Document	Key findings
	this guide.
Quality Control: Percentage double marking	<ul style="list-style-type: none"> - Double marking parameters were defined and a step by step guide to set the parameter was described in a document - Administrators were trained to set percentage double marking parameters and allocation of marking load to examiners - There was need to train examiners on how the percentage double marking approach to quality control works.
Quality Control: Seeding	<ul style="list-style-type: none"> - Seed parameters were defined and a step by step guide to set each of the parameter was described. - Administrators were trained to set seed parameters and allocation of marking load to examiners - There was no evidence of examiner training on how the seed approach to quality control works. - Biology 5008/2 senior markers were given inadequate time for the pre-standardisation meeting for the November 2015 examination (PMS' report, January 2016), compromising the quality of seeds set. - The marking period for the November 2014 examinations was reduced from ten days, reducing the training and standardisation period as well (E-marking program, November 2014). - Senior markers were setting seeds for 5008/4 while the normal markers were doing pre-live training (E-marking report). The normal markers, therefore, were not monitored by senior markers during training. This could have compromised the quality of marking.
Senior marker quick reference guide	<ul style="list-style-type: none"> - The guide trained senior markers to resolve problems escalated to the by all markers and to set seeds - The senior markers were provided with the guide to constantly

Document	Key findings
	remind them of the training content
Question papers	
5008/2 N 2015; J2016; N2016	<ul style="list-style-type: none"> - The paper might need more time for training and standardisation than paper 4 because it is longer. - The PMS report indicated that the paper needed more time for coordination for the 2015 exam. - The pre-standardisation meeting was rushed leading to poor understanding and application of the mark scheme by both normal and senior markers (PMS Report).
5008/3 N2013; N2014	<ul style="list-style-type: none"> - The paper is longer than 5008/4 but shorter than 5008/2. - Mark scheme might take short standardisation period. - Pre-live training might enhance mastery of the mark scheme, and hence quality.
5008/4 N2013 – N2016	<ul style="list-style-type: none"> - The paper is short (1hr long; 40marks). Mark scheme might take short standardisation period.
Mark schemes	-
5008/2 N2015; J206; N2016	<ul style="list-style-type: none"> - Addition of some more answers at standardisation meetings could improve the quality of the mark schemes and the accuracy of marking. - The standardisation meeting could provide guidance to the examiners on how to award marks to candidates' responses - The pre-standardisation time for the N2015 mark scheme was inadequate (PMS' report, January 2016). - The pre-live training enhances mastery of the marking scheme Addition of some more answers at standardisation meetings could improve the quality of the mark schemes and the accuracy of marking.

Document	Key findings
5008/3 N2013; N2014	<ul style="list-style-type: none"> - The standardisation period might have been long as examiners tried to provide specific answers to the questions. - The standardisation meeting could provide guidance to the examiners on how to award marks to candidates' responses - The pre-live training enhances mastery of the marking scheme
5008/4 N2013; N2014; N2016	<ul style="list-style-type: none"> - The standardisation meeting provide guidance to the examiners on how to award marks to candidates' responses - The pre-live training enhances mastery of the marking scheme
Supervisor's report 5008/3 N2013	<ul style="list-style-type: none"> - The standardisation meeting provide guidance to the examiners on how to use the report during marking - The pre-live training enhances mastery and use of the reports.
IDS: 5008/2/4 J2016	<ul style="list-style-type: none"> - Examiners needed to be trained to marks using the mark scheme as guided by the IDS - Examiners go through pre-live training where they practise with the direct and tick marking and the items marked together (interviews; OSM program)
Minutes of meetings	
Commissioning of marking meetings: July 2015; Dec 2016; June 2017	<ul style="list-style-type: none"> - Marking was due to start on the 6th of July 2015 for all subjects but was postponed to 13 July 2015 for small entry subjects with fewer scripts because mark sheets had not been delivered. The mark sheets were to be delivered on the 17th of July 2015 - The department of responsible for inviting examiners indicated that they were going to revoke the invitation letters they had sent to examiners since the starting date had been postponed. - This is evidence of problems associated with PBM or improper planning on the part of ZIMSEC. - .Marking would start period was 5-22 December 2016 for PBM and 3-22 December 2016 for OSM.

Document	Key findings
	<ul style="list-style-type: none"> - The marking exercise was time-framed for both PBM and OSM in Dec 2016. This could impact on the training and standardisation activities, given that the scanning was delayed due to fewer scanners available. - Marking period was 3-18 July 2017 - The marking exercise was time-framed for both PBM and OSM for the June 2017 marking exercise. This could impact on the training and standardisation activities.
ILAP specification document	<ul style="list-style-type: none"> - ILAPs increase workload for senior markers, who could rush over the standardisation sessions - More work for examiners could force the ZIMSEC to reduce the standardisation and training period to accommodate activities such as ILAP checking
Memo from marking administrators	<ul style="list-style-type: none"> - The cutting off of holes punched on scripts could delay scanning, hence reducing the training and standardisation periods as indicated in the November 2014 marking programme.
Memo to A/Dir TDR&E	<ul style="list-style-type: none"> - The cost of hiring computers and venues could force the ZIMSEC to reduce training and standardisation periods

Appendix L: Parameter calculator user guide

Feature	Notes
Author/Creator	Software provider
Context (place and time of document creation)	Place and date not indicated
Intended audience	OSM administrators
Purpose for document creation	Setting of quality control parameters
Type of document (pamphlet, newspaper, memo, etc)	User guide
Main points expressed in the document	<ul style="list-style-type: none">- To support users who have the core data for the marking of a related part to determine sensible and appropriate seeding or double percentage marking parameters.- The seeding model has default values that are incorporated in the mark scheme, which can be amended- For percentage double marking, the marking approach is assessed to determine the appropriate levels for the quota and the four key parameters for double marking (pioneer cap; partnering cap; penalty and suspect cap).

Feature	Notes
	<ul style="list-style-type: none"> - Seeding parameters are set at question level, not at component level because some factors may vary depending on the complexity of the guidance (mark scheme) and the marking tolerance. <p style="margin-left: 20px;">Seeding Input</p> <ul style="list-style-type: none"> - Number of parts to be marked; number of (normal) markers - Length of (core) marking period - Seed settings: seed percentage – e. g 5%; seed window size – e. g 10; seed maximum failures – e. g 3 - A higher seed percentage would probably increase the quota size (the number of questions marked by each examiner) - Qualification settings: qualification – e. g 10; maximum qualification failures – e. g 3; and qualification limit – e. g 3 (optional) - Lowering the qualification size will reduce both the seedbank size and the quota. - Lowering the seed maximum failure will increase the risk of failing seeds and increase the quota. - The settings can be reviewed to determine the impact of changing the values. - Lowering the maximum qualification failures will increase the risk of failing qualification and so may increase the number of seeds and quota size - There are input and output for percentage double marking - There are no default values in the double marking fields - Just like in the seeding approach a percentage is set for double marking, with a maximum of 100%, where all questions are marked by a second examiner. - Expected stop rate is used to set the caps in double marking - There is another quality control mechanism called the S-Process that was not explained in detail.

Feature	Notes
<p>Relevance of main points to research questions:</p> <p>Examiner training and standardisation</p>	<ul style="list-style-type: none"> - The OSM administrators would need to understand the significance and impact of each parameter on quality control. - Interview subject managers on the training they received and use of this guide.
Monitoring quality of marking	<ul style="list-style-type: none"> - The effectiveness of the quality control system depends on the parameters set by the administrators (subject managers) - The parameters can be set to increase or reduce the amount of quality control delivered. - Seed percentage and qualification are the key parameters in the seeding approach to quality control. - The percentage set for double marking and the penalties are key parameters in the percentage double marking approach - It is possible to have all questions marked by a second marker in double marking (check literature for Hong Kong)
Influence of questions and mark schemes on quality of marking	<ul style="list-style-type: none"> - Seeds are used to control quality of marking for constrained examinations and double percentage marking is used for unconstrained examinations (Roan, 2009; Hudson, 2009)
Opportunities of quality control in the OSM environment	<ul style="list-style-type: none"> - Setting of parameters that deliver high quality control - Review of the parameters when they are not working well - Identification of errant markers - Training of errant markers
Challenges of quality control in the OSM	<ul style="list-style-type: none"> - Manipulation of quality control parameters to allow minimum quality control so as to meet marking deadlines.

Feature	Notes
environment	
Conclusion	<p>Quality control parameters can be set and reviewed by OSM administrators. The parameters can be set to increase or reduce the amount of quality control delivered by the system. Seed percentage and qualification are the key parameters in the seeding approach. The percentage and the penalties are the key parameters in double percentage marking.</p>

Appendix M: Editing certificate

EDITING CERTIFICATE

Midlands State University

Department of English and Communication

Zvishavane Campus

Private Bag 9055

Gweru

To whom it may concern

Please be advised that I, Tasiyana D. Javangwe, completed professional academic editing on the following thesis, submitted in accordance with the requirements of the degree of Doctor of Philosophy in Education, for Ebba Masiri:

*Exploring the practice of quality control in the onscreen marking of
Ordinary Level Biology in Zimbabwe*

This included editing of spelling, grammar, register and other language-related items as well as assistance with referencing and layout.

Disclaimer

Please note that all changes were tracked and either accepted or rejected at the client's discretion. I thus take no responsibility for the final document as it may differ from that which I supplied to the client.

Yours faithfully



Prof Tasiyana D. Javangwe

BA, BA Hons (English), MA (English)- UZ, DLitt et Phil (Unisa)

Cell: +263 773 634 138

Email: javangwet@staff.msu.ac.zw or javangwet03@gmail.com

Appendix N: Ethical clearance

UNISA COLLEGE OF EDUCATION ETHICS REVIEW COMMITTEE

Date: 2019/05/15

Ref: **2019/05/15/58526110/17/MC**

Name: Mrs E Masiri

Student: 58526110

Dear Mrs Masiri

Decision: Ethics Approval from
2019/05/15 to 2024/05/15

Researcher(s): Name: Mrs E Masiri
E-mail address: 58526110@mylife.unisa.ac.za
Telephone: +263 77 600 2853

Supervisor(s): Name: Prof MT Gumbo
E-mail address: Gumbomt@unisa.ac.za
Telephone: +27 12 429 3339

Title of research:

**Exploring the Practice of Quality Control in the Onscreen Marking of Ordinary Level
Biology in Zimbabwe**

Qualification: D. Ed in Comparative Education

Thank you for the application for research ethics clearance by the UNISA College of Education Ethics Review Committee for the above mentioned research. Ethics approval is granted for the period 2019/05/15 to 2024/05/15.

*The **low risk** application was reviewed by the Ethics Review Committee on 2019/05/15 in compliance with the UNISA Policy on Research Ethics and the Standard Operating Procedure on Research Ethics Risk Assessment.*

The proposed research may now commence with the provisions that:

1. The researcher(s) will ensure that the research project adheres to the values and principles expressed in the UNISA Policy on Research Ethics.



2. Any adverse circumstance arising in the undertaking of the research project that is relevant to the ethicality of the study should be communicated in writing to the UNISA College of Education Ethics Review Committee.
3. The researcher(s) will conduct the study according to the methods and procedures set out in the approved application.
4. Any changes that can affect the study-related risks for the research participants, particularly in terms of assurances made with regards to the protection of participants' privacy and the confidentiality of the data, should be reported to the Committee in writing.
5. The researcher will ensure that the research project adheres to any applicable national legislation, professional codes of conduct, institutional guidelines and scientific standards relevant to the specific field of study. Adherence to the following South African legislation is important, if applicable: Protection of Personal Information Act, no 4 of 2013; Children's act no 38 of 2005 and the National Health Act, no 61 of 2003.
6. Only de-identified research data may be used for secondary research purposes in future on condition that the research objectives are similar to those of the original research. Secondary use of identifiable human research data requires additional ethics clearance.
7. No field work activities may continue after the expiry date **2024/05/17**. Submission of a completed research ethics progress report will constitute an application for renewal of Ethics Research Committee approval.

Note:

*The reference number **2019/05/15/58526110/17/MC** should be clearly indicated on all forms of communication with the intended research participants, as well as with the Committee.*

Kind regards,



Prof AT Motlhabane
CHAIRPERSON: CEDU RERC
motlhat@unisa.ac.za



Prof PM Sebate
ACTING EXECUTIVE DEAN
Sebatpm@unisa.ac.za



Approved - decision template – updated 16 Feb 2017

University of South Africa
Preller Street, Muckleneuk Ridge, City of Tshwane
PO Box 392 UNISA 0003 South Africa
Telephone: +27 12 429 3111 Facsimile: +27 12 429 4150
www.unisa.ac.za