THE DEVELOPMENT AND EVALUATION OF AFRICANISED ITEMS FOR MULTICULTURAL COGNITIVE ASSESSMENT

by

NOMVUYO NOMFUSI BEKWA

submitted in accordance with the requirements
for the degree of

DOCTOR OF COMMERCE

in the subject

INDUSTRIAL AND ORGANISATIONAL PSYCHOLOGY

at the

University of South Africa

Supervisor: PROF M DE BEER

January 2016

DECLARATION

I, NOMVUYO NOMFUSI BEKWA, declare that **"The development and evaluation of Africanised items for multicultural cognitive assessment"** is my own work, which is submitted in accordance with the requirements for the degree of Doctor of Commerce in the subject, Industrial and Organisational Psychology. All sources that were used were acknowledged as references.

I further declare that ethical clearance to conduct the research has been obtained from the Department of Industrial and Organisational Psychology, University of South Africa. I also declare that the study was carried out in strict accordance with the Unisa Policy on Research Ethics and that I conducted the research with the highest integrity during all phases of the research process, taking into account Unisa's Policy on Copyright Infringement and Plagiarism

---------- ----------

Ms N N Bekwa **Date**

3232 530 4

ACKNOWLEDGEMENTS

If I have seen further, it is only by standing on the shoulders of giants
Isaac Newton, 1676

I know I am standing on the broad shoulders of a great many and I am grateful for whatever role each one played (mentioned or not) because all their contributions spurred me on towards the completion of this project.

- My parents, Charles and Lauraine Bekwa, lovingly planted their seed and diligently did all the spadework necessary to cultivate the soil – I had no choice but to germinate. I humbly dedicate this thesis to you both.
- My promoter, Prof Marié de Beer, who was true to the saying, "*when the student is ready, a teacher will appear*". I appreciate you for trusting me with your idea – the journey has been a gift. I am eternally grateful for your unwavering support, guidance, input, encouragement, kindness and friendship. As the curtain goes down on this round of the project, I proudly take a bow to you.
- My department, Industrial and Organisational Psychology, Unisa, for providing the comfort of knowing I had experts at my disposal. At different times and in different ways, my colleagues provided the necessary advice, sharing of knowledge and experiences, and evidence of what is possible.
- Unisa-SANPAD for providing the necessary compass with the training opportunity of the Doctoral Research Capacity Initiative Programme.
- All the institutions involved in granting permission for the collection of data for this study. I am thankful to you for opening your doors and changing your schedules to accommodate this research. I am grateful to all the personnel of these institutions who embraced the whole process and went the extra mile to ensure that the environment was conducive to the research. More importantly, I am indebted to the participants who gave me the gift of their time by completing the questionnaires, giving feedback; and sharing their views.
- Andries Masenge, for providing statistical support whenever I was confused. I appreciate the fact that you thought I knew more about data analysis than I really do. It pushed me to learn more.

- The editorial support from Moya Joubert and Bahia Singh is greatly appreciated. Moya, for ensuring that my words (content-wise) read well, and Bahia, for ensuring my words (technically) look good.
- My friends who have been my pillars throughout: Phumeza, Nokwanda, Sammy, Munka, Gugu, Martin, and Ophillia. I do believe in telepathy because of you – at every turn, you seemed to offer what I needed exactly when I needed it.
- My dearest siblings, Tsili, Thembisa, Vuyani and Sitsaba and their families; Dabaw' uLulama and Dabaw' uThozama, to me you symbolise the presence of my father which I appreciate immensely, the family tree of the Bekwa clan in its entirety – both living and beyond. I may not have chosen you, but I certainly am happy I was born into this family. Having all of you in my life is a blessing. To each one of you I say, nangamso, Gamede, Wushe, Mjoli, Phathwa, Manyaweni, Mphubane, Nkungwini – abantu abangafani nabany' abantu.

Dear Lord, I know I am because You are; I am grateful for all the above blessings. The journey was long – with life-changing bumps which left irreparable dents; but taken in the context of what has been achieved I hope I will see them as beautiful dimples of growth. Thank you for keeping me steady and on track. Where and when I could not, You did!!!

SUMMARY

Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.

Marie Curie

Debates about how best to test people from different contexts and backgrounds continue to hold the spotlight of testing and assessment. In an effort to contribute to the debates, the purpose of the study was to develop and evaluate the viability and utility of nonverbal figural reasoning ability items that were developed based on inspirations from African cultural artefacts such as African material prints, art, decorations, beadwork, paintings, et cetera. The research was conducted in two phases, with phase 1 focused on the development of the *new items,* while phase 2 was used to evaluate the *new items*. The aims of the study were to develop items inspired by African art and cultural artefacts in order to measure general nonverbal figural reasoning ability; to evaluate the viability of the items in terms of their appropriateness in representing the African art and cultural artefacts, specifically to determine the face and content validity of the items from a cultural perspective; and to evaluate the utility of the items in terms of their psychometric properties.

These elements were investigated using the exploratory sequential mixed method research design with quantitative embedded in phase 2. For sampling purposes, the sequential mixed method sampling design and non-probability sampling strategies were used, specifically the purposive and convenience sampling methods. The data collection methods that were used included interviews with a cultural expert and colour-blind person, open-ended questionnaires completed by school learners and test administration to a group of 946 participants undergoing a sponsored basic career-related training and guidance programme. Content analysis was used for the qualitative data while statistical analysis mainly based on the Rasch model was utilised for quantitative data.

The results of phase 1 were positive and provided support for further development of the *new items,* and based on this feedback, 200 *new items* were developed. This final pool of items was then used for phase 2 – the evaluation of the *new items*. The

statistical analysis of the *new items* indicated acceptable psychometric properties of the general reasoning ("g" or fluid ability) construct. The item difficulty values (*p*-values) for the *new items* were determined using classical test theory (CTT) analysis and ranged from 0.06 (most difficult item) to 0.91 (easiest item). Rasch analysis showed that the *new items* were unidimensional and that they were adequately targeted to the level of ability of the participants, although there were elements that would need to be improved. The reliability of the *new items* was determined using the Cronbach alpha reliability coefficient (α) and the person separation index (PSI), and both methods indicated similar indices of internal consistency (α = 0.97; PSI = 0.96). Gender-related differential item functioning (DIF) was investigated, and the majority of the *new items* did not indicate any significant differences between the gender groups. Construct validity was determined from the relationship between the *new items* and the Learning Potential Computerised Adaptive Test (LPCAT), which uses traditional item formats to measure fluid ability. The correlation results for the total score of the *new items* and the pre- and post-tests were 0.616 and 0.712 respectively. The *new items* were thus confirmed to be measuring fluid ability using nonverbal figural reasoning ability items. Overall, the results were satisfactory in indicating the viability and utility of the *new items*.

The main limitation of the research was that because the sample was not representative of the South African population, there were limited for generalisation. This led to a further limitation, namely that it was not possible to conduct important analysis on DIF for various other subgroups. Further research has been recommended to build on this initiative.

Key terms: multicultural cognitive assessment; fluid ability; item development; African art and cultural artefacts; culture-fair assessment; item analysis; Rasch analysis; reliability; validity; item difficulty; differential item functioning.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Our sense of elevation at this moment also derives from the fact that this magnificent product is the unique creation of African hands and African minds
Today it feels good to be an African
I am an African
Thabo Mbeki

(Excerpt from “I am an African” speech by President Thabo Mbeki)

CHAPTER 1

SCIENTIFIC ORIENTATION TO THE RESEARCH

Begin with the ending in mind. - Stephen Covey

1.1 INTRODUCTION

The field of psychological testing and assessment is a fascinating one which generally arouses curiosity and intrigue about such measurement – relating responses to a set of items or questions and interpreting the results to form a meaningful profile about individuals' personality, general ability, aptitudes, interests and values. The major decisions about selection, placement, promotion, training, career paths or choice of field of study, which are usually made on the basis of the results of testing and assessment, can have a life-changing impact on individuals and can also influence groups, organisations and the nation at large. Because of this impact, the way in which these items and instruments are developed and how their quality, accuracy and applicability across cultures are evaluated, continues to be of critical importance, especially in a multicultural and multilingual society such as that of South Africa – hence the topic of the current study, namely the development and evaluation of new Africanised items for multicultural cognitive assessment.

In chapter 1, the scientific orientation to this study is outlined, starting with the background to and motivation for the research as well as the problem statement. This is followed by a discussion of the aims of the research, the paradigm within which the research was contextualised and the research design. The layout of the chapters is also presented.

1.2 BACKGROUND TO AND RATIONALE FOR THE RESEARCH

Assessment of cognitive functioning has a history that goes back centuries, as do the debates about how best to test and assess people from different contexts and backgrounds (Anastasi & Urbina, 1997; Gregory, 2007; Kgosana, 2012). Some

issues in the debate have included the effects of language and socioeconomic background on test performance, culture fairness of the tests and the interpretation and use of the results (Anastasi & Urbina, 1997; Kgosana, 2012; Papalia & Olds, 1988). Similar debates have been evident in South Africa, and these were further fuelled by the history of racial discrimination which saw test results being used to exclude and disadvantage the majority of the population from educational and economic opportunities (Claassen, 1997; Kanjee, 2006; Nzimande, 1995).

In light of political and social changes that occurred in South Africa in the 1990s, testing and assessment were one of the areas that received critical scrutiny. There were "anti-psychometric sentiments" (Claassen, 1997, p. 303), with voices, both soft and loud, questioning the need and importance of continuing with the use of such tests (Claassen, 1997; Kanjee, 2006; Nzimande, 1995) creating the impression that testing and assessment were under threat. However, rather than simply discontinuing the use of tests and assessments, in 1998, the government addressed the potential flaws and shortcomings of testing and assessment by promulgating the Employment Equity Act 55 of 1998 (EEA) (Government Gazette, 1998, p. 7), which included the following stipulation:

> *Psychological testing and other similar forms of assessments of an employee are prohibited, unless the test or assessment being used (a) has been scientifically shown to be valid and reliable; (b) can be applied fairly to all employees; and (c) is not biased against any employee or group.*

Initially, with the introduction of this Act, some people misunderstood it to mean that testing and assessment were banned. However, this was not the intention, because the Act placed emphasis on finding ways to better assess South African citizens validly, reliably, fairly and without bias (Government Gazette, 1998). The Act therefore opened debate, encouraging creativity and at the same time requiring more responsibility and accountability on the part of test developers and test users. The positive consequence of the Act was that it promoted improvement, change and

transformation of tests and testing practices to be fair and scientifically defensible in the multicultural and multilingual South African society.

Some of the consequences and challenges that were emphasised by the promulgation of the Act were the need to develop new instruments, to validate existing instruments for all groups in South African society and to ensure improved testing practices (Van de Vijver & Rothmann, 2004). The Act is also credited with increasing the number of studies on cultural differences, test adaptation, test bias, fairness and multicultural assessments (De Beer, 2004; Donald, Thatcher, & Milner, 2014; Foxcroft & Aston, 2006; Malda, Van de Vijver, & Temane, 2010; Paterson & Uys, 2005; Theron, 2007), in particular with regard to cognitive assessment. These studies investigated and highlighted some of the issues that continue to warrant research on the impact on test performance of factors such as South Africa's 11 official languages, different socioeconomic backgrounds, schooling and cultural differences and rural versus urban differences (De Beer, 2004; Donald et al., 2014; Foxcroft & Aston, 2006; Malda et al., 2010; Paterson & Uys, 2005; Theron, 2007; Van Dulm & Southwood, 2013).

On the basis of the above background, the challenge for assessment research as a field to critically and creatively contribute to the development of new instruments and testing practices is clear –hence the rationale for this study.

1.3 PROBLEM STATEMENT

From a history of bias and unfairness to the present legislatively regulated testing and assessment environment and practices, considerable improvement is evident (Maree, 2010). However, as recommended in various studies, still more needs to be done, such as investigating different methods or item formats to better assess individuals in the multicultural and multilingual South African context (Donald et al., 2014; Foxcroft, 2004; Foxcroft & Aston, 2006; Malda et al., 2010; Paterson & Uys, 2005; Rothmann & Cilliers, 2007). Cultural groups, language, schooling experiences and rural versus urban backgrounds have all been highlighted as crucial factors in

designing test items (De Beer, 2004; Foxcroft, 2004). Maree (2010) also highlighted the importance of considering the extent (or rather the lack) of Afrocentrism in the psychological tests used in South Africa. This study was undertaken in order to address some of these issues.

In this study, the development of general nonverbal figural reasoning items and item formats based on inspirations from African art and cultural artefacts (indigenous artefacts) was explored, firstly, to determine the viability of such items to improve the culture fairness of cognitive assessment by using stimulus material that is more familiar to the majority of local individuals than the content traditionally used in such tests. Such art and artefacts are celebrated around the world (Azeez, 2011; Hagg, 2010; Nettleton, 2010), thus giving rise to the interest in using them as inspiration for the new items (see the examples in figure 1.1). Secondly, in the study, the utility of the items was evaluated for measuring general nonverbal figural reasoning ability in terms of their psychometric properties. It is important to note that the focus of the study was intentionally limited to the specific phase of item development and not as a full-scale project to develop a new instrument. Test development requires specialised assessment and research expertise and much larger samples to meet the requirements for the validation of a measure. Such a full-scale process was beyond the scope of this doctoral project, which could be viewed as doing the groundwork for future development of a new measure that could make use of similar item content as that developed for and evaluated in the present study. For the purposes of this research, the newly developed items will be referred to as the *new items* throughout the thesis.

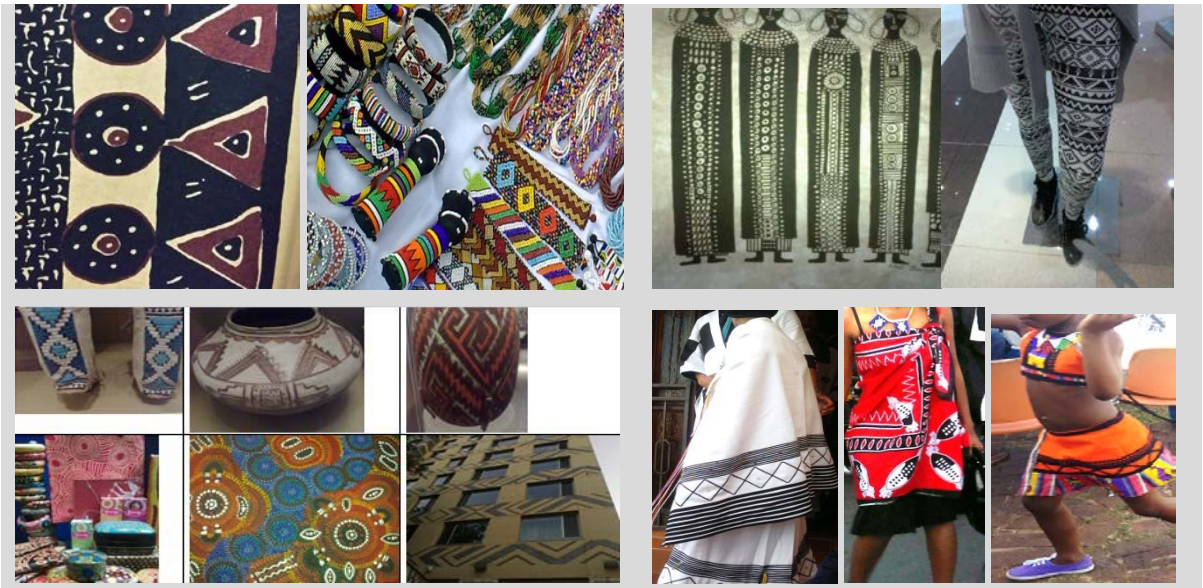


Figure 1.1. Examples of African art and cultural artefacts

The research question was formulated as follows: What is the viability and utility (in terms of psychometric properties) of the *new items* that were developed, using African art and cultural artefacts as inspiration, for measuring general nonverbal figural reasoning ability?

1.4 AIMS OF THE RESEARCH

The primary aim of this research was to develop and evaluate the viability of *new items* inspired by African art and cultural artefacts and also to evaluate the utility of these *new items* in measuring general nonverbal figural or fluid reasoning ability.

The specific aims formulated for the study addressed both the theoretical and empirical aspects of the research as follows:

- To conceptualise intelligence in order to gain an understanding of the theories, principles and debates on fluid reasoning ability measurement
- To review the criteria for the evaluation of tests
- To develop items inspired by African art and cultural artefacts to measure general nonverbal figural reasoning ability

- To evaluate the viability of the *new items* in terms of their appropriateness in symbolising African art and cultural artefacts, specifically to determine the face and content validity of the *new items* from a cultural perspective
- To evaluate the utility of the *new items* in terms of their psychometric properties
- To compare the results obtained on the *new items* with the results obtained from another measure of general nonverbal figural reasoning using the more traditional format items, specifically to determine construct validity
- To draw conclusions about the *new items* based on the aims of the study
- To acknowledge the limitations of the study and make recommendations for future research

The contribution of the study to the field of industrial and organisational psychology is specifically in psychometrics as the study entails the development and evaluation of *new items*. The novelty of the contribution is in how practical creations of art and culture are combined with a scientific process of developing items for psychological testing. A further contribution is the implementation in practical terms of the concept of Africanisation into cognitive assessment, thus contributing to fairness in psychological assessment as required by the EEA legislation (Government Gazette, 1998).

1.5 CONTEXT OF TESTING AND ASSESSMENT

The diversity of people defines the reality of South Africa, therefore playing a key role in the differences that are found in test results (Grieve & Foxcroft, 2013; Owen, 1992). Issues such as culture, socioeconomic status, language and so on, become significant in settings of multicultural assessment. Van Dulm and Southwood (2013) highlighted additional diversity issues that are due to immigration to South Africa, resulting in further language barriers. It is therefore necessary to emphasise the need for understanding the context in which testing and assessment occur. Fulop and Robert (2015) noted the importance of identifying what the contextual factors are, as this provides clarity for when and where interventions are required.

For this study, the context of testing and assessment encompasses the interconnected levels of macro (environmental), meso (organisational) and micro (individual) levels (Fulop & Robert, 2015; McShane & Von Glinow, 2005; Miller, 1993). Applying these levels to the context of testing and assessment would mean that test development and all the processes for the scientific evaluation of tests represent the micro level, test use by test administrators and organisations represent the meso level while, the legislation, political environment and socioeconomic factors represent the macro level (McShane & Von Glinow, 2005; Miller, 1993). According to McShane and Von Glinow (2005) and Miller (1993), the levels have an iterative impact, where a change at one level creates a ripple effect in the whole system. As illustrated in figure 1.2.

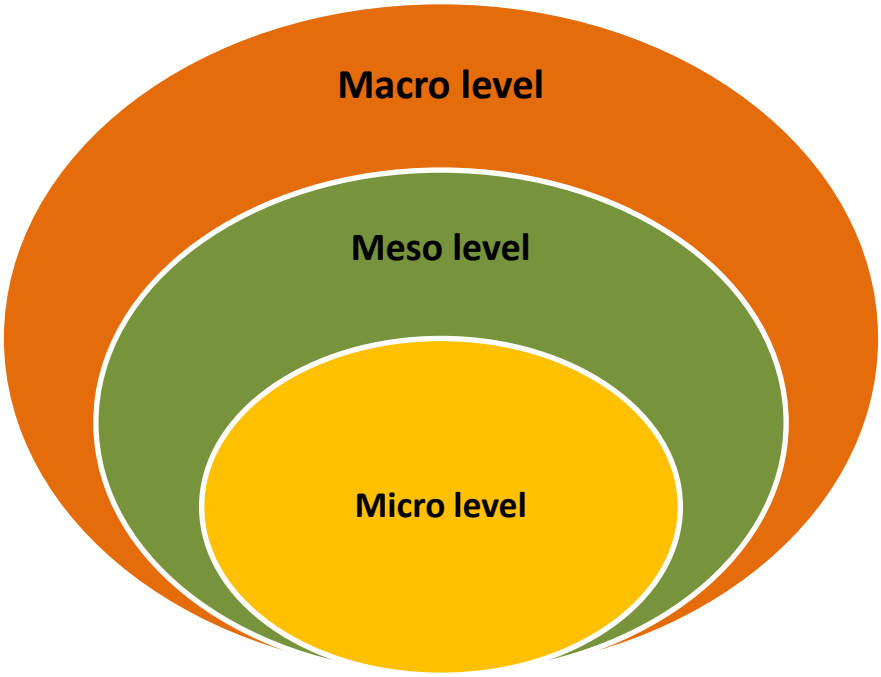


Figure 1.2. Context of testing and assessment

In the development and evaluation of tests, the interconnectedness of the context of testing and assessment has to be specifically considered. For example, when the EEA (Government Gazette, 1998) was introduced at macro level, the test administrators and organisations responsible for testing had to address the requirements laid down to ensure adherence to the legislation professionally (meso

level) and whatever actions and processes taken to address the specific issues of concern would fall within the micro level. In this study, the *new items* (micro level) were constructed using African art and culture artefacts (macro level) as inspiration. The *new items* were evaluated from a cultural perspective (macro level), while the comparison of the newly developed items and traditional standardised items (micro level) was done using participants representative of the different gender and culture groups, and different educational backgrounds (meso level).

1.6 PARADIGM PERSPECTIVE

According to Kuhn (1970), a paradigm is a model for conducting research founded on rules and regulations that clarify the boundaries for the researcher regarding what should be researched and how the research should be conducted. Kuhn (1970) and Mustafa (2011) also explained the use of paradigms as important determinants in terms of what would be regarded as valid research solutions. According to Mustafa (2011), through a paradigm philosophical and conceptual boundaries are set out to guide the research in terms of what the reality is, how to study it and the methods that can be used to study it. Both the psychological paradigm (to clarify what was researched) and the research paradigm (to clarify how the research was conducted) are discussed below. The discussion also covers the disciplinary relationship, the applicable subfields, models and theories, concepts and constructs, methodological convictions and hypotheses (Mislevy, 1996).

1.6.1 Disciplinary relationships

Disciplines in the academic environment refer to the fields of study that cover a certain subject matter or body of knowledge, have their own methods and theories to acquire and organise that knowledge and should have their own norms and values on how content in that field is studied (Buanes & Jentoft, 2009; Moran, 2010). When conducting research it is necessary to indicate the discipline from which arguments and viewpoints originate so that control boundaries are clarified (Moran, 2010). The current research falls within the discipline of industrial and organisational psychology

(IOP) – a field of study that "utilises principles and assumptions of psychology to study and influence human behaviour in the work context" (Bergh, 2009, p. 19). The specific subfield focus is psychometrics because the study involved constructing and evaluating *new items*. Psychometrics entails "the systematic and scientific way in which psychological measures are developed and the technical measurement standards (e.g. validity and reliability) required of measures" (Foxcroft & Roodt, 2013a, p. 4). Other areas of application of psychological assessment are research, personnel psychology, career psychology, counselling psychology and cross-cultural psychology because of the applications of testing and assessment in a multicultural context.

1.6.2 Psychological paradigm: Cognitive psychology

According to Mislevy (1996), the cognitive psychology paradigm caters for research into human abilities. The main assumptions of cognitive psychology entail understanding people as problem solvers, decision makers, continuous processers of information and self-regulators (Das, Naglieri, & Murphy, 1995; Theron, 2009). This study involved assessment research focusing mainly on cognitive assessment – hence cognitive psychology was applicable. This paradigm provides the framework for the literature review.

1.6.3 Research paradigm: Critical realism

Egbo (2005) described critical realism as a framework that is ideal for critical social scientific inquiry. It is a philosophical approach associated with philosophers such as Roy Bhaskar (1978, 1989), Rom Harre (1986) and Margaret Archer (1995), who worked on overcoming social structures of dominance and oppression (Archer, 1995; Bhaskar, 1978, 1989; Carspecken, 1996). According to Mustafa (2011), this paradigm is emancipatory as it balances the power relations, and transformative, because of its advocacy for change. Moerdyk (2015) argued for the use of this paradigm in quantification as he acknowledged that in the real world, measurement is not confined to physical science, but has to include the social sciences. He further noted that the real world "… is shaped and construed by our life experiences, value

systems and the cognitive schemata and categories that we bring to bear on issues … there is some kind of reality out there, although we all interpret it slightly differently as a result of our socialisation experiences" (Moerdyk, 2015, p. 12). Therefore, although the existence of the real world is acknowledged, the interpretation thereof is believed to be based on multiple realities and value-laden views.

Taking into account the context of testing and assessment in which this study was conducted, the three domains of critical realism, namely real, actual and empirical (Bhaskar, 1978; McEvoy & Richards, 2006; Mingers, Mutch, & Willcocks, 2013; Wilson & McCormack, 2006) were applicable. Figure 1.3 below indicates the interconnectedness of the domains.

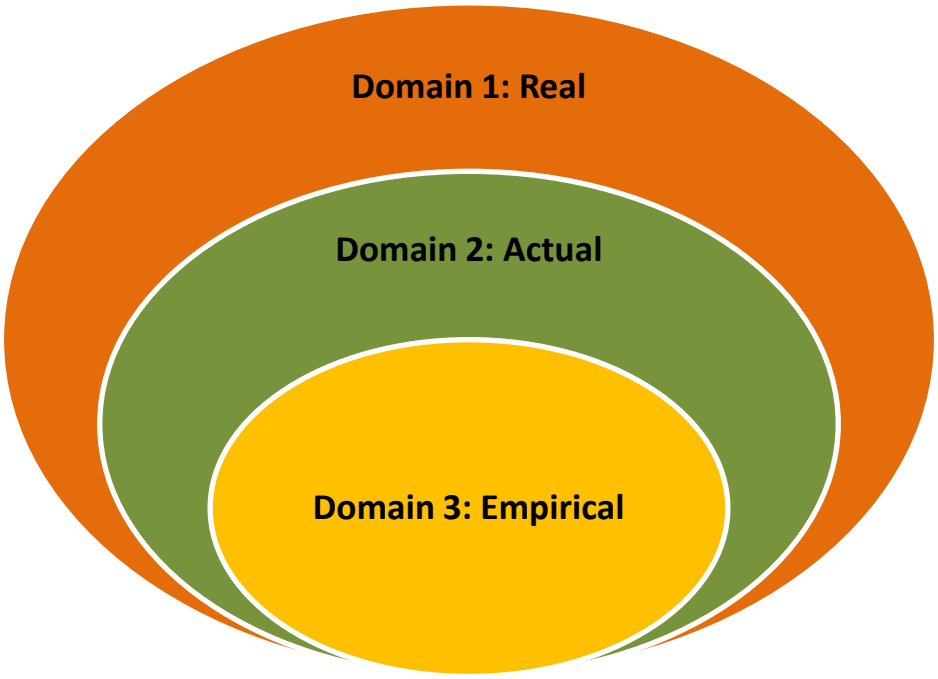


Figure 1.3. Domains of critical realism

The real (aligned to the macro level) comprises mechanisms, events and experiences (e.g. South African legislation – the EEA which represents the external reality); the actual (aligned to the meso level) consists of events that do (or perhaps do not) occur that take into consideration human experiences (e.g. test administrators and organisational contexts); and the empirical (aligned to micro level)

includes those events that are observed as human actions, reflections or theorising (e.g. *new items*) (McEvoy & Richards, 2006; Mingers et al., 2013; Pratt & Gutteridge, 2014). For example, tests are developed to determine behaviours and traits of individuals and groups (empirical domain), by people who use their experiences and knowledge (actual domain) to ensure that the tests are used in adherence to legislative requirements and ethical standards (real domain).

Critical realism is ideal for this study because it offers a platform for various methods that ensure both the depth and breadth in research results (Mustafa, 2011; Venkatesh, Brown, & Bala, 2013). As indicated by Bazeley (2004) and McEvoy and Richards (2006), the choice of methods within the critical realism paradigm is dependent on the research problem, which for this study is the development (qualitative methods) and evaluation (quantitative and qualitative methods) of the *new items*.

1.6.4 Theories and models

Theories are defined as explanations of relationships, predictions of events and understanding of the domain within a field of study (Wacker, 1998), while models are collections of concepts that provide a better understanding of the functioning and insights of the system under study (Baumgärtner, Becker, Frank, Müller, & Quaas, 2008). The focus of the study was mainly on multicultural cognitive assessment, item development and item evaluation (in terms of cultural acceptability or viability and item analysis or utility).

1.6.4.1 Multicultural cognitive assessment

Various theories that contribute to the understanding of the nature and assessment of intelligence are briefly discussed, but the theory that was mainly used is the theory of fluid and crystallised general ability developed by Raymond B. Cattell (Cattell, 1963). This theory is based on a factor analysis approach where underlying relationships between sets of variables are measured (Gregory, 2007). Cattell (1963) proposed two types of cognitive abilities, namely fluid intelligence (*gf*) and

crystallised intelligence (*gc*). Fluid intelligence, on which the nonverbal figural reasoning ability items in this study were based, requires adaptation to new situations where ability is perceived independent of previous learning, but is focused on solving problems logically and creatively as the problems are usually new and independent of acquired knowledge (Cattell, 1963; Gregory, 2007). Crystallised intelligence (*gc*), by contrast, is based on acquired skills and knowledge, where previous learning is important (Cattell, 1963; Gregory, 2007).

1.6.4.2 Item development

Item development was based on the generic steps of developing a scale or psychological measure (Foxcroft, 2013; Moerdyk, 2015), but the process was adapted for item development and evaluation. The adapted process (discussed in detail in chapter 5) included planning (identifying the purpose, sourcing ideas and inspirations); item writing (creating the *new items*, first draft and second draft items); assembling, evaluation (viability in terms of cultural appropriateness) and pretesting (provisional items); and item analysis (evaluation of utility based on item difficulty, reliability and validity of the full item bank) (Foxcroft, 2013; Moerdyk, 2015). The evaluation of the *new items* was focused on determining the viability and utility of the *new items* in terms of their appropriateness as inspirations from African art and cultural artefacts, on the one hand, and the psychometric properties, on the other.

1.6.4.3 Item evaluation

The item evaluation entailed, firstly, evaluation in terms of viability (qualitative evaluation of cultural appropriateness), and secondly, evaluation in terms of utility (quantitative item analysis). The viability and cultural appropriateneness of the items were evaluated by means of a qualitative process involving feedback on the appropriateness of the *new items* from different groups. The utility of the *new items* was evaluated on the basis of quantitative item analysis. Psychometric properties can be evaluated using measurement theories that include classical test theory (CTT), the item response theory (IRT) and the Rasch analysis model (Bond & Fox, 2007; Foxcroft, 2013; Hambleton & Jones, 1993; Zumbo, Gelin, & Hubley, 2002).

CTT entails a “set of procedures that are based on the notion that an observed score is comprised of a true score and an error score” (Zumbo et al., 2002, p. 25). These procedures include, inter alia, the difficulty of items, reliability and item-total correlations (Foxcroft, 2013; Wiberg, 2004; Zumbo et al., 2002). The limitations of CTT were identified as its dependence on the content of the test and on a specific sample, which means comparisons of individuals can only be made on the same tests (Hambleton & Jones, 1993; Wiberg, 2004). By contrast, IRT is not dependent on the content and sample specifics (Foxcroft, 2013; Wiberg, 2004).

The main assumption of IRT is that “the higher an individual’s ability is, the greater the individual’s chances are of getting an item correct” (De Kock, Kanjee, & Foxcroft, 2013, p. 92). The Rasch model, which is related to IRT, is a probabilistic model that estimates item locations independent of the sample, while allowing for inferences to be made about the test, regardless of the distribution of the sample (Bond & Fox, 2007; Furr & Bacharach, 2008; Linacre, 2005). It is expressed in terms of “the probability that an individual with a particular trait level will correctly answer an item that has a particular difficulty” (Furr & Bacharach, 2008, p. 318). According to Lantano (2010), the Rasch model is appropriate for evaluating the psychometric properties of items and tests.

1.6.5 Concepts and constructs

Concepts are defined as the “representation of a chunk of knowledge that can be used to categorise and understand a domain of objects, events, or processes” (Cohen & Murphy, 1984, p. 28). According to Baumgärtner et al. (2008), concepts are accepted standards, processes and principles in a discipline. Constructs are defined as theoretical, intangible or unobservable qualities or traits along which individuals differ (Gregory, 2007; Zumbo et al., 2002). According to Bacharach (1989), clarifying concepts and constructs for the research is essential because they are components of theories. The main concepts and constructs that were used in the study are discussed below:

1.6.5.1 Testing and assessment

It is generally understood that psychological testing and assessment entail an objective and standardised process of measuring or assessing the extent to which people possess a particular characteristic or attribute (Anastasi & Urbina, 1997; Foxcroft & Roodt, 2013a; Moerdyk, 2015).

1.6.5.2 Intelligence and cognitive functioning

Intelligence is defined as the "combined faculties of understanding, reasoning, cognition, apprehension, attention and ideation" (Rapoport, 1999, p. 149). Cognitive functioning (a term used interchangeably with "cognitive abilities") is understood to broadly entail both the measurement of intelligence and aptitude of an individual (Van Eeden & De Beer, 2013). Moerdyk (2015) summarised the characteristics of intelligent people as having the ability to learn and adapt, evaluate and judge, solve problems and comprehend relationships. However, one should note that a common definition of intelligence has not been agreed upon, although most researchers do agree that the construct is relatively stable (Moerdyk, 2015; Van Eeden & De Beer, 2013).

1.6.5.3 Reliability and validity

According to Furr and Bacharach (2008), reliability and validity are fundamental to the quality of items and crucial to the evaluation of psychometric properties. These two concepts are specifically noted as a requirement for any test used in South Africa as per the EEA. Reliability is about the consistency and precision of scores (Furr & Bacharach, 2008; Moerdyk, 2015; Roodt, 2013a), while validity is the extent to which the measure truly reflects the construct measured (Furr & Bacharach, 2008; Roodt, 2013b). Although there are various types of reliability, this study dealt with internal consistency (inter-item) reliability, which is based on the consistency of the responses to all items (Roodt, 2013a). For validity, the study dealt with content description procedures and construct identification procedures, specifically the correlation with another similar measure of the same construct (Roodt, 2013b). One should note that the relationship between these concepts, in which reliability has a

limiting influence on validity, which means a test may be reliable but not valid (Roodt, 2013b).

1.6.5.4 Fairness and bias

Fairness and bias are also specifically mentioned as requirements in the EEA – hence their importance in this study. Fairness is subjective in its approach, based on value judgement, while bias is statistical in approach, based on the systematic error that gives one group an advantage over another (R.B. Kline, 2013; Kurnaz & Kelecioğlu, 2008; Moerdyk, 2015). The considerations of these two concepts make it possible for a test to be viewed as unfair, while it has been shown statistically not to have bias (R.B. Kline, 2013).

1.6.5.5 Viability and utility

Viability is synonymous with feasibility, capability of being done and practicality. According to Peter, Leichner, Mayer, and Krampen (2015), viability entails exploring the appropriateness of the test content and the initial subjective experience in the application of a test. Utility refers to usefulness, effectiveness, substantive meaning, explanation and prediction (Anastasi & Urbina, 1997; Bacharach, 1989). According to Weiner (2013), utility taps into the relevance of the domain of the construct and provides information about the validity and precision of decisions. The main aim of this study was the evaluation of the viability and utility of the *new items*.

1.6.5.6 Item difficulty level

According to Gregory (2007), item difficulty is depicted by the *p*-value and is the proportion of individuals who respond to the item correctly. The higher the proportion of those who respond to the item correctly, the easier it is, while when fewer individuals respond to it correctly, it can be viewed as a difficult item (Foxcroft, 2013; Gregory, 2007).

1.6.5.7 Africanising or Africanisation

Africanising or Africanisation is a term associated with African heritage, African influence and African culture (Franke & Esmenjaud, 2008; Louw, 2009; Msila, 2009). It has been viewed as a process of applying and inculcating the African influences or culture in various environments (Franke & Esmenjaud, 2008; Louw, 2009; Msila, 2009).

1.6.6 Methodological convictions

Methodological convictions include procedures and techniques used to answer the research questions with precision (Salkind, 2014). For this study, these included the administration of tests and *new items*, and item analysis (Foxcroft, 2013; Moerdyk, 2015).

1.6.6.1 Administration of tests

It is imperative that the administration of tests is handled the same (standardised conditions) for everyone (Moerdyk, 2015). The process is divided into three stages, namely the preparation that needs to be done before the testing session, such as checking the testing venue and testing materials; the activities performed during the testing session, such as establishing rapport with test takers and ensuring that the instructions are clearly understood; and lastly, the responsibilities after the testing session, such as collecting and securing the testing material (Griessel, Jansen, & Stroud, 2013; Moerdyk, 2015).

1.6.6.2 Item analysis

Item analysis is the examination of the quality of each item to ensure reliability and validity (Anastasi & Urbina, 1997; Foxcroft, 2013). The procedure helps with the improvement of items in that items are changed, adapted or eliminated on the basis of the results (Foxcroft, 2013).

According to Trafford and Leshem (2008), the choices and clarifications made in the paradigm perspective section are important for setting boundaries and creating clear

territorial margins on which research claims, arguments and strategies for the research are founded.

1.7 RESEARCH DESIGN

A research design is a strategic framework that serves as a link between the research questions and implementation of the research, in order to ensure maximum validity and minimum error in the study (Babbie & Mouton, 2010; Durrheim, 2006). The research journey depicted in figure 1.4 is based on the adaptation of the magic circle model (Trafford & Leshem, 2008), which provides a strategic overview of the research process. Based on the critical realism paradigm, in which both the qualitative and quantitative methodologies are accepted (Krauss, 2005), a mixed method approach was used in this study, specifically the exploratory sequential mixed method research design with quantitative research processes embedded in it (Caruth, 2013; Creswell, Klassen, Plano Clark, & Smith, 2011). The research design and research questions are discussed in detail in chapter 4 of this thesis.

Figure 1.4 depicts the interconnectedness of all the parts of the research process (Trafford & Leshem, 2008). As per the interconnectedness explained for the context of testing and assessment, and the critical realism research paradigm, the magic circle provides a similar premise where each part of the research is connected to the overall research. The adaptation of the magic circle to the study is discussed in chapter 7, where the conclusions, limitations and recommendations for the study are presented. That discussion indicates how the gap in knowledge was addressed by the research. According to Syed, Mingers, and Murray (2009), it is essential to ensure a balance between research rigour and relevance in practice.

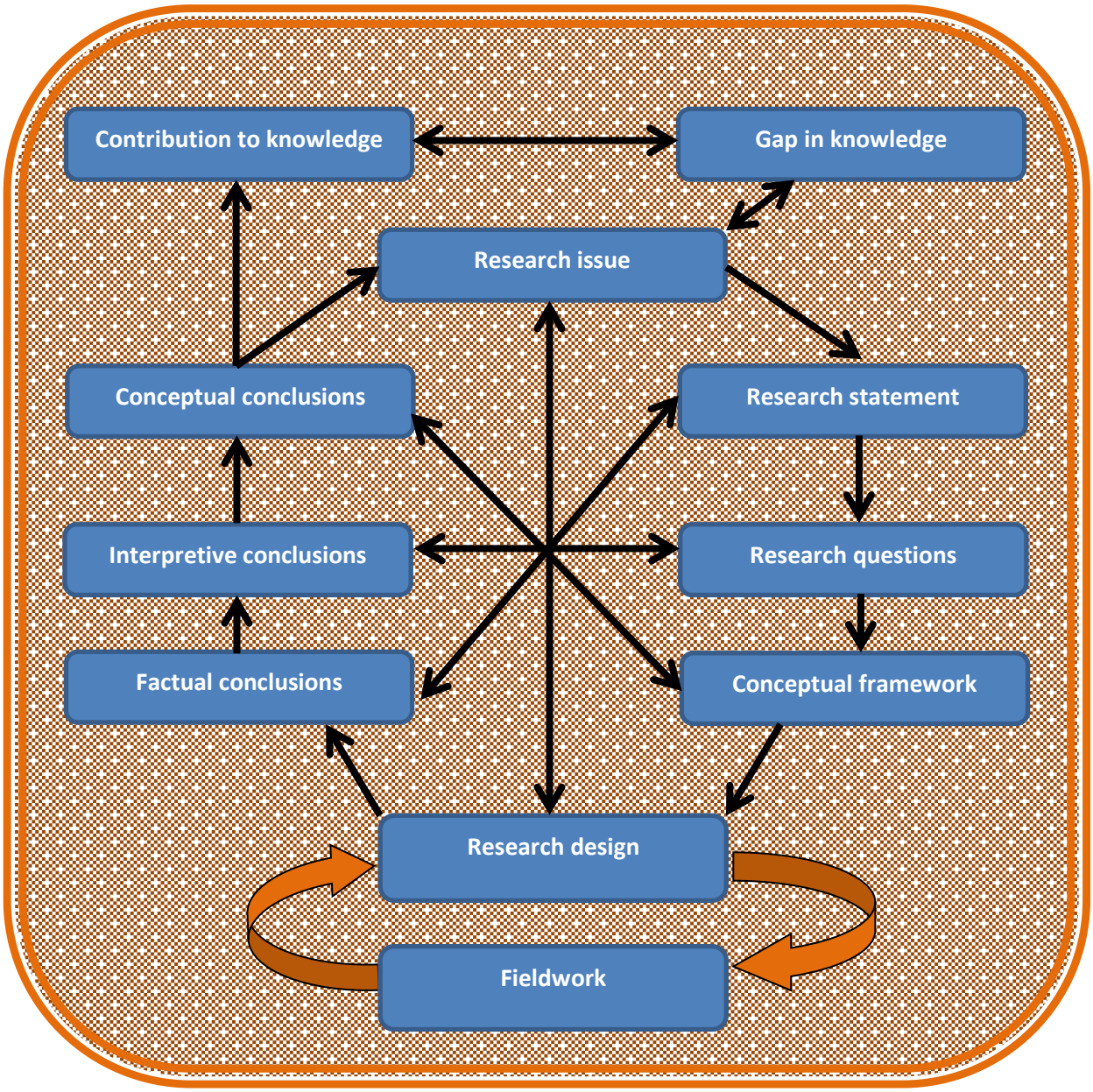


Figure 1.4. The magic circle (adapted from Trafford & Leshem, 2008, p. 170)

1.8 CHAPTER LAYOUT

The study is presented in seven chapters, with each chapter having its own title and discussion of relevant content, as indicated below. Depending on the discussion, the information is also presented in figures, tables, photographs and diagrams, which are included as part of the chapters. Appendices are included at the end of this thesis, and these will be referred to in specific chapters.

Chapter 1: Scientific orientation to the research

In this chapter, the background to and rationale for the research were discussed. The problem statement, the aims of the research, the paradigm perspective of the research and the research design were identified.

Chapter 2: Multicultural cognitive assessment

The focus of this chapter is to gain an understanding of the construct on which the *new items* were based. An overview of the measurement of cognitive functioning, with specific focus on the challenges and debates on multicultural issues, is presented. Theories that contributed to defining and understanding the assessment of cognitive functioning are highlighted, with particular emphasis on the theory of fluid and crystallised intelligence. This chapter helps to clarify and define the construct that was measured by the *new items*.

Chapter 3: Criteria for evaluating a measure

This chapter provides an overview of measurement, reliability and validity in relation to their importance as part of the development of a test or scale. The discussion of the measurement theories, item analysis and differential item functioning are also included in so far as they relate to the particular focus of the present study.

Chapter 4: Research design and methodology

This chapter deals with how the research was conducted, with details of the sample, the measures (instruments) used and the procedures followed to collect and process the qualitative (viability) and quantitative (utility) data. The precautions taken to ensure reliability and validity and the ethical considerations are discussed.

Chapter 5: Development of the new format items

This chapter explains the motivation for developing the *new items*. The process of planning and writing the *new items* is discussed, and the different item types that were included in the final item pool are presented.

Chapter 6: Results

In this chapter, both the qualitative and quantitative results of the study are presented, thus providing answers to the stated research questions regarding the viability and utility of the *new items*. The results are also debated using previous research to substantiate the meaning and interpretation of the results.

Chapter 7: Conclusions, limitations and recommendations

In this chapter, the interpretation and implications of the results are discussed, as well as the limitations of the study and recommendations for future research.

1.9 CHAPTER SUMMARY

In this chapter, an orientation to the research was presented starting with the background to and rationale for the research. The problem statement, general research question and aims of the study were outlined in detail. Subsequent to that, the paradigm perspectives within which the research was conducted were explained and the research design discussed. The chapter concluded with the chapter layout. In terms of the quotation at the beginning of the chapter, the presentation provided a vision of what the research was about, what it was founded on and the pathway to achieving the research aims. The bigger picture of the research was set out in this chapter (Trafford & Leshem, 2008). Chapter 2 highlights the strides made in multicultural cognitive assessment.

CHAPTER 2

MULTICULTURAL COGNITIVE ASSESSMENT

You're born with a gift. If not that, then you get good at something along the way.
Quotation from the motion picture – "Along came a spider."

2.1 INTRODUCTION

Research in the field of cognitive functioning and its assessment has a fairly lengthy history. Although huge strides have been made in the development of various instruments to assess different aspects and elements of cognitive functioning, the fundamental debates and concerns are not much different from the debates of decades past. Debates such as how best to test and assess intelligence – specifically of people from different contexts and backgrounds – continue. The concerns such as the lack of a common definition of intelligence, the effects of socioeconomic and educational background on test performance, the interpretation and use of the test results and the need for the culture fairness of the tests (Anastasi & Urbina, 1997; Claassen, 1997; Gregory, 2007; Kanjee, 2006; Papalia & Olds, 1988), which were issues of concern in earlier years, are more important now than ever before. These debates continue to be a source from which vital research has emanated.

As new research is embarked on, it is always advisable to have some sense of the past before venturing into the future. This helps to put in perspective how much has already been done, acknowledging the foundations on which new research is or should be built and highlighting the development areas that new research should be ready to address. In line with the aim of this research to develop and evaluate *new items*, it is necessary to start with a clear understanding of what the *new items* are intended to measure – hence the focus of chapter 2.

This chapter provides an overview of the measurement of cognitive ability, which highlights the early developments of cognitive assessment both internationally and in South Africa, the key elements in various definitions of intelligence, the theories that

have contributed to defining and understanding intelligence, and the debates and challenges relating to or emphasised by multicultural cognitive assessment.

2.2 OVERVIEW OF THE MEASUREMENT OF COGNITIVE FUNCTIONING

A lot of work has been done on testing and assessment, dating as far back as biblical times, ranging from astrology to graphology, from looking at physical features to palm reading – all these were attempts to assess human attributes (Foxcroft, Roodt, & Abrahams, 2013b). These attempts were driven by various pioneering people, but the discussion for this research will start in the 1800s with people such as Francis Galton and James Mckeen Cattell, who initiated quantitative measurements of intelligence in controlled environments in order to ensure the possibility of replication of the studies (Gregory, 2007). Once this background has been outlined, the discussion will focus on the ongoing progress in the development of cognitive measurements.

2.2.1 The early developments internationally

The development of theories on or the understanding of intelligence during the earlier years is discussed below based on the adaptation of the periods as presented by Ryans (1938). The first period entails work done between 1890 and 1910, a second period between 1910 and 1920 and the last period between 1920 and 1938, when the article was published.

2.2.1.1 The period between 1890 and 1910

The ground work for this period was laid early on by people such as Jean Esquirol, Francis Galton and Wilhelm Wundt, where Esquirol worked on differentiating between insanity and mental retardation in 1838; Galton worked on classifying individuals based on their differences in 1869; and Wundt established an experimental laboratory for psychological research in 1879 (Moerdyk, 2015; Ryans, 1938; Sattler, 2001).

According to Moerdyk (2015) and Sattler (2001), the research by Francis Galton focused mainly on human evolution and intelligence assessment. The first experiments conducted by Galton were biological and physical measurements which were linked to the assumption that knowledge acquisition was related to the levels of sensory discrimination and motor abilities (Galton, 1869; Moerdyk, 2015; Ryans, 1938; Sattler, 2001). The sensory perceptions were related to levels of judgement and intelligence (Anastasi & Urbina, 1997; Galton, 1869; Gregory, 2007). Therefore the higher the sensory discrimination and motor abilities (psycho-physical abilities) the individual indicated, the higher their intelligence levels were deemed to be (Anastasi & Urbina, 1997; Galton, 1869; Gregory, 2007; Sattler, 2001).

Studies from controlled environments, such as the laboratory settings, highlighted the fact that measurement could be conducted objectively if standardised procedures were used, thus enabling comparison and replication (Gregory, 2007). Conducting laboratory-based studies was essential because the characteristics of objectivity, comparison and replication are still important for testing and assessment in the present day. Other contributions from the work of Galton were in the use of questionnaires, rating scales and statistical methods for analysing individual differences (Anastasi & Urbina, 1997; Galton, 1869), which have also been and are still important in testing and assessment.

In 1890, Cattell continued on a similar path to Galton and administered tests which included tasks such as naming colours while being timed, judgement of time, memory (number of letters remembered) and reaction time to sound (Anastasi & Urbina, 1997; Gregory, 2007; Sattler, 2001). Cattell contributed the term “mental tests” when he first used it in psychological literature (Anastasi & Urbina, 1997; Gregory, 2007). However, his work was limited by the lack of validity confirmation for intelligence testing (Anastasi & Urbina, 1997).

In terms of conceptualisation of intelligence, it was during this period that Charles Spearman came up with his two-factor theory (Smit, 1996). According to Spearman, there was one single general factor (*g*) which encompassed all mental activities (Spearman, 1930; Ryans, 1938). This theory will be discussed later under the theories of intelligence (section 2.4).

The work of Alfred Binet has been widely acknowledged, and the test he and Theodore Simon developed and published in 1905 (Anastasi & Urbina, 1997; Binet & Simon, 1916) was the first to be recognised as a test that measures intelligence (Anastasi & Urbina, 1997; Gregory, 2007; Moerdyk, 2015). According to Benjamin and Baker (2004), this test, unlike the previous ones, indicated a strong correlation with academic performance. The test, known as the Binet-Simon scale, was used to identify children who needed and could benefit from special education programmes. It was thus developed to predict academic performance (Binet & Simon, 1916; Gregory, 2007; Moerdyk, 2015). Through this work, Binet established that intellectual ability was better represented and measured by various tasks such as judgement, comprehension and reasoning (Anastasi & Urbina, 1997; Binet & Simon, 1916; Ryans, 1938). The format of testing adopted was to present the questions in sequence, according to their difficulty levels, starting with easier items and progressing to more difficult ones, and with testing continuing for as long as the answers that were given were correct (Binet & Simon, 1916; Moerdyk, 2015). This format in some way initiated the use of item difficulty, which remains a significant feature in applications using item response theory (IRT) (De Beer, 2006; 2007).

As mentioned earlier, most of the initial work by Binet and Simon was done with children with mental challenges in an effort to determine whether they could benefit from special education programmes, thus determining their potential to learn to some extent (Anastasi & Urbina, 1997). Hence the assumption was that intelligence levels could improve as the child grows and is exposed to environments that further include training (Anastasi & Urbina, 1997; Binet & Simon, 1916; Ryans, 1938). Contrary to the view that intelligence was dependent on hereditary factors only, Ryans (1938) suggested that the work of Binet and Simon highlighted an alternative understanding of intelligence as entailing continuous adaptation by the individual.

The revised versions of the scales of Binet and Simon were extended to normal children, therefore providing variety in the group composition and making it possible to compare results of children from different backgrounds (Anastasi & Urbina, 1997; Binet & Simon, 1916). For example, when testing a ten-year-old, the results would be compared and interpreted based on whether the child performed within, below or

above average compared to that age group. Through this, the use of mental levels for comparisons was introduced (Anastasi & Urbina, 1997; Binet & Simon, 1916).

Clearly, the overall contribution made during this period between 1890 and 1910, was that of an evolving understanding of intelligence from only a concept into a construct that could be described by theories and measured objectively by means of standardised procedures and instruments (Ryans, 1938).

2.2.1.2 The period between 1910 and 1920

The second period was mostly characterised by an increase in test development, test administration and more revisions and adaptations of existing tests (Smit, 1996).

Binet's work was clearly the foundation of many tests during this time – it was being used to evaluate people in different countries, and a variety of work was conducted with different people in different environments and from different backgrounds, thus making contributions to the better understanding of measuring intelligence and adding to the body of research evidence (Anastasi & Urbina, 1997). As more research was undertaken on Binet's scales, more lessons were learned and the initial scales were revised and improved (Anastasi & Urbina, 1997).

Henry Goddard was one of the people who assumed responsibility for translating and standardising the Binet scales and publications for the American population (Foxcroft et al., 2013b). As noted in Foxcroft et al. (2013b), Goddard's test administrations were to immigrants who at the time of being tested had had little rest; were responding to tests that were administered in languages that had been translated maybe three or more times from the original language (French); and using norms from France. Not surprisingly, most of the immigrants showed low intelligence levels. This work highlighted the importance and impact of language, test sophistication, test administrator characteristics, illiteracy, the well-being of the test taker, conditions of testing and norms used for the test results, which are still issues that are presently considered very important in testing.

Lewis Terman also made adaptations to the Binet-Simon scale, which resulted in the introduction of a new test, which he named the Stanford-Binet as he was working at Stanford University at the time (Anastasi & Urbina, 1997; Moerdyk, 2015). He also ventured into the comparison of mental levels as Binet did, but used the ratio between mental and chronological age to devise the formula for the intelligence quotient or IQ, a term first used by William Stern (Anastasi & Urbina, 1997; Moerdyk, 2015). The use of IQ as a term and a calculated value has stood the test of time as it is still used today; however challenges and misconceptions about the term continue to raise interest in research and in the acceptance of such tests by the public in general (Anastasi & Urbina, 1997).

As the use of tests increased and the research on them progressed, more participants varying between children, military staff and immigrants were tested (Anastasi & Urbina, 1997). The questions of language, characteristics of the test administrator, issues of disability, illiteracy, the well-being of the test taker and conditions of testing were raised and concerns about the appropriateness of test use were voiced (Gregory, 2007). Contributing to these concerns was the fact that the original test was in French and had to be translated, the tests were mostly verbal, testing was conducted through translators, in certain circumstances not much rapport was established with the test takers which were children (some with challenges), military staff and immigrants who were anxious about the unfamiliar environment of testing (Anastasi & Urbina, 1997; Foxcroft et al., 2013b; Gregory, 2007). It is important to note that the questions and the issues of concern raised then are still important today, particularly in the diverse South African context.

Most of the earlier initiatives were focused on individual tests. The work of people such as Robert M. Yerkes and Arthur S. Otis was therefore significant in the culmination of group testing (Anastasi & Urbina, 1997; Gregory, 2007). The year credited for the introduction of group testing is 1917 (Anastasi & Urbina, 1997; Gregory, 2007; Smit, 1996). Most of the work was done in a military setting and the tests were mainly used on military recruits for placement purposes (Gregory, 2007; Smit, 1996). The two tests that were first recognised as group intelligence tests were the Army Alpha, which contained verbal content and was appropriate for those

recruits who had English as their first language; and the Army Beta, which was nonverbal in content in an effort to cater for recruits who had English as a second or third language or those recruits who were illiterate (Anastasi & Urbina, 1997; Gregory, 2007; Smit, 1996). The challenges of language use in tests and the use of nonverbal format tests to address them is still relevant today – hence the use of nonverbal content items for this study to avoid language which could impact on the culture fairness of the stimulus material.

2.2.1.3 The period between 1920 and 1938

The last period Ryans (1938) noted was that between 1920 and 1938 which mainly focused on analysing and critiquing the instruments and theories used at the time. In 1935, Louis Thurstone used the technique of factor analysis in developing his theory of primary mental abilities in which he challenged the existence of one primary ability as theorised by Spearman, and instead introduced the existence of multiple primary abilities (Moerdyk, 2015; Smit, 1996; Thurstone, 1936). Clearly, this theory was in contrast to the single general factor proposed by Spearman. The discussion on the theories of intelligence will be continued in section 2.4.

Another significant development during this period was the work of David Wechsler, who developed a test of his own to test verbal and performance abilities with scales for both children and adults (Moerdyk, 2015; Smit, 1996; Wechsler, 1939). Most of this work was published in 1939 and later (Anastasi & Urbina, 1997).

2.2.1.4 The period from 1938 to the present

Once the measurement of intelligence had been initiated, conceptualised and operationalised in the period prior to and during 1938, research continued with further developments of new tests and revisions of the original ones (Gregory, 2007). However, as the acceptance of testing grew, so did the criticisms. Part of the acceptance of group tests arose from the convenience of test administration as they were ideal for mass recruitments and selection decisions where both cost and time were saved, and similar conditions of testing for all persons could be achieved (Anastasi & Urbina, 1997). The criticisms related to the increase in the use of group

tests without proper caution being exercised regarding the limitations of the tests (Anastasi & Urbina, 1997). Most criticisms of the tests were about the extent to which they were dependent on language which disadvantaged many who were not well versed in the language of the test (Anastasi & Urbina, 1997; Gregory, 2007; Foxcroft et al., 2013b). However, language issues were only a start; over the years, the focus was broadened to include multicultural issues of which language was one. This is discussed in detail in section 2.7.

As mentioned earlier, the test developed by David Wechsler had a performance test section which did not require language proficiency (Foxcroft et al., 2013b; Moerdyk, 2015). Similar developments were made as time progressed and different understandings of the nature and structure of intelligence came to the fore. According to Foxcroft et al. (2013b), criticisms regarding language, verbal skills, illiteracy and so forth, can be viewed positively when they prompt considerations and improvements in the development and use of tests.

2.2.2 Early developments in South Africa

The development and use of psychological tests in South Africa followed a similar route to that in Europe and the United States of America except for the context. South Africa was characterised by racial segregation and unequal distribution of specifically socioeconomic and educational resources (Claassen, 1997; Foxcroft et al., 2013b, Nzimande, 1995). According to Foxcroft et al. (2013b), the discriminatory context of South Africa led to questionable use of psychological tests in their administration and standardisation, and the interpretation and use of results.

In contrast to Anastasi and Urbina's (1997) assertion that intelligence tests should not be used to label people, Foxcroft et al. (2013b) noted how words such as superior and inferior were used in Fick's reporting of the differences between the black and white children he had tested. Regrettably, the tests were being used to confirm and propel the labelling of one race as being better than another. According to Claassen (1997), tests were separately developed for Afrikaans- and English-speaking groups, thus excluding the other language groups which were in the

majority in South Africa. Apartheid legislation and policies continued to fuel misuse of tests, where norms from imported tests were used for South African groups and decisions were made on the basis of test results without considering the environmental impact (Foxcroft et al., 2013b).

Huysamen (1996) noted how South Africa was more open to the use of group tests than individual tests in contrast to how other countries received the large-scale testing. Many of the tests used in South Africa were adaptations from imported tests. Smit (1996) cited the following as examples: the Binet-Simon-Goddard-Healy-Knox Scale by Dr Moll; the Grey revision of the Binet-Terman Intelligence Test by Dr Eybers; and the Fick Scale based on Stanford-Binet Scale by Dr Fick. The concern raised about all these adaptations was the suitability of using such tests for different cultural groups (Claassen, 1997). Tests were seen to be used to exclude certain groups of people from the job market, while promoting another group as better. South Africa thus also experienced the negative perceptions and rejection of tests (Foxcroft et al., 2013b) that were experienced internationally.

2.2.3 Developments in South Africa after 1994

Over the years, with sociopolitical circumstances changing, trends in testing also changed, such as South Africa starting to appreciate the importance of cross-cultural issues (Schaap & Vermeulen, 2008). More emphasis was placed on research that empirically investigated test bias, the fair use of tests for all cultural groups and validation studies of existing tests for different culture and language groups (Foxcroft, 2004; Foxcroft & Aston, 2006; Paterson & Uys, 2005; Theron, 2007; Van de Vijver & Rothmann, 2004). The focus of testing started to shift from separation and exclusion to inclusion of all groups. Furthermore, research had shown that culture and language were important moderators of test performance (Schaap & Vermeulen, 2008), which meant more caution had to be exercised to ensure better use of tests and encouraging more research on the use of tests in multicultural and multilingual contexts.

The introduction of the Employment Equity Act 55 of 1998 (EEA) also ensured that such a shift was boosted as people working in the field had to adhere to the requirements (Government Gazette, 1998). The EEA-specific requirements (Government Gazette, 1998, p. 7) were covered in chapter 1 (see section 1.2)

According to Van de Vijver and Rothmann (2004), the EEA promoted improvement, change and transformation of tests and testing practices. The EEA is also credited with the increase of the number of studies on cultural differences, test adaptation, test bias, fairness and multicultural assessments (De Beer, 2004; Foxcroft, 2004; Foxcroft & Aston, 2006; Paterson & Uys, 2005; Theron, 2007; Van de Vijver & Rothmann, 2004), specifically with regard to cognitive assessment.

2.3 UNDERSTANDING INTELLIGENCE

Generally, intelligence is understood to be related to performance – if children do well academically at school, then they are regarded as intelligent. If people perform well at work, they are deemed competent, which translates into being intelligent. Similarly, if people do well in life without having any scholarly qualification, they are still regarded as intelligent. People are at times deemed intelligent within a specific environment; for example, an urban-raised person may be deemed intelligent in urban areas, but may be seen as less intelligent in the rural environment, based on the factors that are regarded as important in each of those environments. Another example of different views of intelligence is that of “book smart” and “street smart” people. The abilities of people who are book smart are linked to scholastic learning and qualifications, while street smart people are considered wise enough to perform well because they are able to adapt to the ways of life and the world. Therefore, as emphasised by Moerdyk (2015), the definition of intelligence depends on the people who hold power in that society as they give value to the behaviours deemed intelligent.

2.3.1 Defining intelligence

Moerdyk (2015, p. 149) traced the origins of the word “intelligence” to Latin, specifically the “verb *intelligere*, which means to understand”. Synonyms for the word are comprehending, recognising, knowing and being aware of. It might seem simple enough to view intelligence as the ability to understand, but in reality, intelligence as a construct to be measured has been difficult to define. Over the years, various definitions of intelligence have been formulated, with no one single view being recognised as a generally accepted one (Van Eeden & De Beer, 2013).

For the purposes of this study, instead of listing all the different definitions in the literature, the discussion focuses on the common elements that feature in the different definitions, which will relate to defining the construct as it will be used in the *new items* for the present study. However, it is important to note that caution was exercised in thinking of both the “real definition which entails the true nature of the construct” and the “operational definition which entails how the construct is measured”, as differentiated in Gregory (2007, p.164).

Gregory (2007) and Ryans (1938) credited Binet and Simon as the developers of the first acknowledged test of intelligence, and the definition they gave of intelligence as the ability to judge well, to understand well and to reason well. Such a definition embodies and relates to the simple meaning of the original Latin word *intelligere* (Moerdyk, 2015).

Eysenck (1988) framed the understanding of the term “intelligence” by looking at three different aspects of intelligence, namely biological, psychometric and social. Biological intelligence entails physiology, biochemistry and genetics, while psychometric intelligence includes the standardised psychometric measurements and tests (comprising items that are similar to those in this study), while social intelligence refers to the practical competence of interacting with others (Anastasi & Urbina, 1997; Eysenck, 1988). Psychometric intelligence is based on a group of items given in a test for people to respond to; and then assuming that responding correctly to those items indicates understanding – hence intelligence (Eysenck,

1988; Van Eeden & De Beer, 2013). This study focused on psychometric intelligence.

According to Gregory (2007), learning and adaptation are the core elements of most definitions of intelligence. Other aspects of intelligence highlighted by various definitions encompass the ability to gain knowledge, to adapt to new situations, to think abstractly, to solve problems and to make deductions (Eysenck, 1982; Gregory, 2007; Li, 1996; Smit, 1996). Similarly, Moerdyk (2015) highlighted being able to understand or comprehend, recognise patterns, discover rules, solve problems and process information as important defining aspects of intelligence. These are the same aspects mentioned by McGrew and Flanagan (1998, p. 14) in their discussion of general ability, specifically fluid reasoning ability, where they refer to the mental processes that are used "to form and recognise concepts, identify relations, perceive relationships among patterns, draw inferences, comprehend implications and solve problems".

The above discussion captures the elements of intelligence that were relevant to this study. The items developed and evaluated for this study are nonverbal figural items where participants need to be able to recognise the patterns used in order to complete the various formats of questions; they need to be able to discover what rules are embedded in the item patterns and the relationships between the figures to be able to solve the problem or complete the pattern or series.

2.3.2 The nature of intelligence

A child who passes all his or her subjects with distinctions would generally be considered to be intelligent. However, if the majority of the children in a particular school pass in a similar manner, some questions may be asked in an attempt to understand such excellent performance. For example, questions could be asked about the difficulty levels of the test; or the access to the questions by the children prior to their taking the test, how they were selected to be in that particular school, or what teaching methods were used. The fact is, these questions would be raised because it is generally accepted that the levels of performance vary from above

average, average to below average, with similar levels of performance on intelligence tests usually also distinguishable.

As Gottfredson (1998) suggested that people are born with differing levels of intelligence, in the same way as it is accepted that some people are taller or more athletic than others. However, it is not as simple, because as much as intelligence is seen as stable, it is also viewed as modifiable owing to interactions with the environment (Anastasi & Urbina, 1997). So a person might be born with a certain level of intelligence, but through interventions such as teaching and support, his or her level of measured intelligence may be improved. This raises the question of the differences in intelligence being based on whether people are born or made intelligent, which over the years, has been labelled the “nature versus nurture” debate (Van Eeden & De Beer, 2013).

People such as Galton, Goddard and Jensen held the view of intelligence as a genetically inherited construct (Anastasi & Urbina, 1997; Moerdyk, 2015; Ryans, 1938). Studies conducted on the basis of the assumption that the role of genetics in the family was important in ensuring that offspring were of similar abilities did not clearly provide the evidence to support such assertions (Anastasi & Urbina, 1997; Fancher, 1985; Moerdyk, 2015). Searching for patterns of success in family trees, twins and siblings was included in these studies in attempts to support the hereditary (nature) view (Fancher, 1985). The results of such studies were not conclusive. The work of Binet, however, was based on the assumption that people could be trained – hence the belief that intelligence could be modified for the better (Anastasi & Urbina, 1997; Moerdyk, 2015; Ryans, 1938; Van Eeden & De Beer, 2013). Fancher (1985) noted the work of John Stuart Mill, who believed that intelligence was mainly dependent on the power of training and the support one receives. Mill was taught by his father from an early age, and he attributed his cognitive development to that support and training, as opposed to believing he had been born with the gift (Fancher, 1985).

There has not been unchallenged evidence that either hereditary or environmental factors are the sole contributors to intelligence levels. Anastasi and Urbina (1997) cautioned that answers to individual and group differences may not be limited to nature or nurture explanations, but may also tap into factors such as motivation, emotional and attitudinal variables. With a drive to succeed, stable emotions and a positive attitude, so much can be achieved – which would have an impact on how one performs in the tests. This is why the focus cannot be limited to the instruments only, but would need to include the processes and procedures followed during the administration of the tests, such as ensuring conducive testing conditions (Griessel et al., 2013). Responsibilities like building rapport, motivation and reducing anxiety are vital to ensure that people have been optimally prepared to do well.

Publications hypothesising racial superiority as a result of genetic factors raised controversies and backlashes for IQ testing (Graham & Lily, 1984; Gregory, 2007). One such publication was the book entitled *The bell curve,* which raised earlier stereotypes and misconceptions of ethnic and gender differences in intelligence (Anastasi & Urbina, 1997). History has shown publications and activities associated with such beliefs to have harmful consequences, because laws and policies were implemented to discriminate against people using that kind of information as the motivation and explanation (Gregory, 2007; Kanjee, 2006).

Although intelligence naturally manifests at different levels – similar to a host of other human characteristics – the explanation of how those differences come about continues to garner interest. Gottfredson (1998) commented on how the explanations have been derailed by political and social power plays. The nature of intelligence is also described by theories which provide a framework from which measurement of the construct can be developed.

2.4 THEORIES OF INTELLIGENCE

Numerous theories have contributed to the description and explanation of the nature of intelligence. These theories provide the necessary framework needed to make decisions about how to measure the intelligence construct and interpret the results.

However, only a few theories will be highlighted in this discussion, based mainly on their relevance to the study and some for their contributions to the understanding of intelligence. According to McGrew and Flanagan (1998), most theories can be categorised into three main approaches, that is, the structural (factor analytic), information processing and cognitive modifiability approaches.

The structural (factor analytic) approaches to intelligence are the oldest and most researched approaches that rely on factor analysis to determine the underlying relationships between a variety of cognitive ability tasks and tests (McGrew & Flanagan, 1998; Moerdyk, 2015). Factor analysis is a statistical technique that is used to analyse the intercorrelations between psychological tasks to determine the underlying structure of the theoretical construct being assessed (Anastasi & Urbina, 1997; McGrew & Flanagan, 1998; Moerdyk, 2015; Roodt, 2013b; Smit, 1996). According to Moerdyk (2015), the reason for various structural approaches is that each one uses a different type of factor analysis. The theories highlighted here are Spearman's two-factor theory, Thurstone's theory of primary mental abilities and Cattell's fluid and crystallised abilities theory (Moerdyk, 2015).

According to Moerdyk (2015), the structural approaches made huge contributions in describing the nature and the structure of intelligence but that turned out to be seen as a limitation by some who believed there should be an explanation of how this intellectual performance is possible, thus referring to the information processing (or cognitive) theories of intelligence. These latter approaches focused on how people mentally represent and process information (Gregory, 2007), which means how the input, processing and output of information are successfully managed. The theory discussed here, which falls into this category, is Sternberg's triarchic theory of intelligence.

The cognitive modifiability (or cognitive developmental) approaches describe the increase in complexity of cognitive functioning and were highlighted by the contributions of Jean Piaget and Lev Vygotsky (Cohen & Swerdlik, 2002; Gregory, 2007). According to McGrew and Flanagan (1998), these approaches are significant because they focus on the adaptable nature of intelligence, and the dynamic

assessment techniques were mostly based on their assumptions. The theory discussed here is Piaget's cognitive developmental theory.

2.4.1 Spearman's two-factor theory

Charles Spearman was the first person to propose a theory to explain the nature of the structure of intelligence using the psychometric principles of factor analysis (Moerdyk, 2015; Smit, 1996; Spearman, 1904; 1930). Through the factor analysis technique, Spearman was able to conclude that all intellectual activities share a single common factor, that is, the general factor or *g,* and various specific factors, termed *s,* which were viewed as being strictly specific to separate activities – hence the designation, the two-factor theory (Anastasi & Urbina, 1997; Gregory, 2007; Spearman, 1904). The positive correlations found between intellectual tasks, which Spearman named the positive manifold, were viewed as indicating the presence of the so-called "*g*" factor (Anastasi & Urbina, 1997; Spearman, 1904, 1930). According to Anastasi and Urbina (1997), these high correlations were significant as they strengthened the basis of predictive validity in which the performance of an individual could be predicted from one situation to another.

Spearman (Gregory, 2007; Ryan, 1938; Spearman, 1904, 1930) described *g* as entailing three mental activities, that is, educing relations, educing correlates and self-recognition. According to Gregory (2007, p. 177), the term "educing" means the "process of figuring things out". In explaining these principles, Gregory (2007) stated the following: people faced with a problem are able to identify the significant elements of the problem through their own perceptions and understanding (self-recognition); they can take the identified significant elements and try to determine the relationships between them and make inferences about those relationships (educing relations); and they can apply such inferences to new problems or situations (educing contexts).

Tests that were developed on the basis of the measurement of *g* were Raven's progressive matrices (RPM) and Cattell's culture-fair intelligence test (CFIT) (Anastasi & Urbina, 1997). Since the items used for this study were nonverbal figural

items requiring the use of general reasoning skills, this two-factor theory was relevant.

2.4.2 Thurstone's theory of primary mental abilities

The multiple-factor theory by Thurstone presented an opposing view from the single general factor by Spearman and argued for a number of overlapping but independent factors, which he termed the primary mental abilities (Anastasi & Urbina, 1997; Moerdyk, 2015; Ryans, 1938; Smit, 1996; Thurstone, 1938). According to Gregory (2007), the factor analysis procedures used by Thurstone enabled him to search for correlations to indicate the existence of group factors. Thurstone's theory came with multidimensional content for intelligence tests (Sattler, 2001; Thurstone, 1938) because he identified seven primary mental abilities, which included verbal comprehension, word fluency, numerical ability, spatial ability, associative memory, perceptual ability and reasoning ability (Smit, 1996). This theory is less applicable for the items used in this study because of its multidimensionality.

2.4.3 Cattell's theory of fluid and crystallised intelligence

Owing to the fact that Raymond B. Cattell did not fully agree with the single *g* factor, in his theory, he split the *g* factor into two factors, namely fluid intelligence (*gf*) and crystallised intelligence (G_c) (Cattell, 1963; Gregory, 2007). Interestingly, the explanations of these factors were mentioned by Spearman (1930), but with the use of different terms, namely eductive and reproductive abilities respectively. According to Cattell (1963), fluid intelligence is relevant for tests that require adaptation to new situations independently of acquired knowledge, where nonverbal abilities such as problem solving, pattern recognition, learning and abstract reasoning are tested, similar to Spearman's eductive abilities. According to Cattell (1963), the capacity of perceiving relations and educing correlates, as discussed by Spearman, relates to the behaviours proposed for fluid intelligence. Crystallised intelligence, like Spearman's reproductive abilities, is associated with the application of prior learning with verbal abilities such as vocabulary, verbal reasoning and general knowledge being tested (Cattell, 1963; Moerdyk, 2015).

Fluid intelligence is more constant and biologically determined, while crystallised intelligence, which is a product of the environment, is more culturally determined (Cattell, 1963). Moerdyk (2015) highlighted the investment theory of intelligence in which Cattell explained how everybody is born with both abilities of *gf* and *gc*, but mostly using *gf* at a younger age. The balancing comes later on in life when people start investing fluid intelligence in different activities. Depending on the level of mental stimulation of the activities, intellectual growth will be increased, maintained or might even decline (Moerdyk, 2015). Crystallised intelligence (*gc*) and fluid intelligence (*gf*) are assumed to be on par before maturity (15 to 20 years of age), and any differences between them, would be due to cultural opportunity and interest, whereas in adults, the differences are credited to age, experience and time decay of *gf* (Cattell, 1963). It is assumed that fluid intelligence reaches its maximum between the ages of 14 and 15, while crystallised intelligence, depending on the opportunities to learn, can further increase between the ages of 18 and 28, and beyond (Cattell, 1963).

This theory was crucial to the current study because the items that were used focus on measuring the *g* factor, specifically fluid intelligence. The items are independent of educational content and context and have no verbal content, and the requirement is to identify the patterns or sequences and to deduce relationships.

2.4.4 Sternberg's triarchic theory of intelligence

The focus of this theory was on the processing of information and how what has been processed is used and/or adapted in the real-world environment (Gregory, 2007). Sternberg proposed three aspects of intelligence, namely componential, experiential and contextual intelligence (Gregory, 2007; Moerdyk, 2015; Sternberg, 1984). Componential intelligence entails the analytical elements used to analyse, compare and evaluate familiar problems in order to solve them (Moerdyk, 2015; Sternberg, 1984). Gregory (2007) highlighted the important components relevant to this type of intelligence as the ability to plan, reason and acquire a suitable vocabulary of words. Experiential intelligence – also referred to as the creative intelligence – involves solving new problems, and has as its core characteristics, creation, invention and design (Moerdyk, 2015; Sternberg, 1984). Contextual

intelligence was viewed as being concerned with practical adaptation to everyday situations (Gregory, 2007), thus linked to the work of Binet and Wechsler (Moerdyk, 2015).

2.4.5 Piaget's theory of cognitive development

According to Cohen and Swerdlik (2002), Piaget conceptualised intelligence as a cognitively driven process of assimilation and adaptation to the environment, which is both active and constructive in nature. The theory is based on four stages of cognitive development, with each stage characterised by its own unique pattern of thought (Gregory, 2007; Piaget, 1972; Van Eeden & De Beer, 2013). The stages are identified as sensorimotor, pre-operational, concrete operational and formal operational (Gregory, 2007). According to Berk (2000), the stages progress from using the senses to learn about the environment, representing the environment symbolically and in language, using logic appropriately, and lastly, reasoning logically and drawing conclusions. The basis of this theory is that as new information is perceived, new systems are developed in an effort to understand the world better (Berk, 2000; Cohen & Swerdlik, 2002). Some questions have been raised about the relevance of the stages as some children do not necessarily progress according to the stages of the theory (Cohen & Swerdlik, 2002).

2.4.6 Vygotsky's theory of the zone of proximal development

The focus of this theory is on the changeable nature of intelligence, and how the current level of functioning of a person can reach without help is differentiated from the level of functioning a person can reach with help (Van Eeden & De Beer, 2013; Vygotsky, 1978). The difference between functioning without help and functioning with help is called the zone of proximal development (Vygotsky, 1978). The basis of this theory is the link between learning and development where the learners are presented with progressively difficult tasks to complete so that learning potential could be determined (Shabani, Khatib, & Ebadi, 2010). The process incorporates dynamic assessment where both the current and future level of performance are determined through a test-train-retest approach (Shabani et al., 2010; Van Eeden & De Beer, 2013). Shabani et al. (2010, p. 240) referred to this process as the

determination of “abilities that are fully matured as well as those that are still in the process of maturing.” According to Shabani et al. (2010), Vygotsky developed this theory in response to criticisms of static measures which he felt were limited to fully matured ability.

This theory was important to the current study because the measure used to determine the construct validity of the *new items* is a dynamic computerised adaptive measure for the measurement of learning potential.

As mentioned previously, the theories of intelligence discussed above are not exhaustive, but even from the few discussed, one can see how each theory provides a different view of intelligence. The approaches discussed place the focus on different factors, such as the content, the process, the context and the development of intelligence. The measurement methods and interpretation of results are guided by the fundamental assumptions of the theory chosen.

2.5 MEASUREMENT OF INTELLIGENCE

According to Smit (1996), the progress in the measurement of intelligence was driven mainly by a response to educational, clinical and research needs. It may in fact have started in order to identify children with special schooling needs (Binet & Simon, 1916; Gregory, 2007), but testing has continued beyond that. According to Paterson and Uys (2005), some of the reasons cited by many practitioners in recent times for using assessment included selection and recruitment, training and development, succession planning, placements, career counselling and development. Although there are various reasons for testing, the underlying common element is that the measure must be appropriate for the defined construct and for the particular group and context. This is where the theories of intelligence help to narrow the elements that need to be included in the test battery. According to Anastasi and Urbina (1997), because tests measure a sample of behaviour, the selected theory should provide the framework for that sample of behaviour. As noted earlier, there

are many theories and definitions of intelligence, which have resulted in many more tests being developed for testing intelligence.

For this study, as mentioned previously, the theoretical framework for the development of the *new items* was Cattell's theory of fluid intelligence. According to Cattell (1963), the core function of fluid intelligence is abstract reasoning ability which measures the logical thinking involved when solving problems that identify patterns and similarities between shapes and figures. Abstract reasoning, also known as the ability to reason inductively and deductively, is the process of looking from the specific to the general for inductive reasoning – or vice versa for deductive reasoning – and being able to recognise the rules governing the pattern changes and relationships between figures and symbolic materials (Smit, 1996).

From the different methods used for testing intelligence listed by Moerdyk (2015), namely the series items, matrix items, odd one out, general knowledge items, assembly tasks, group and individual assessment, verbal and performance scales, the series, pattern completion and matrix items were deemed relevant for this study. The *new items* were developed using these methods by presenting a series or matrix of figures and shapes – inspired by African art and cultural artefacts – where participants would be expected to complete the patterns using either size, shape, position and/or quantity with logical principles that could be used to make inferences. According to Lohman (2005), this format of nonverbal reasoning testing requires participants to label the figures by shape or form (e.g. triangle, circle), note their attributes (e.g. size, colour) and recognise the rules that generate the changing features or elements (e.g. figure becomes bigger). Fluid general intelligence is therefore measured by means of nonverbal testing formats which are appropriate for cross-cultural applications (Lohman, 2005). Nonverbal tests are further discussed in section 2.7.4.

2.6 USES OF INTELLIGENCE TEST RESULTS

As Gregory (2007, p. 266) concisely puts it, an intelligence test is a "neutral, inconsequential tool until someone assigns significance to the results derived from

it." It is therefore not only about the quality of the test, but also about how its results are used. Over the years (as explained in the earlier sections), there have been numerous misconceptions, misunderstandings and controversies surrounding the use of intelligence tests, starting with the meaning of IQ (Anastasi & Urbina, 1997; Gregory, 2007; Owen, 1998). People tend to forget that IQ scores, like other tests, are based on samples of behaviour and should not be inflated to defining people's worth (Gregory, 2007). It would be erroneous to think that intelligence test results are 100% accurate, ignoring the fact that the test can only measure a sample of behaviour representative of the psychological construct rather than the construct in all its diversity (Owen, 1998; Van Eeden & De Beer, 2013). Furthermore, IQ scores should not be viewed as being unchangeable – nor should they be seen as exhaustively defining the index of diverse intellectual abilities (Gregory, 2007; Owen, 1998). Tests only provide a part of the information that can be generalised outside of the test situation, which leads to probabilities rather than certainties (Owen, 1998). Anastasi and Urbina (1997) highlighted the need to regard IQ scores for descriptive purposes, rather than as labels of adequacy or inadequacy. They also noted the importance of viewing IQ scores as comprising various functions, as opposed to regarding them as a unitary ability.

The use of test results is vital in making appropriate decisions about education, work and life. Such results therefore have to be understood in the context and conditions in which they were obtained. Grieve and Foxcroft (2013) distinguished between the biological context (e.g. age, physical impairments and other physical bodily functions), the intrapsychic context (e.g. personal experience and feelings) and the social context (schooling, language, socioeconomic status, urbanisation, etc.). Understanding these contexts and how they can impact on test performance would make it possible to use the test results with a better understanding of the potential limitations.

Although the use of test results had previously not been entirely to the benefit of the majority of people in South Africa, this is changing. In the meantime, the ongoing challenge in both the measurement of intelligence and the use of intelligence tests

lies in the underlying difficulty, namely the multicultural context in which the testing takes place – hence the focus of the next section.

2.7 ISSUES IN MULTICULTURAL COGNITIVE ABILITY TESTING

With globalisation and all the technological innovations and advancements, the world has become quite small. The exposure to and understanding of different environments and contexts have increased as technology has made it possible to communicate with people all over the world. People emigrate, become expatriates or could be refugees. These, amongst other factors, make countries more likely to be multicultural in nature. South Africa is no different, but even without all the above, the South African population has been characterised by multiple cultures (Van Dulm & Southwood, 2013). South Africa is often praised for its 11 official languages, which were recognised after 1994, but within those languages groups various cultures are embedded. Eleven languages may seem a lot. However, Van de Vijver and Phalet (2004) noted that if one considers the 200 languages that are spoken in the public schools of Chicago, then one begins to understand the challenge of testing in a multicultural context.

This section is therefore important as it will provide an overview of the issues of concern in multicultural cognitive ability measurement. Some of the key concepts that are synonymous with multicultural assessment research are bias and equivalence (Meiring, Van de Vijver, Rothmann, & Barrick, 2005; Van de Vijver & Rothmann, 2004). These concepts will be discussed in detail in the next chapter. In this section, Owen's (1992) suggested reasons for bias, namely culture, socioeconomic status, language and cognitive styles, will be dealt with as part of multicultural assessment. The discussion starts with a brief overview of cross-cultural psychology, which is included here, because it is through this subdiscipline that cross-cultural comparisons and culture-fair testing have been highlighted. Multicultural assessment is discussed next, after which the role of acculturation is explained.

2.7.1 Cross-cultural psychology

Cross-cultural psychology is based on the principles of anthropology and psychology where the systematic relationships between behavioural variables and ethnic-cultural variables are studied (Matsumoto, 2001; Yau-Fai Ho, 1994). It is a study of similarities and differences in and between groups of people in terms of the role that cultural factors and socialisation play in the shaping of human behaviour (Bergh, 2009). Matsumoto (2001) highlighted processes such as the perception and language of colour, the processes of language acquisition, the principles of cognition, thinking and learning; gender differences, and gender stereotypes as important in this subfield of psychology. Being more aware of cultural similarities and differences would help make it possible to develop culturally sensitive and inclusive methods of psychological assessment, which is essential in the understanding of processes and in promoting relationships between people and psychological interventions such as psychological assessment (Bergh, 2009; Matsumoto, 2001).

Cross-cultural psychology was also significant as a context for the present study to appropriately develop and evaluate the *new items* concerned because it comprises cross-cultural use and culture-fair testing as its traditions (Poortinga, 1995). According to Poortinga (1995), cross-cultural comparison studies were embarked upon to determine the universality of human intelligence, while culture-fair testing was seen as a response to overcoming the challenges in fair and nondiscriminatory use of tests. The culture-fair label replaced the initial references to “culture-free” test attempts, which failed to gain approval because tests could not be free of culture because culture is infused in most if not all environments (Anastasi & Urbina, 1997).

2.7.2 Multicultural assessment

People from different cultural groups may vary in terms of cultural values, language and language styles, views of life and death, problem-solving strategies, mental health and illness, and the stages of acculturation (Gregory, 2007). They also differ in terms of their test performance. It is necessary to determine which tests and assessments are more culture appropriate based on test/item bias, equivalence and fairness considerations (Van de Vijver & Rothmann, 2004). Although acknowledging

many other factors that could contribute to differences in cognitive test performance, Owen (1992) highlighted culture, socioeconomic status, language and cognitive style as the most important.

2.7.2.1 Culture

It is a challenge to test people from different backgrounds and this can be considered a handicap when the people concerned have to compete and succeed in a cultural context that is not their own (Anastasi & Urbina, 1997). This is applicable to the South African context, where the majority of the people were not exposed to testing but found themselves having to perform on tests that were not standardised for them and being compared with groups that were familiar with the language and format of testing. In fact, it has been said that the test reflects the cultural background of the test developer, which might create a disadvantage for those who come from a different culture (Anastasi & Urbina, 1997; Grieve & Foxcroft, 2013). Furthermore, Foxcroft (2004, p. 10) highlighted how the understanding and operationalisation of the construct of intelligence is different per cultural background, with Western cultures regarded as “mentally sharp and quick thinking”, while Eastern cultures are viewed as “thoughtful and reflective”. Therefore, as much as culture might be related to cognitive development and the acquisition of skills, it could also influence where and how people use those skills (Owen, 1992). Consequently, culture is often acknowledged as having a considerable impact on test performance (Paterson & Uys, 2005).

2.7.2.2 Socioeconomic status

Research has mostly focused on evaluating and addressing the equity and fairness aspects of testing and assessment, but Bradley and Corwyn (2002) suggested that the importance of socioeconomic status should not be overlooked. Socioeconomic status entails factors such as education, occupation and income which, on the negative side, are associated with poverty, ill health, poor schooling facilities and lack of support, to mention but a few (Grieve & Foxcroft, 2013; Owen, 1992). This means that the level of socioeconomic status to some extent determines the resources and opportunities to learn that would be available and accessible to

people (Grieve & Foxcroft, 2013). Since the level of socioeconomic status is related to the presence or lack of opportunity for cognitive development (Bradley & Corwyn, 2002), it means people can either be advantaged or disadvantaged in test performance on the strength of their level of socioeconomic status (Grieve & Foxcroft, 2013; Owen, 1992).

2.7.2.3 Language

One should bear in mind that language has a moderating effect on test performance when the language of the people being tested is different from the language used in the test (Anastasi & Urbina, 1997; Foxcroft & Aston, 2006; Grieve & Foxcroft, 2013; Owen, 1992). With the eleven official languages in South Africa, it becomes more difficult to ensure that tests are available in all these languages, as translations are not necessarily cost and time effective (De Kock, Kanjee & Foxcroft, 2013). In addition, some English words are difficult to translate in one or two vernacular words that might be required, thus making translations more difficult. It has also been highlighted that translations may affect the difficulty level of the test and the issue of various dialects further complicates the possibility of fair and proper translations (De Kock et al., 2013). Another challenge with language is that although English is not a first language to many in South Africa, people are being taught in that language and become accustomed to being tested in it, thus making it difficult to decide what the best language is for testing (Grieve & Foxcroft, 2013). Historically, nonverbal content has been used to address the issue of language in an attempt to level the testing experience (Anastasi & Urbina, 1997). Nonverbal testing will be discussed in section 2.7.4.

2.7.2.4 Cognitive styles

Cognitive styles refer to people's preferences in terms of how they perceive, remember, think and solve problems (Anastasi & Urbina, 1997), and have been identified as contributing to differences in test performance (Owen, 1992). The main styles identified are the field dependence and field independence styles, which entail how people organise experiences either from an analytical orientation or a global

perspective (Owen, 1992). The field independent people are seen as active participants in their learning, while field dependent people take the role of spectator (Anastasi & Urbina, 1997). Owen (1992) differentiated between these two styles by identifying people using the field independent style as being able to focus on the relevant stimuli independently from other irrelevant more attractive stimuli. They have been found to perform better in psychometric tests that use hidden and embedded figures and block designs (Anastasi & Urbina, 1997; Owen, 1992).

2.7.3 The role of acculturation in assessment

The old saying, "*when in Rome, do as the Romans do*", befits the discussion of acculturation. In relation to testing and assessment, the question of interest would be whether *people can perform as well as the Romans do just because they are in Rome* – meaning, the effect that acculturation has on testing and assessment.

Acculturation occurs when people from different cultures interact on a continuous basis resulting in changes in each of those cultures (Van de Vijver & Phalet, 2004). Haslberger (2005) described it as a process in which people become competent in functioning successfully in a culture that is not originally their own. Similarly, Grieve and Foxcroft (2013) referred to it as a process in which people become integrated into a culture. Van de Vijver and Phalet (2004) identified four strategies of acculturation, in which people

- establish a good relationship with and adapt to the new (host) culture – called an integration strategy,
- maintain a good relationship with the old (original) culture – called the separation strategy,
- let go of the original culture in exchange for a good relationship with the host culture – called the assimilation strategy, and lastly,
- are indifferent to the host or the original culture – called the marginalisation strategy.

As shown by these different strategies of acculturation, it cannot be assumed that the process of acculturation occurs at the same speed or along the same lines

(Grieve & Foxcroft, 2013; Van de Vijver & Phalet, 2004). In terms of how acculturation affects test performance, there is some evidence that the gap in the IQ scores of black and white students is closing (Grieve & Foxcroft, 2013). Claassen (1997) had earlier attributed the identified differences between English- and Afrikaans-speaking groups to changes in the socioeconomic status of the Afrikaans-speaking groups, while in a similar vein, Grieve and Foxcroft (2013) attributed the closing gap between black and white students to the improved quality of education and educational opportunities for black students. Acculturation therefore has the potential to affect the magnitude of differences between cultural groups.

2.7.4 Nonverbal figural items

Throughout this chapter, reference has been made to the *new items* of this study, which were nonverbal and figural in format. These were developed and evaluated to ensure that the multicultural issues of the South African testing context are taken into consideration. The use of these types of items (figure series, matrixes and pattern completion) has been found to be ideal for multicultural contexts such as that of South Africa (Schaap & Vermeulen, 2008). According to Owen (1998), nonverbal intelligence tests were developed as a reaction to criticisms of tests with verbal content being seen as culturally biased. Nonverbal tests were therefore an alternative that was perceived to be more appropriate to address the concerns of people whose language was different to the one used in the test. Using nonverbal tests to measure fluid intelligence meant the items provided a culturally reduced format of testing (Gregory, 2007; Lohman, 2005; Schaap & Vermeulen, 2008). Gregory (2007) defined nonverbal reasoning assessment as being concerned with the ability to understand complex concepts, analyse new information and solve problems using visual stimulus and perceptual strategies. These items appear in diagrammatic form and entail identifying relationships, similarities and differences between shapes and patterns, and recognising visual sequences and relationships between objects (Gregory, 2007; Lohman, 2005).

Lohman (2005) cautioned on the use of nonverbal tests because of the assumption that they are void of culture. He argued that it would be a mistake to think abilities

can be measured in a way that excludes culture, motivation and experience. Anastasi and Urbina (1997) claimed that intelligence is rooted in culture. This claim is supported by Lohman (2005) who believed that even though the items are figural, people understand and label the figures in their language, and if they cannot find the word, then they will use the perceptual strategy to solve the problem – which is limiting for the more difficult items. In his discussion of the culture fairness of these items, Lohman (2005) emphasised culture and cognitive styles as significant factors in the same way as Owen (1992) did. Lohman (2005) noted that for tests to be culturally fair, they need to have items that can be assumed to be familiar across cultures, and cognitive processes that are required to solve the problems should be equally familiar. He therefore encouraged the use of tests combined with other sources of information to make decisions (Lohman, 2005).

2.8 CHAPTER SUMMARY

Based on the discussions in this chapter, the opening quotation seems to be true of intelligence tests and intelligence test results, because for some, intelligence *is a gift, but for others, they get good at it along the way.* As the different viewpoints from the historical development of intelligence testing were presented, showing the continuous progress in testing, acknowledging the lessons learned and challenges faced, chapter 2 focused on exploring the construct to be measured. Intelligence was defined, highlighting the key elements relevant to this study. The discussion indicated the significance of both the views of nature and nurture, and frameworks for understanding the construct of intelligence were discussed as theories. The measurement and uses of the results of intelligence tests were highlighted and the focus on multicultural testing as a crucial growth opportunity emphasised. Having discussed the measurement of cognitive functioning, at this juncture, it would be apposite to address measurement in terms of what it entails and how measures are developed. This is the topic of chapter 3.

CHAPTER 3

CRITERIA FOR EVALUATING A MEASURE

If someone offers to furnish a sure test, ask what the test was which made the sure test sure – Henry S. Haskins (1875-1957)

3.1 INTRODUCTION

Signposts are fundamental to the interconnectedness of development and evaluation. According to Trafford and Leshem (2008), signposts provide indicators of quality during the various stages of research. Similarly, signposts are available in the context of testing and assessment, in which psychometric properties are used to ensure that the criteria for sound quality measurements are met (Gregory, 2007; Moerdyk, 2015).

Since the aim of this study was to develop and evaluate *new items*, it was crucial that the requirements be considered from the outset. This chapter therefore focuses on the concepts that were essential to ensure that the *new items* measured what they were intended to measure in a manner that was fair, consistent, accurate and without bias. The chapter starts with a brief discussion of what measurement is, and important cautions about measurement and measurement errors are highlighted. This overview of the essential elements of measurement provides a vital baseline discussion for the context of measurement and for a common understanding of the other technical and psychometric concepts that are referred to in the chapter. Reliability, validity, fairness and bias are defined and discussed, the differences between fairness and bias clarified and an overview provided of the applications of these for a broader understanding of the concepts. Throughout the discussion, the significance of these concepts for the acceptance and better use of tests in specific contexts is highlighted. In the concluding section in this chapter, these concepts are consolidated for application in a multicultural assessment context. This was deemed necessary because it would bring to the fore the multicultural contextual realities under which the *new items* were developed and evaluated.

3.2 AN OVERVIEW OF MEASUREMENT

Measurement in psychological testing is about the assignment of numbers to unobservable or latent traits (Baker, 2001). Huysamen (1996) referred to measurement as a standardised procedure for assigning numbers to reflect differences in some attributes, while Moerdyk (2015) defined measurement as the process of assigning numbers to observations in order to categorise and quantify meanings. When measuring any psychological attribute, Wright and Stone (1999) agreed that the observations are qualitative, but numbers are assigned to quantify the observations. The rules that are set to control how these observations are made and the meanings attached are important – hence the theory of measurement (Baker, 2001; Krebs, 1987; Wright & Stone, 1999).

Krebs (1987, p. 1834) explained the theory of measurement as the "conceptual foundation of all scientific decisions". This therefore entails the rules of measurement, set assumptions and definitions that provide the framework for valid measures. According to Suppes and Zinnes (1962), the theory of measurement falls within the domain of applied mathematics, and is deemed important for measurement and data analysis. However, caution is necessary, as the measurements used are not necessarily fully representing the traits being measured (Krebs, 1987; Suppes & Zinnes, 1962; Wright & Stone, 1999). The difficulty associated with psychological measurement can be attributed to the invisibility or indirect measurement of the psychological constructs (Baker, 2001; Bond & Fox, 2007; Smit, 1996; Wright & Stone, 1999). As a result of the indirectness of this kind of measurement or the invisible nature of psychological constructs, measurement in such a context cannot be directly compared to measurements in the physical sciences which are usually visible and observable (Baker, 2001; Bond & Fox, 2007; Smit, 1996; Wright & Stone, 1999). Hence the meaningfulness of and deductions made from any psychological measurement should be understood in the context of probabilities, estimation and approximation of what could be considered the true measure of the construct concerned.

McMillan (2000) viewed measurement in terms of evidence and evaluation. He explained measurement evidence as the statistical procedures used for the assignment of numbers and regarded this as necessary in providing a description of the trait of interest, while evaluation refers to the value judgements and interpretations applied to the measurement scores. Moerdyk (2015) also gave a similar explanation of measurement and evaluation and went on to credit the critical realism perspective for this view of measurement. Since the current study was based on the critical realism paradigm, it was therefore deemed appropriate to gain an understanding of measurement in which the importance of the reality of numbers and the meaning attached to it as evidence are recognised. However, of equal importance is the undertaking not to isolate those numbers from practical interpretations and use.

3.2.1 Levels of measurement

According to Gregory (2007), decisions about which levels of measurement are used are important in test construction because each level impacts on the statistical procedures that can be performed. Levels of measurement were first introduced by Stevens in the 1940s (Baker, 2001; Bond & Fox, 2007; Gregory, 2007; Moerdyk, 2015; Wright & Stone, 1999). The four measurement levels are nominal, ordinal, interval and ratio – with each level having its own properties (Gregory, 2007; Roodt, 2013a). The *nominal* measurement level is for labelling. In the current study, for example, for group membership such as gender, female = 1 and male = 2, however, these numbers have no meaning other than to sort information into convenient categories (Gregory, 2007; Roodt, 2013a). The *ordinal* level assigns numbers to reflect the ranking order, for example, school grades, grade 9 = 1, grade 10 = 2 and grade 11 = 3. Therefore the ordinal level has the property of magnitude (moreness – although the numeric value or rank is not a direct indication of the magnitude of the difference between different values). As a result of their limited properties, the nominal and ordinal levels are often referred to as categorical data while the interval and ratio levels are called continuous measurement. The *interval* level has the properties of equal interval and magnitude (Gregory, 2007; Roodt, 2013a). This is the scale generally used in psychological tests, as opposed to the *ratio*, because this

latter scale has the properties of equal interval and magnitude as well as an absolute zero. Interval scales are preferred in psychological tests because a score of 0 does not mean absolute zero or an absence of what is being measured.

3.2.2 Measurement errors

In the theory of measurement it is acknowledged that with every measure of any attribute there is some element of error (Moerdyk, 2015). Therefore the accuracy and consistency of measurement do not necessarily imply a 100% perfect score (Moerdyk, 2015; Wright & Stone, 1999). This element of error is called the measurement error which according to Anastasi and Urbina (1997) is any factor irrelevant to the purpose of the test which affects test results. Kane (2010) cautioned that although measurement errors can be expected, they should not be present to the extent of causing misinterpretation of scores.

Measurement errors include random and systematic errors (Moerdyk, 2015; Roodt, 2013a). Figure 3.1 is an adaptation of the illustration by Moerdyk (2015, pp. 37 & 47) that has been included to explain how measurement, its scores and the variations interact.

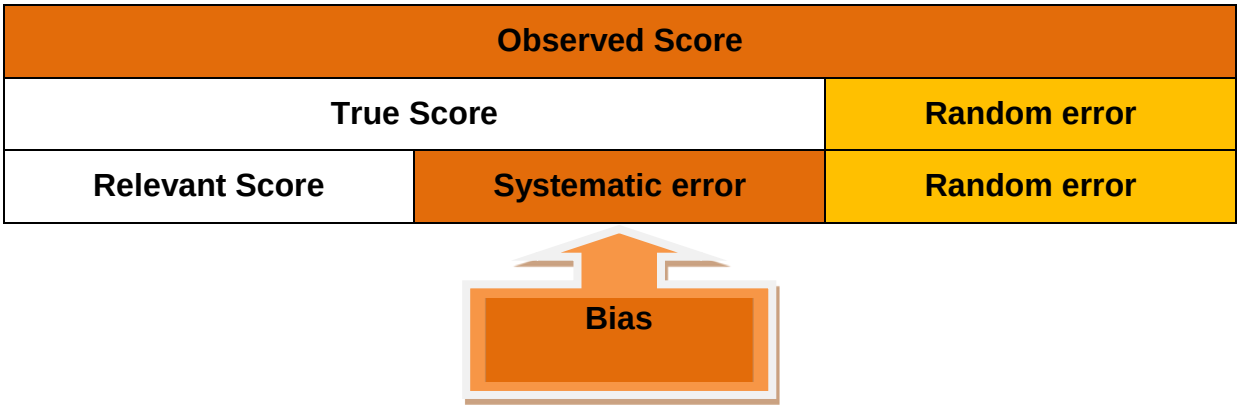


Figure 3.1. Measurement and its errors (Adapted from Moerdyk, 2015, pp. 37 & 47)

According to Moerdyk (2015), the score obtained when testing is called the observed score. From this observed score it is assumed there is some embedded random error. Random error as the name implies, occurs at random or by chance and has uncontrollable effects (Moerdyk, 2015). Taking random error into account helps to

determine what the true score is and within the true score it is alleged that there is systematic error (Moerdyk, 2015). Gregory (2007, p. 100) referred to systematic error as a “veritable ghost”. A systematic error, also known as bias, is consistent and affects one group (e.g. based on demographics) more than another (Moerdyk, 2015). More details will be furnished about bias in the later sections. As shown in figure 3.1, test results should not be taken at face value. Accepting the observed score as the real score would not be justifiable because various error measurements need to be considered.

3.2.3 Sources of measurement error

Krebs (1987) purported that all components of the test environment are potential sources of error. The important ones that have been highlighted are the test itself, test administration, test takers and test scoring (Gregory, 2007; Krebs, 1987; Moerdyk, 2015; Roodt, 2013a). The *test* itself can be a source of error because the items used or developed for the test are crucial to the quality of the test (Gregory, 2007). This would include the pool of items, the item alternatives, the format of items, the clarity or ambiguity of the items, the instructions, the scoring procedures, et cetera (Gregory, 2007; Moerdyk, 2015). For *test administration*, Roodt (2013a) emphasised that standardised assessment procedures would be affected if there are variations in instructions, assessment conditions, interpretation of instructions and scoring. Any processes that are not followed, such as time limits, conditions (i.e. temperature) in the testing room, noise levels, lighting, et cetera, can impact on the performance and scores (Gregory, 2007; Moerdyk, 2015).

Another possible source of error is in the personal characteristics of the *participants,* since factors such as their mood, language proficiency in understanding instructions, experience of testing (test wiseness and test sophistication) and response biases, could impact on the magnitude of error in measurement (Gregory, 2007; Moerdyk, 2015; Roodt, 2013a). Participants’ demographics such as gender, race, level of education, socioeconomic status, age, et cetera, can also be related to errors. The last possible further source of error is *test scoring,* which is more relevant for tests that do not use a set format for scoring and thus need some subjective judgement from the rater, although clerical or human errors may also affect even standardised

scoring (Gregory, 2007). Roodt (2013a) cautioned that any scoring should be checked for correct editing, coding and processing to ensure the integrity of the data.

Krebs (1987) asserted that with maximum control of these errors it is possible to decrease the occurrence (prevalence) of errors and ensure that the construct of interest is measured as accurately as possible. Furthermore, how the results are interpreted and used would to a certain extent curb any negative impact of measurement errors. As Kane (2010) asserted, acknowledging the existence of measurement errors provides opportunities for improvement in measurement processes and better informed decisions on measurement interpretations.

3.2.4 Measurement in practice

Having looked at the context and challenges of measurement, it is clear that psychological testing by its very nature is not a perfect science. Foxcroft and Roodt (2013a) also acknowledged this, which is why they cautioned that test results are merely one source of information, and an approximation at that. Hence any decisions made should be based on multiple and multidimensional information. Notwithstanding that caution, tests need to be continuously improved to ensure that their added value to decision making is maintained or enhanced. Smit (1996) drew attention to the critical detail necessary for the compilation of tests and the precision that needs to be maintained in the application of those tests. He regarded these as the key aspects to the success and effectiveness of measurement in terms of providing evidence and evaluation. The framework as suggested by Roodt (2013a) would entail the following three elements that anchor consistency in measurement: (1) clarifying the construct to be measured, (2) identifying the nature of the scale or measure to be used, and (3) applying a set of instructions on how measurement should be administered, scored and interpreted. For the current study, the construct to be measured by the *new items* was general nonverbal reasoning (fluid) ability, which is measured using dichotomously scored multiple-choice items. No training or preparatory studying is required beforehand and the example items given during the instructions before testing commences are considered to provide sufficient practice for answering the test items (or questions). These considerations are important as rules for the measurement process, because such standardisation of procedures

promotes perceived fairness (Kane, 2010). Fairness is further discussed in section 3.4.

3.2.5 Measurement: Concluding remarks

Even though measurement is a fairly broad concept, an attempt was made in the above discussion to highlight the important aspects that would provide an overview of measurement in order to achieve a common understanding. It is also vital to have insight into measurement and its challenges in order to appreciate the intricacies of adhering to the requirements of the EEA (Government Gazette, 1998). However, in the context of the discussion, the EEA is used to broadly integrate a number of concerns relating to measurement and to set the requirements as a quality control benchmark for the *new items* – in particular, in the multicultural and multilingual South African context.

3.3 RELIABILITY AND VALIDITY

The first requirement of the EEA is to ensure that the tests used are shown to be scientifically reliable and valid. Recognising the importance of these concepts is not new. Wright and Stone (1999) mentioned that reliability and validity had been regarded as the key concepts of measurement for almost a century. In fact, the work of Binet was also driven by the need to find an instrument that could accurately predict academic performance (Anastasi & Urbina, 1997), which shows that even in those early days, predictive validity was considered important for test development. In this section, the meaning of these concepts and how they can be measured scientifically will be discussed.

3.3.1 Reliability

Reliability refers to the consistency of measurement where similar scores are expected from the same group of people when they are retested with the same or an equivalent test (Anastasi & Urbina, 1997; Gregory, 2007; Roodt, 2013b). According to Wright and Stone (1999), reliability is the extent to which test scores are without measurement errors. The degree of consistency is illustrated by the relative

influence of the true and error scores on observed test scores calculated mathematically as a correlation coefficient (Anastasi & Urbina, 1997; Gregory, 2007). Using classical test theory[1] (CTT) principles, Harvill (1991), Huysamen (2006) and Moerdyk (2015) defined reliability as being equal to the true score divided by the observed score [R = (O – E) / O] (see figure 3.1). The reliability coefficient is the traditional quantification of consistency (Wright & Stone, 1999). Different methods can be used to determine the reliability coefficient, and these entail aspects of stability, equivalence and internal consistency (Anastasi & Urbina, 1997; Gregory, 2007; Huysamen, 2006; Roodt, 2013b; Wright & Stone, 1999). The other determinant of reliability is the standard error of measurement which is based on the variability of scores (Harvill, 1991; Huysamen, 2006). The different types of reliability are discussed below.

3.3.1.1 Types of reliability

The different types of reliability are listed as test-retest, alternate-form, split-half, inter-item and scorer reliability (Anastasi & Urbina, 1997; Huysamen, 2006; Roodt, 2013b; Wright & Stone, 1999).

(a) *Test-retest reliability* is calculated as the correlation between the scores from administering the test twice to the same group of test takers on different occasions (Anastasi & Urbina, 1997; Huysamen, 2006; Roodt, 2013b; Wright & Stone, 1999). It is also referred to as the coefficient of stability. Problems associated with this type of reliability include the question of time that is ideal for retesting after the first testing session, whether the conditions of testing (including the emotional factors and physical environment) can be duplicated for both sessions and how much impact effects such as practice and memory will have on the second session of testing (Roodt, 2013b).

(b) *For alternate-form reliability,* two equivalent forms are administered to the same group of test takers on separate occasions, thus providing the coefficient of equivalence (Anastasi & Urbina, 1997; Huysamen,

[1] Classical test theory is discussed in detail in section 3.7.1.

2006; Roodt, 2013b; Wright & Stone, 1999). The challenge with this type of reliability is ensuring that the two forms are truly equivalent in terms of the number of items, level of difficulty of the items, similar content and scoring procedures (Roodt, 2013b). Another challenge is the time invested in construction as a result of the increased number of items that have to be designed for the alternative form (Roodt, 2013b).

(c) *Split-half reliability* involves the correlation coefficient calculated when two equivalent halves are obtained from splitting one measure to ascertain internal consistency (Anastasi & Urbina, 1997; Huysamen, 2006; Roodt, 2013b; Wright & Stone, 1999). According to Roodt (2013b), the challenge with this type of reliability is in deciding how best to split the measure to ensure equivalent halves.

(d) *Inter-item reliability,* like the split-half type, is also based on ascertaining internal consistency by determining the stability of the responses to all items (Anastasi & Urbina, 1997; Huysamen, 2006; Roodt, 2013b; Wright & Stone, 1999). Depending on whether the items are dichotomous or polytomous, the Cronbach alpha or the Kuder-Richardson will be used respectively to calculate the coefficient of internal consistency (Roodt, 2013b). Internal consistency was considered to be relevant and applicable to the current study because determining whether the items are homogeneous is a vital consideration for the development of the *new items*.

(e) *Scorer reliability* is more relevant for the scores of raters by either determining the consistency between raters (inter-scorer) or between the scores of a single rater (intra-scorer) (Anastasi & Urbina, 1997; Huysamen, 2006; Roodt, 2013b; Wright & Stone, 1999). According to Roodt (2013b), because ratings have an element of subjectivity of the scorer, determining how reliable such ratings are is of vital importance, and confirming the consistency of the different ratings conveys credibility.

The reliability coefficients from the above types of reliability range from 0.0 to 1.0, where the coefficient of 0 means unreliable and the coefficient of 1 means no

measurement errors, and are therefore totally reliable (Gregory, 2007). Hence any coefficient closer to 1 would indicate a higher or stronger reliability, while a coefficient closer to 0 would indicate lower or weaker reliability. However, deducing that a test is reliable or not may not mean much, because what carries more weight in this context is what is considered acceptable or unacceptable for the practical use of tests.

3.3.1.2 Standard error of measurement

Since error is acknowledged as part of measurement, reliability can also be expressed through the standard error of measurement (Anastasi & Urbina, 1997; Huysamen, 2006; Roodt, 2013b). According to Huysamen (2006), measurement errors are not mistakes, but are accepted variations from the observed scores. The standard error of measurement (SEM) is the standard deviation associated with test scores (Gregory, 2007; Harvill, 1991). Smit (1996) described SEM as being independent of the inconsistency of the sample. It is the range of acceptable fluctuation that scores can have as a result of error (Huysamen, 2006). SEM is important as it provides a safeguard against using a single numerical score and can be used to determine the confidence intervals for interpretation of scores (Anastasi & Urbina, 1997; Gregory, 2007; Roodt, 2013b). The use of SEM provides a boundary for the variability of scores, which means any score that falls within the score bands will be acceptable. SEM has been recommended as the better option than coefficients in reporting on reliability scores (Anastasi & Urbina, 1997; Harvill, 1991).

3.3.1.3 Reliability and its meaning

What does a score of the reliability coefficient mean? At what level of reliability does one accept the test to be sufficiently consistent and thus appropriate to use? Since an element of error is expected, what degree of error is acceptable? What can be done to improve the level of reliability?

Before responding to these questions, it is necessary to highlight as a reminder that scores should be seen as estimates rather than absolutes. Nunnally and Bernstein (1994) suggested that reliability scores of 0.7 or higher should be the goal. According

to Anastasi and Urbina (1997), the acceptable range of a reliability coefficient for standardised tests is at 0.8 to 0.9, while Huysamen (1996) suggested 0.85 for decisions on individuals and 0.65 or higher for decisions on groups. However, it should be noted that the acceptability of reliability is not only about the number, but also about the decisions being made on the basis of the results of standardised tests. Smit (1996) was cautious about suggesting specific values because such values could be construed as targets rather than minimum standards. It is often recommended that decisions about the reliability of a test should not be based on one type of estimation, but should be more open to different types in order to give a broader picture of consistency (Anastasi & Urbina, 1997; Roodt, 2013b).

With regard to SEM, because the value obtained displays the degree (amount) of variance, the smaller the SEM, the better the reliability will be and the larger the SEM, the lower the reliability will be (Harvill, 1991). There are no clearly set levels of acceptable measurement error, but one way of improving reliability, proposed by Nunnally and Bernstein (1994), is increasing the number of items, because the more items included in the test, the more reliable the test is expected to be. However, this should be done within reason, because test takers would not be motivated to respond to long tests.

The other key consideration to ensure sound quality measurement is to ensure that the test measures the construct it is supposed to measure.

3.3.2 Validity

The importance of validity for testing and assessment is never really questioned. Wright and Stone (1999) confirmed that validity is the primary consideration in the evaluation of a measure. The question of whether the measurement is measuring what it set out to measure and how well it does so is often the focal point of concern for validity (Anastasi & Urbina, 1997; Roodt, 2013c). Moerdyk (2015) defined it as the ratio of the relevant score to the observed score (see figure 3.1).

Historically, the focus in understanding validity was on “positivistic conceptions of science” (Schmidt, 2006, p. 59). The technical, objective and value-free meaning of

the definition were emphasised in the conceptualisation and determination of validity, thus limiting the concept to science and separating it from practice (Schmidt, 2006). The expansion of the understanding and application of validity is in acknowledging that validity is not only about the measure being valid, but also about the deductions and interpretations drawn from the results (Cohen & Swerdlik, 2002; Gregory, 2007; Huysamen, 1996; Pan, 2009; Schmidt, 2006; Smit, 1996; Wright & Stone, 1999; Zumbo, 1999). Huysamen (1996) and Smit (1996) also acknowledged that tests are valid for specific applications as opposed to the somewhat limiting description of seeing the test (as an instrument) being valid. According to them, a definition that does not mention the purpose would give the impression that validity is a test-specific property, which is not an accurate reflection. The understanding of validity was that of a context-dependent phenomenon (Huysamen, 1996; Smit, 1996).

Borsboom, Mellenbergh, and Van Heerden (2004) presented different arguments on the conceptualisation of validity where they concluded that both the test-specific properties and test score interpretations were significant. McMillan (2000) underlined the importance of both evidence and evaluation in measurement. The test-specific properties would therefore provide the evidence, whilst the inferences made from the test would provide the evaluation part of measurement. It is necessary to ensure the appropriateness, meaningfulness and usefulness of the test itself and the inferences made from its scores (Anastasi & Urbina, 1997; Gregory, 2007; Schmidt, 2006; Zumbo, 1999). For that reason, validity should be understood from both the content- and context-dependent views.

According to Zumbo (1999), the use of tests and the decisions based on test results have to be a primary part of the validation process. Whether test validity is seen from the traditional view of technical properties or the functional view of consequences and values, what is important is that validity evaluation is a continuous process that is never completed (Anastasi & Urbina, 1997; Bond, 2003; Gregory, 2007; Zumbo, 1999). It is said that the validation process is a long-term commitment that starts when the construct is defined, even before items are written, then taken through to item analysis, and continues when the measure is in use with validation studies of different groups and contexts (Bond, 2003; Zumbo, 1999). Such a commitment ensures that tests are continuously evaluated and improved and can maintain their

usability and relevance. In determining the degree of validity, different types of investigations are conducted, which are discussed below.

3.3.2.1 Types of validity

The traditional categories of validity as Zumbo (1999) referred to them, are content, criterion-related and construct validity. One should note that these are still relevant in ascertaining validity, although not exclusively (Cohen & Swerdlik, 2002; Schmidt, 2006; Zumbo, 1999).

(a) *Content validity* is a nonstatistical type of validity that is used to determine whether the content covers a representative sample of the behaviour domain being measured (Anastasi & Urbina, 1997; Gregory, 2007). Evaluation of content validity typically involves use of a panel of subject experts to provide judgemental evidence of domain relevance (Roodt, 2013c; Zumbo, 1999). In the same category, but in no way scientific, is face validity. This type of validity is content based as well, but not validity in the technical sense, since it is concerned with whether the test “looks valid” to those who are completing it or other nonexpert stakeholders exposed to the measure (Anastasi & Urbina, 1997; Moerdyk, 2015; Roodt, 2013c).

(b) *Criterion-related validity* entails both concurrent validity and predictive validity. According to Wright and Stone (1999), this would entail comparing the test results with an external criterion that has been deemed adequate as a base of comparison. Concurrent validity refers to the extent to which the test scores correlate with scores of another criterion that can be obtained at the same time, while predictive validity refers to when the criterion scores are obtained at some future date (Moerdyk, 2015; Roodt, 2013c; Zumbo, 1999).

(c) *Construct validity* entails the extent to which the test (measure) scores accurately correlate with other measures of the same construct as defined by theory (Anastasi & Urbina, 1997; Moerdyk, 2015). This type of validity can be determined in various ways, such as investigating

correlations with other measures that test the same theoretical construct, by using exploratory and confirmatory factor analysis or by convergent and discriminant validity (Visser & Viviers, 2010; Zumbo, 1999).

The choice of the validity procedure/s to be used depends on the purpose for which decisions or inferences have to be made. One should bear in mind that validation is not a process of selecting one type of validity from the above (Zumbo, 1999), but that these various types of validity should be seen as providing an integrated picture of test validity at different stages of the construction process (Cohen & Swerdlik, 2002; Foxcroft, 2004). Although it is necessary to know the different types of validity, it is more important to understand the meaning of the results of such investigations and what is acceptable as evidence of validity.

3.3.2.2 Validity and its meaning

The definitions above highlighted the importance of the purpose for which the test is used in relation to validity. The question of what level of validity is acceptable is a difficult one to answer. According to Gregory (2007, p. 120), since presenting validity as “one single, tidy statistic” score is almost impossible, he suggested a continuum ranging from weak to strong. Zumbo (1999) had also referred to the continuum in an earlier discussion, as he proposed that the traditional view of seeing a test as valid or invalid was outdated. Anastasi and Urbina (1997) and Smit (1996) opted for the suggestion of acceptable statistical significance levels of 0.05 and 0.01. However, Huysamen (1996) and Moerdyk (2015) cautiously suggested the coefficients to be acceptable at 0.5 and as low as 0.2 for selection criteria purposes. However, it would seem that placing validity on a continuum is more applicable to the context-sensitive understanding of validity.

Another aspect of giving meaning is through validity generalisation. This is described as an extension of validity to other populations that may be different from that used to initially validate the test (Anastasi & Urbina, 1997; Moerdyk, 2015; Roodt, 2013c). Foxcroft, Paterson, Le Roux, and Herbst (2004) raised the importance of cross-cultural validity to guide test use in the multicultural context of South Africa. Similarly,

Moerdyk (2015) underscored the need for ecological validity (contextual validity), which determines whether the results of the assessment are meaningful and useful outside the setting in which they were obtained. All these are seen to be linked to studies of fairness and bias (Foxcroft et al., 2004; Moerdyk, 2015; Zumbo, 1999) because of the assurances and evidence needed to confirm that tests work similarly for different groups.

3.3.3 Reliability and validity in practice

As mentioned previously, determining reliability and validity is not about choosing one type of investigation and reporting on it. Gregory (2007) acknowledged that most tests report on multiple sources of information about these properties. It is also important to highlight the power relations between reliability and validity that have often been accepted as reliability having a limiting effect on validity (Roodt, 2013c). However, Zumbo (1999) preferred to view reliability as that essential part of validity that contributes to the reduction of measurement error. Superseding this relationship is the question of whether the properties of tests in terms of reliability and validity are also applicable across different groups. If some part of what is being measured is irrelevant to the defined construct, or prediction cannot be applied across different groups, then questions of bias and fairness crop up (Foxcroft et al., 2004; Moerdyk, 2015; Pan, 2009; Roever, 2005; Schmidt, 2006; Zumbo, 1999).

3.3.4 Reliability and validity: Concluding remarks

As discussed above, ensuring precision and accuracy in measurement has been shown to be vital to psychological testing. The significance of these properties is not confined to the tests themselves, but to the inferences and decisions made on the basis of test results (Gorin, 2007). It therefore makes sense that the EEA extended its focus to broader applications such as fairness.

3.4 FAIRNESS

A further requirement of the EEA is fairness in application. According to Moerdyk (2015), fairness involves the extent to which outcomes are used in a way that does not discriminate against particular individuals or groups. Fairness is attributed to a test that affords all test takers from different groups an equal opportunity to exhibit skills and knowledge or performance relevant to the purpose of the test (Perrone, 2006; Roever, 2005). Kriek (2001) asserted that fairness is more about a perception rather than a scientifically determined concept, as it is focused on and concerned with the impact of decisions in the social context. McMillan (2000) mentioned the absence of bias, equitable treatment and equality in outcomes as key aspects of fairness, while Pan (2009) raised the inclusion of social, ethical, legal and philosophical views as important in the discussion of fairness. Kunnan (2004) provided a test fairness framework in which he highlighted five qualities of fairness, namely validity, absence of bias, access, administration and social consequences. In his discussion, he saw fairness as an overarching concept over validity, bias and any consequences that may occur.

3.4.1 Fairness and bias

The necessity for defining fairness and bias with clear distinguishing characteristics is undisputed as the erroneous (or incorrect) interchangeable use of these concepts is seen as the source of some of the controversies surrounding and criticisms of tests (Gregory, 2007; Owen, 1992). Angoff (1993) and Kriek (2001) clarified the differences between the two by describing bias as a statistical interpretation and fairness as being based on social interpretation. Test bias is an “objective, empirical question not a matter of personal judgement” (Gregory, 2007, p. 268). Similarly, Schmidt (2006) referred to bias as a technical issue that is in the science domain, while fairness was seen as being outside the domain of science and more in the domain of values. Hence fairness is deemed to be embedded in social values of test usage, while bias is the systematic error of the items or content of the test that impact on one group in favour of the another (Gregory, 2007; R.B. Kline, 2013; Kriek, 2001; Kurnaz & Kelecioğlu, 2008; Moerdyk, 2015; Visser & Viviers, 2010).

Kriek (2001) implied that it is possible for a test with all the psychometric properties required to still have its usefulness questioned if it is perceived to be unfair. This is so because with fairness, the emphasis is placed on the application of the test results in terms of the decisions made by test users, based on the test results. With this understanding of fairness and bias, it is therefore possible for a test to be deemed biased but used fairly, just as much as a test deemed unbiased can be used unfairly (R. B. Kline, 2013; Moerdyk, 2015). According to Plake and Jones (2002), test users are supposed to protect the integrity of test scores. Visser and Viviers (2010) concurred with this view, as they noted that fairness is not necessarily controlled by the test developers, but instead that responsibility rests more on the shoulders of the test users and how they make decisions based on the test results. This explanation is in line with the value placed on the decisions of validation, which clearly shows the shift of focus from test-specific properties to the context and manner in which test results are used.

3.4.2 Approaches to fairness

Various approaches are followed to make decisions about the fairness of tests, which include value judgements and those based on the social consequences of using tests (Anastasi & Urbina, 1997; Gregory, 2007). Depending on what the societal values are, the decisions made will then be judged as fair or unfair. These are briefly discussed below.

3.4.2.1 Unqualified individualism

The unqualified individualism position entails making decisions based on which individual is the best choice in terms of predicted criterion scores rather than considering other factors such as race, gender or any other socially desirable targets (Gregory, 2007; Moerdyk, 2015). Anastasi and Urbina (1997) suggested a similar strategy for fair test use in that the sole basis of the decision is on the test performance of each individual excluding any other goals, thus providing equal opportunities to all individuals.

3.4.2.2 Quotas

Anastasi and Urbina (1997) cited the need to increase the demographic mix and representativeness, thus quotas, while Gregory (2007) described it as decisions based on ratios, which imply that not all those selected would necessarily have the highest scores. Similarly, Moerdyk (2015) referred to quotas as a preset number of decisions that are to be made in favour of specific groups of people in order to meet racial, gender and other socially desirable targets.

3.4.2.3 Qualified individualism

The argument advanced by Moerdyk (2015) in justification of qualified individualism is that if testing is acknowledged as being less than 100% accurate, then group membership should be regarded as important in decision making. Qualified individualism is portrayed as a form of affirmative action because preference in decision making favours groups that were previously disadvantaged by past inequities (Anastasi & Urbina, 1997; Moerdyk, 2015).

3.4.3 Fairness in practice

Since fairness is mainly driven by perceptions and values in the social context, the manner in which measurement errors as encompassed in the assessment process, testing situation and test administration are handled would also be significant. Providing controls and guidelines may impact on the perceived fairness of the testing experience. Plake and Jones (2002) took into account all stakeholders (test organisations, test developers, test administrators and test takers) and how they could work together to improve the use of tests. Roever (2005) made a few suggestions, which included guidelines such as treating people with respect, avoiding stereotyping, being sensitive to cross-cultural issues and showing elements of diversity in the items. Foxcroft, Roodt, and Abrahams (2013b) and Griessel et al. (2013) also wrote chapters on the legalities and ethical standards of testing and on guidelines for administering tests fairly. In these chapters, they clearly set out statutory controls of test use; best practices in testing and assessment; rights and responsibilities of both practitioners and test takers; and procedures and duties of

what needs to be done before, during and after testing. Many of these guidelines have little to do with the test itself, but are more about the use and application of it.

3.4.4 Fairness: Concluding remarks

In addition to being aware of how precision and accuracy in measurement are negatively affected by the number of errors, and this being to the benefit of testing to ensure that any errors that can be reduced, are reduced, understanding the context of testing and the precautions needed in a multicultural setting are important to show commitment to fairness. Although fairness has been shown to be socially and value based, it does involve consideration of both the content of tests and the context of testing.

3.5 BIAS

Ensuring that bias against any group is curbed is the third and last requirement of the EEA. According to Kurnaz and Kelecioğlu (2008), bias is a systematic error in the measurement process which results in differences between subgroups. These differences (bias) occur when test takers from different groups are of similar (or comparable) ability, but their performance in the test is lower for one subgroup (Geranpayeh, 2008; Perrone, 2006; Roever, 2005). Pedrajita and Talisayon (2009) defined bias as a systematic error that causes invalidity by giving one group an advantage in performance over other groups. Items are biased when they contain elements that are not relevant to the construct being measured, or at other times, it is because the content of the items is completely different from the life experiences of one group (Crane, Van Belle, & Larson, 2004; Pedrajita & Talisayon, 2009; Zumbo, 1999). This would mean the items with bias are more related to the characteristics of the group rather than to the trait of interest. Bias has also been associated with validity, or invalidity to be specific (Clauser & Mazor, 1998; Visser & Viviers, 2010). Gregory (2007) referred to differential validity as a form of bias because the scores of subgroups do not fall on the same regression line. As a result, differential validity predicts differently for different groups – hence the relation to bias.

3.5.1 Bias and other related concepts

In the same way as it was necessary to distinguish between bias and fairness, the same needs to be done for bias and equivalence. These two concepts are considered fundamental to cross-cultural comparisons (Visser & Viviers, 2010). It may be easier to say that bias and equivalence are opposites, but in cross-cultural comparisons, a specific effort should be made to separately define and establish the existence of equivalence and bias (Schaap & Vermeulen, 2008; Visser & Viviers, 2010). It has been ascertained that an equivalent item does not necessarily mean that the item is without bias (De Kock, Kanjee & Foxcroft, 2013; Meiring et al., 2005; Schaap & Vermeulen, 2008; Van de Vijver & Tanzer, 2004). However, Van de Vijver and Rothmann (2004) noted that the two have an inverse relationship because as scores are found to be without bias, they are more likely to be equivalent for the relevant subgroups.

Equivalence entails the comparison of tests conducted on different cultures (Meiring et al., 2005; Van de Vijver & Rothmann, 2004; Van de Vijver & Tanzer, 2004; Visser & Viviers, 2010). Equivalence is often described from a measurement level perspective, whereby comparison of scores can be made for different groups with similar ability in the construct (De Kock et al., 2013; Meiring et al., 2005; Schaap & Vermeulen, 2008; Van de Vijver & Tanzer, 2004). Equivalence is important and commonly used when translating and adapting tests for cross-cultural use (De Kock et al., 2013). Equivalence is a way of determining that the construct of interest is the same in meaning and properties for all groups and subgroups (Van de Vijver & Tanzer, 2004). Equivalence that ensures that the same construct is measured for all groups is referred to as construct equivalence, while measurement unit equivalence entails scales with the same unit across population groups and the equivalence that has the same measurement unit and the same origin for all population groups is called scalar or full-score equivalence (Van de Vijver & Rothmann, 2004; Van de Vijver & Tanzer, 2004; Visser & Viviers, 2010).

Other related concepts that need to be clarified are item impact, item bias, differential item functioning (DIF) and adverse impact (Moerdyk, 2015). *Item impact* entails the differing group probabilities of responding to items correctly, and such

differences are seen to be based on true group differences underlying the ability being measured (Angoff, 1993; Moerdyk, 2015). *Item bias* can be seen as the opposite of item impact because the group differences in the case of item bias are due to characteristics not relevant to the purpose of the test (Gregory, 2007; Moerdyk, 2015). *DIF* entails differences in the functioning of items from different groups that have similar abilities. An important note made by Angoff (1993) on DIF and item bias, was that these two were distinguished because items with DIF could not be assumed to have item bias. This will be discussed further in section 3.5.4. *Adverse impact* entails the inequality of the decisions made on the basis of test performance (Kriek, 2001). Kriek (2001) associated adverse impact with incidences in testing and assessment where the number of individuals selected from a specific group is lower because of noncompliance with the criterion – which may or may not be fair. According to Angoff (1993), items that discriminate as a result of the construct (item impact), as opposed to personal characteristics (item bias), are deemed more appropriate.

3.5.2 Types of bias

Owen (1992) asserted that the common goal for all methods and techniques used to detect bias is in identifying the systematic error in the estimation of the true value for a particular group of individuals. He discussed the three types of bias, namely bias in construct validity, bias in predictive validity and item bias, while Gregory (2007) categorised the types in terms of bias in content validity, bias in predictive or criterion-related validity and bias in construct validity. Kurnaz and Kelecioğlu (2008) and Zumbo (1999) pointed to internal and external bias, while Meiring et al. (2005), Van de Vijver and Rothmann (2004), Van de Vijver and Tanzer (2004) and Visser and Viviers (2010) highlighted construct bias, method bias and item bias.

Gregory (2007) and Owen (1992) suggested similar types of bias and their explanations seem to overlap. *Bias in construct validity* can be present in two situations – either indicated when the test measures different constructs for one group than another; or when the test measures the same construct for groups, but with differing degrees of accuracy. *Bias in content validity* is confirmed when the items are more difficult for one group than another, whereas their general abilities

are similar (Gregory, 2007). *Bias in criterion-related validity* is present when there is systematic error in the estimation of performance on some specified criterion measure or on predictions made (Gregory, 2007), based on test performance.

Kurnaz and Kelecioğlu (2008) differentiated between two types of bias, namely internal and external bias. They described internal bias as a consequence of the psychometric characteristics of the items, thus making it item and content dependent, while external bias is dependent on the testing conditions. The external bias would therefore be group and context dependent. Zumbo (1999) asserted that the external bias provides evidence of a different relationship between the scores and the criterion for various groups, while internal bias is about the internal item relationships. According to Angoff (1993), the internal item relationships would indicate different functioning because the item contained specific knowledge and skills relating to a certain subgroup that are not equally possessed by the other groups of test takers.

Meiring et al. (2005), Van de Vijver and Rothmann (2004), Van de Vijver and Tanzer (2004) and Visser and Viviers (2010) suggested three types of bias. *Construct bias* occurs when the construct being measured is not the same across cultures (Meiring et al., 2005; Van de Vijver & Rothmann, 2004; Visser & Viviers, 2010). According to Van de Vijver and Rothmann (2004), at times, the test does not cover all the elements to fully represent the behaviours related to or representative of the construct being measured, and this could contribute to bias. *Method bias,* as the name implies, has to do with the bias of methods and procedures (Meiring et al., 2005). This bias entails method-related issues such as sample, instrument and administration bias. *Item bias* is specific to the items in the sense that items are not the same or are not applicable across cultures (Schaap & Vermeulen, 2008). Item bias occurs when one group is less likely to answer an item correctly than another group because of some characteristic of the item (Angoff, 1993; Zumbo, 1999). The concept of item bias is often used interchangeably with differential item functioning (DIF), but Gómez-Benito, Hidalgo, and Guilera (2010) differentiated between the two as they asserted that DIF is necessary but not sufficient for item bias.

3.5.3 Differential item functioning (DIF)

Magis and De Boeck (2010) referred to DIF as an unwanted phenomenon in which an item gives an unfair advantage to some or disadvantage over other groups. Zumbo (2007, p. 223) defined DIF as the lack of "measurement invariance". Although various definitions are generally provided and accepted for DIF, they all confirm that it is present when success on any item depends on group membership, despite the groups being of similar ability (Abedalaziz, 2010; Ariffin, Idris, & Ishak, 2010; De Kock et al., 2013; Gierl, 2004; Glickman, Seal, & Eisen, 2009; Guo, Rudner, & Talento-Miller, 2006; Li & Zumbo, 2009; Tan, Xiang, Dorans, & Qu, 2010). Beaujean and Osterlind (2008) referred to DIF as the differing item parameters across subgroups.

3.5.3.1 An overview of DIF

The overview of DIF is presented through three generations by Zumbo (2007), and the discussion provided highlights the critical issues of how DIF has evolved.

First-generation DIF is characterised by the repositioning of DIF, item bias and item impact. According to De Kock et al. (2013), debates about the interchangeable use of DIF and item bias had been intensive, which gave rise to the decision to replace the statistical detection of item bias with the term "DIF". Hence for DIF analysis, statistical analytical procedures are used to indicate the differences between groups in answering items correctly, while the difference itself was explained by either item impact or item bias (De Kock et al., 2013; Zumbo, 2007).

Second-generation DIF is characterised by the development of statistical methods and computer software to determine DIF. According to Abedalaziz (2010), various methods can be used to detect DIF, but with limited guidelines provided on which method to select. The methods include examination of differences between groups, based on item difficulty, item discrimination, item characteristic curves (ICCs)[2], distribution of incorrect responses and multivariate factor structure analysis

[2] A detailed discussion of ICC is provided in section 3.6.5.

(Abedalaziz, 2010). Some of the statistical methods are Mantel-Haenszel, item response theory, simultaneous item bias test, factor analysis, item difficulty, chi-square, standardisation procedures and logistic regression (Abedalaziz, 2010; Clauser & Mazor, 1998; De Kock et al., 2013; Geranpayeh, 2008; Kurnaz & Kelecioğlu, 2008; Magis & De Boeck, 2010; Tan et al., 2010; Zumbo, 2007).

According to Osterlind and Everson (2009), DIF is recognised for its detection of systematic differences, and the subsequent contribution to test validity. DIF and validity are seen to be interrelated because investigations of DIF are essential for content bias analysis and DIF can be a barrier to valid inferences (Araffin et al., 2010; Frederickx, Tuerlinckx, De Boeck, & Magis, 2010). According to Anastasi and Urbina (1997), it is necessary for DIF to be analysed during both the initial and final stages of test development. If problem items can be detected early on and are eliminated or corrected, this will help to improve the end product. This is similar to the current study in which *new items* were analysed to determine their viability and utility. It has been suggested that items identified with DIF should be eliminated as this would most likely result in an increase in the reliability and validity of scores (Araffin et al., 2010; De Kock et al., 2013; Frederickx et al., 2010; Magis & De Boeck, 2010). However, if DIF is understood as not necessarily meaning bias, then items should not be summarily eliminated without further investigation – hence the introduction of the third generation of DIF.

The last generation of DIF focuses on explaining the reasons behind the existence of DIF. DIF can take two forms, namely judgemental (substantive) and statistical analysis (Anastasi & Urbina, 1997; De Kock et al., 2013; Gierl, 2004; Zumbo, 1999). Statistical analysis, as mentioned above, would entail statistical methods to identify items that may result in bias, while judgemental analysis involves a review by subject experts of the items with the potential to indicate bias (De Kock et al., 2013; Gierl, 2004; Zumbo, 1999). However, one should emphasise that using both forms of analysis is recommended because statistically detected DIF is not enough to fully explain bias, and substantive analysis to determine and clarify possible reasons for DIF can then be used (Yu, Lei, & Suen, 2006). Gierl (2004) recommended a

combination of the two forms as beneficial to systematically identifying and interpreting the reasons for DIF.

3.5.3.2 Types of DIF

In cases where there is no DIF, the ICCs will be the same for each of the groups, whereas if there is DIF, the ICCs will differ (Zumbo, 1999). According to Zumbo et al. (2002, p. 26), the ICC is “a curve showing the probability of a correct response as a function of the trait being measured”. A more detailed discussion on ICCs is provided in section 3.6.5. The two types of DIF are uniform and nonuniform (Kurnaz & Kelecioğlu, 2008; Van den Noortgate & De Boeck, 2005). According to Kurnaz and Kelecioğlu (2008), uniform DIF shows consistency in the difference of the probability of success between groups across ability levels. This means the one group is consistently at an advantage over the other group in responding correctly to the items. Nonuniform DIF, however, is when the difference is not consistent between groups across ability levels and the items advantage the one group only in some parts of the ability continuum (Kurnaz & Kelecioğlu, 2008; Van den Noortgate & De Boeck, 2005). Gómez-Benito et al. (2010) adopted the explanation of ICCs to explain the differences between these two types of DIF, where the ICCs for the uniform DIF do not cross, while for the nonuniform DIF, the ICCs do. In their explanation, Gómez-Benito et al. (2010) used the interaction or no interaction of the ICCs to account for the relationship between probability of success and the latent trait of interest levels. Therefore, if the difference in probability of a correct response to an item of each group is the same for all different measurement levels of the trait, then there is uniform DIF, while if the difference in probability of the correct response to an item by each group is not the same for all different measurement levels of the trait, then there is nonuniform DIF.

3.5.3.3 Studies on DIF

According to Zumbo (2007), earlier research on DIF was mainly on comparisons of groups based on gender and race. However, that has changed over the years to include other groupings for comparison. A few of these studies conducted in the last

decade are mentioned to show the variety of research in bias analyses. Previous studies investigated the presence of potential bias with respect to a number of demographic characteristics such as the following: *gender* (Abad, Colom, Rebollo, & Escorial, 2004; Abedalaziz, 2010; Ariffin et al., 2010; Crane et al., 2004; De Beer, 2004; Einarsdóttir & Rounds, 2009; Li & Zumbo, 2009; Pedrajita & Talisayon, 2009; Schnohr, Kreiner, Due, Currie, Boyce, & Diderichsen, 2007); *language* (Allalouf, 2004; Balluerka, Gorostiaga, & Gómez-Benito, 2010; Foxcroft & Aston, 2006; Glickman et al., 2009; Meiring et al., 2005; Schaap & Vermeulen, 2008); *age* (Crane et al., 2004; De Beer, 2004; Geranpayeh, 2008; Schnohr et al., 2007); *race and ethnicity* (Ariffin et al., 2010; Crane et al., 2004; De Beer, 2004; Meiring et al., 2005; Visser & Viviers, 2010; Wang, 2000); *country and geographic region* (Polikoff, May, Porter, Elliott, Goldring, Murphy, 2009; Schnohr et al., 2007; Yu et al., 2006); and *ability levels*, *years of education* and *school level* (Crane et al., 2004; De Beer, 2004; Pedrajita & Talisayon, 2009; Polikoff et al., 2009).

3.5.4 Bias and DIF in practice

As has been highlighted with all the other requirements of the EEA, restriction of bias should not be a once-off consideration, but should be included from the beginning of the test construction process. According to De Kock et al. (2013), item writers should not only be competent in the content area, but also need to be aware of cultural and other subgroup sensitivities that can impact on performance. This requires caution, even during the item writing phase. DIF and bias analysis are also recommended for new instruments during the item analysis phase because biased items can be removed or revised early on (Gómez-Benito et al., 2010; Pedrajita & Talisayon, 2009; Zumbo, 2007). Item analysis is discussed in section 3.6 below.

Crane et al. (2004) asserted that the presence of a large number of items with DIF threatens the construct validity of tests. Frederickx et al. (2010), Gómez-Benito et al. (2010), Kunnan (2004), Kurnaz and Kelecioğlu (2008), Magis and De Boeck (2010) and Van de Vijver and Tanzer (2004) held similar views. In fact, Gómez-Benito et al. (2010) maintained that the analysis of DIF accumulates evidence for validity. Identifying DIF and bias results in the elimination of items that are not relevant to the

construct, therefore promotes fairness and equity in testing (Gómez-Benito et al., 2010; Tan et al., 2010). Gómez-Benito et al. (2010) suggested that DIF and bias provide the necessary evidence for the internal structure of the instrument and its construct validity.

It has been suggested that in the same way that information on reliability and validity is provided in the test manual of the instrument, the same should be done for research results in respect of bias and equivalence (Van de Vijver & Rothmann, 2004; Van de Vijver & Tanzer, 2004). The cultural appropriateness of the instrument, including the research on reducing cultural loadings, should be clearly stated to ensure that those instruments are used in set contexts (Van de Vijver & Tanzer, 2004). According to De Beer (2004), consideration of DIF and bias analysis outcomes is necessary, but should not be the only consideration, as differential validity also has to be determined to ascertain the utility of the instrument.

3.5.5 Bias: Concluding remarks

Bias infers error, and any reduction of such an error would improve the quality of the test to measure appropriately. According to Pedrajita and Talisayon (2009), detecting and eliminating bias also means the elimination of construct irrelevant elements therefore resulting in tests that are more reliable, valid and fair. Improving these qualities of instruments in a multicultural context is worth striving for and is part of the requirements of the EEA (Government Gazette, 1998).

3.6 ITEM ANALYSIS

Item analysis is a phase in test construction that is performed to determine the best items in relation to the construct and content domain that the test aims to measure (Foxcroft, 2013). It is the process of investigating which items best represent the construct of interest and which deter adherence to psychometric requirements (De Beer, 2004; Gregory, 2007). Item analysis was one of the core aspects of this study because it covers the evaluation part of the *new items*. The information gathered

from the item analysis could then be used to select best items towards further test development processes.

Two broad categories of measurement theories can be used for item analysis, namely the test-based (classical test theories [CTT]) and items-based (item response theories [IRT]) (Fan, 1998; Hambleton & Jones, 1993; Pour & Ghafar, 2009; Thompson & Barnard, 2009). These are explained below, together with specific discussions of the item characteristic curve (ICC) and information functions.

3.6.1 Classical test theory (CTT)

Classical test theory (CTT) has dominated the test development sphere for decades (Fan, 1998; Hambleton & Jones, 1993; Wiberg, 2004). CTT is also referred to as the theory of true and error scores as Charles Spearman was responsible for its historical foundation (cited in Gregory, 2007). CTT uses the difficulty and the discrimination power of the item to evaluate the appropriateness of the item and test for a particular purpose (Foxcroft, 2013). According to Foxcroft (2013), item difficulty (*p*-value) is the proportion of individuals who select the correct response, while the discrimination value is determined by using the performance on the item in comparison with the performance on the total measure. Based on this explanation, the higher the percentage of people with correct responses, the easier the item is. The converse is also true, that is, the lower the percentage of people with correct responses, the more difficult the item is (Fan, 1998). It can therefore be concluded that the emphasis is on the correct responses of a particular group rather than the characteristics of the item – it is thus seen as a sample-dependent theory (Foxcroft, 2013; Hambleton & Jones, 1993). CTT is described as a linear model whose basic assumption is the fact that the observed score is reflective of the individual's true score plus measurement error (Hambleton & Jones, 1993; Huysamen, 2006). Gregory (2007) summarised this to mean that the theory assumes that there are factors that either contribute to the consistency or inconsistency of the measurement.

Some of the problems with CTT were noted (Fan, 1998; Hambleton & Jones, 1993; Thompson & Barnard, 2009) as follows:

- The person and item statistics are item and sample dependent respectively. For example, in a situation where the test is difficult, it may be assumed that the group has low ability levels, which may not be a true reflection.
- Since CTT is test rather than item oriented, no valid basis is provided for individual responses to items.
- Item difficulty and person abilities are not measured on a common scale – the one is on a proportion correct scale (for item difficulties), while the other is on a number correct scale (for person abilities)
- There are difficulties with CTT application for some measurements such as the equating of test scores and computerised adaptive testing.

Despite the noted problems, CTT has stood the test of time and has been applied successfully in various measurement situations (Fan, 1998; Hambleton & Jones, 1993). According to Hambleton and Jones (1993, p. 259), CTT has a number of benefits:

- Smaller sample sizes are required for item analysis.
- It has simpler mathematical analyses.
- Model parameter estimation is conceptually straightforward.
- Analyses do not require strict goodness-of-fit studies to ensure good fit of the model to the test data

3.6.2 Item response theory (IRT)

IRT was first developed in the 1960s, but only gained momentum in recent decades as it started being positioned as a theory that addresses the weakness of CTT (Baker, 2001; Fan, 1998; Foxcroft, 2013; Hambleton & Jones, 1993; Pour & Ghafar, 2009; Wiberg, 2004). IRT is also known as latent trait theory because it assumes that latent traits lead to consistent performance in tests (Gregory, 2007). The use of IRT has increased and its application in test development, equating of test scores, identification of bias are but some of the examples of its continued and increased use (Hambleton & Jones, 1993; Li & Zumbo, 2009).

IRT is a statistical theory that consists of mathematical models which function simultaneously for both the person and test item characteristic (Thompson &

Barnard, 2009; Wiberg, 2004). According to Baker (2001), the focus of the IRT is not on the test score, but on the individual item and whether or not the test taker has answered the item correctly. The basic assumptions of IRT are (1) local independence, which implies a correct response on one item is independent from responses to other items; and (2) unidimensionality, which means all the items in the test measure one construct (Fan, 1998; Winberg, 2004; Zhang, Shen, & Cannady, 2010).

Hambleton and Jones (1993, p. 259) highlighted the following benefits of IRT:

- Item statistics are independent of the groups from which they were estimated.
- Scores describe examinee proficiency that is not dependent on test difficulty.
- Test models provide a basis for matching test items to ability levels.
- Test models are not restricted to parallel tests for assessing reliability.

IRT has its own problems as well, such as the complexity experienced with the mathematical models and how to report on it, which is deemed difficult (Hambleton & Jones, 1993). De Kock et al. (2013) also commented on the complicated mathematical models used which require expensive software programs. The large sample sizes that are required for IRT can also be a practical limitation (De Kock et al., 2013; Hambleton & Jones, 1993).

3.6.3 Comparing CTT and IRT

Differences between these two theories are not discussed in any way in order to decide which one is superior to the other. Fan (1998) discredited some of the assumptions which supported the superiority of IRT over CTT, as the findings from the study indicated comparability between person and item statistics; and the degree of invariance of item statistics across samples also appeared to be similar. Wiberg (2004) compared CTT and IRT for the evaluation of the theory test in the Swedish drivers' licence test and concluded that each provided different but valuable information for evaluating items. She also noted that the estimates made were valid for both CTT and IRT.

Some of the differences may have already been noted in the preceding subsections. According to Einarsdóttir and Rounds (2009), IRT has been able to provide solutions to many problems experienced in CTT, such as modelling the interaction of the person and the individual items to the latent trait. They also asserted that IRT models can be used to determine if two groups with the same level of latent trait respond differently to an item. IRT has also been used to open doors for computerised adaptive testing (Baker, 2001; De Beer, 2004, 2006, 2007; De Kock et al., 2013), which is important for the technologically driven world of today. In adaptive testing, fewer items can be administered without compromising the reliability of the test (Baker, 2001; Einarsdóttir & Rounds, 2009).

3.6.4 Some common IRT models

Although there are various models for dichotomous and polytomous items, the focus of the current study was on dichotomous items, and only these will be discussed. According to Zhang et al. (2010), the dichotomous IRT models cater for models where there is only one correct response. They highlighted the most common models as the Three-Parameter Logistic Model (3PL), Two-Parameter Logistic Model (2PL) and One-Parameter Logistic Model (1PL). The parameters are the b-parameter (difficulty value), the a-parameter (discrimination value) and the c-parameter (pseudo-guessing level) (Baker, 2001; Gregory, 2007; Hambleton & Jones, 1993; Thompson & Barnard, 2009; Zhang et al., 2010). As the names imply, the 3PL considers all three parameters, while the 2PL has the difficulty and discrimination values, and the 1PL only considers the item difficulty. Both the 2PL and 1PL assume there is no guessing, while in the 3PL, guessing is acknowledged as a fact of life that some people may respond correctly to multiple-choice questions purely by chance (Baker, 2001).

The choice of which model to use depends on a combination of reasons such as the size of the available sample, the characteristics of the items, the choice of estimation procedures, sufficient statistics and the fit of data (Thompson & Barnard, 2010; Zhang et al., 2010). According to McCamey (2014), the Rasch model is appropriate for exploratory data analysis where items can be constructed to measure a person's ability independently from the items used. The Rasch model (or Rasch

measurement) was developed by George Rasch (1961), a Danish mathematician, with the aim of providing interval measures and monitoring the adherence of scales to scientific measurement principles (Bond & Fox, 2007; Rasch, 1979; Zhang et al., 2010). According to Zhang et al. (2010), the Rasch model has observable sufficient statistics for the model parameters and a relatively small sample requirement for parameter estimation – hence its use in this study. The Rasch model, according to Linacre (2005), is practically the same as the 1PL, except for a few conceptual differences.

The Rasch measurement model is seen as the "simple and elegant application of IRT" (Gregory, 2007, p. 110) and is based on the assumption that all items measure one common trait (unidimensionality) and equally discriminate but differ in difficulty levels (Gregory, 2007; Zhang et al., 2010). Bond and Fox (2007) highlighted unidimensionality, equal item discrimination and low inclination to guessing as fundamental to the requirements of the Rasch measurement. According to Bond and Fox (2007), the Rasch measurement can be used to evaluate possible DIF), based on the responses of the different groups to specific items and groups of items. Furthermore, using the Rasch measurement provides evidence of construct-related validity (Ding, 2014).

3.6.5 The item characteristic curve (ICC)

According to Baker (2001), each item has its own item characteristic curve (ICC).The ICC is an *S*-shaped graphical illustration (see figure 3.2) of the function, depicting the probability of a correct response and the position of the test taker in the underlying trait being measured by the test (Gregory, 2007; Van den Noortgate & De Boeck, 2005). The probability of a correct response thus depends on the item characteristics and the ability of the person. Figure 3.2 indicates the latent trait (ability) on the X-axis and the probability of answering the item correctly on the Y-axis.

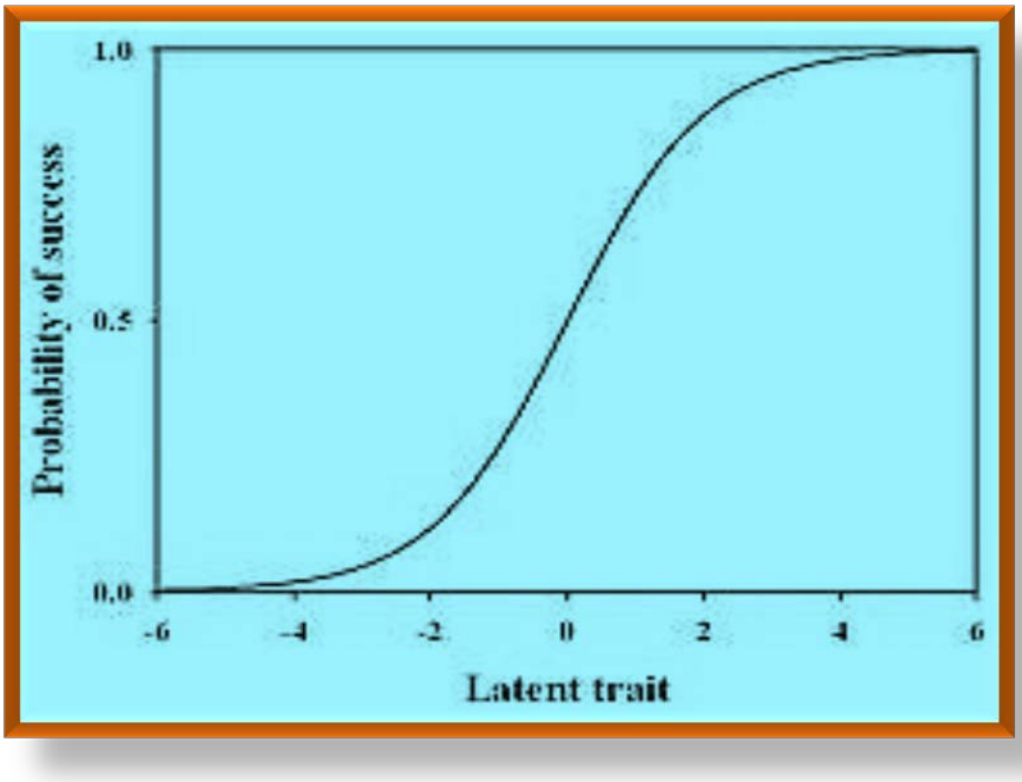


Figure 3.2. Example of the ICC[3]

As a mathematical equation, the ICC is known as the item response function (IRF) and has estimate parameters that determine the shape of the curve for the ICC, namely the *a, b* and *c* parameters (Baker, 2001; Gregory, 2007; Hambleton & Jones, 1993). Zumbo (1999) noted the importance of the ICC as it provides information about all the aspects of an item. The *b*-parameter (difficulty value) represents the level of difficulty of the item and provides the location index for the curve. The *a*-parameter (discrimination value) represents the differentiation level of the item and provides the slope for the curve. The *c*-parameter (pseudo-guessing level) represents the probability that a person with little of the latent trait will answer the item correctly. According to Baker (2001), the slope of the curve (*a* = discrimination level) is the steepest at the ability level corresponding with the item difficulty (*b*-value). Baker (2001) further noted that in the 1PL, the slope of the curve stays the same, and only the location (levels of difficulty) of the items differs. Baker (2001)

[3]Adapted from http://www.rehab-scales.org/rasch-measurement-model.html.

asserted that ICCs provide the basic building blocks of IRT, which are ideal for test design and construction.

3.6.6 Information functions

As discussed above, the IRF (or ICC) represents the probability of success associated with a particular latent trait. It is this information that contributes to what is known about each of the items, which determines the precision of measurement (Baker, 2001; Wright & Stone, 1999). This means each item can provide information across the range of levels measured, and such information can be used to describe the capability of each item to differentiate between people at or close to a particular ability level (Baker, 2001; Gregory, 2007). The item information function can be mathematically derived by converting and collating the IRF of each item and can be graphically illustrated (Gregory, 2007).

According to Baker (2001, p. 109), “test information at a given ability level is simply the sum of the item informations at that level”. It portrays how precise the test is in estimating ability over a whole range of ability levels (Baker, 2001; Gregory, 2007; Hambleton & Jones, 1993; Wright & Stone, 1999).The test information functions can be used for the management of reliability and error of measurement (Anastasi & Urbina, 1997; Baker, 2001; Gregory, 2007; Thompson & Barnard, 2009). Hambleton and Jones (1993) argued that the more information furnished by a test at a particular ability level, the lower the magnitude (degree) of errors associated with the estimation at that level will be. Baker (2001) cautioned that the interpretation of a test information function is dependent on the reciprocal relationship between the amount of information and the variability of the ability estimates.

3.6.7 Item analysis in practice

Item analysis is critical for identifying items that are of good quality or those that may need to be eliminated (Zumbo, 1999). According to Gregory (2007), it is usually expected that the original item pool should be almost double the final required number of items, to ensure that there are enough items from which to choose the best ones. The final set of items used is often based on information from CTT, IRT

and DIF analysis. Zumbo (1999) asserted that DIF is important as an item analysis methodology. Depending on which measurement theory and models are used for statistical analysis, the best items can be identified by determining the item difficulty index (proportion of people who answer the item correctly), item discrimination index (the difference between the proportion of people who arrived at the answer correctly and the proportion of those who responded incorrectly), item reliability index (consistency of the item), item validity index (usefulness of item) and the item characteristic curve (graphical display of the relationship between the probability of a correct response and the person's position on the construct being measured) (Anastasi & Urbina, 1997; Foxcroft, 2013; Gregory, 2007; Hambleton & Jones, 1993).

3.6.8 Item analysis: Concluding remarks

Item analysis puts into action all the important aspects of identifying and ensuring that the items are of a good quality in terms of reliability, validity and (no) bias. It focuses on strengthening the internal structure of the test by ensuring that the items meet the required standard in terms of measuring accurately, precisely and with limited measurement errors. As such, it is a vital phase in the test development process.

3.7 MULTICULTURAL ASSESSMENT

According to Osterlind and Everson (2013), there is interconnectedness in all these concepts of reliability, validity, fairness and bias or DIF. Reliability is assumed in valid measurements, fairness is expected in valid inferences and absence of bias is expected in fair measurements. These are all essential elements for the success of multicultural testing and assessment as they provide the necessary boundaries for balancing the performance of groups. This was confirmed in the qualitative study by Paterson and Uys (2005), where participants highlighted predictive validity, cross-cultural fairness, relevance and reliability as significant standards to achieve for new instruments. Adherence to the requirements of the EEA makes it possible to have equitable testing, which is essential in the multicultural South African context.

3.8 CHAPTER SUMMARY

This chapter highlighted the difficulty if not the impossibility of finding a test that can be 100% accurate, reliable, valid, fair and unbiased in its measurement. This was credited to the challenges of measurement in the psychological context, which is not based on visible constructs, and further exacerbated by the complexities and challenges associated with assessment in a multicultural and multilingual context such as that of South Africa. However, there are acceptable boundaries and requirements that have been set as benchmarks, which need to be strived for. These were discussed in the context of the EEA requirements of reliability, validity, fairness and bias. Item analysis was also elucidated as the process that is used to evaluate the items to ensure that they meet the requirements of the EEA. Adherence to the requirements enhances the acceptability of tests in a multicultural context such as that in South Africa, as well as in the international context, in terms of meeting international standards and requirements. As per the quotation at the beginning of the chapter, where *a sure test* is questioned, chapter 3 did not provide a sure test, but a close enough set of benchmarks to help ensure a measure that is fair, consistent, accurate and without bias.

Having conceptualised what multicultural cognitive assessment is all about and reviewed the benchmarks of assessment, chapter 4 deals with the research design and methodology used in the current research study.

CHAPTER 4

RESEARCH DESIGN AND METHODOLOGY

If you can't describe what you are doing as a process, you don't know what you are doing – W. Edward Deming

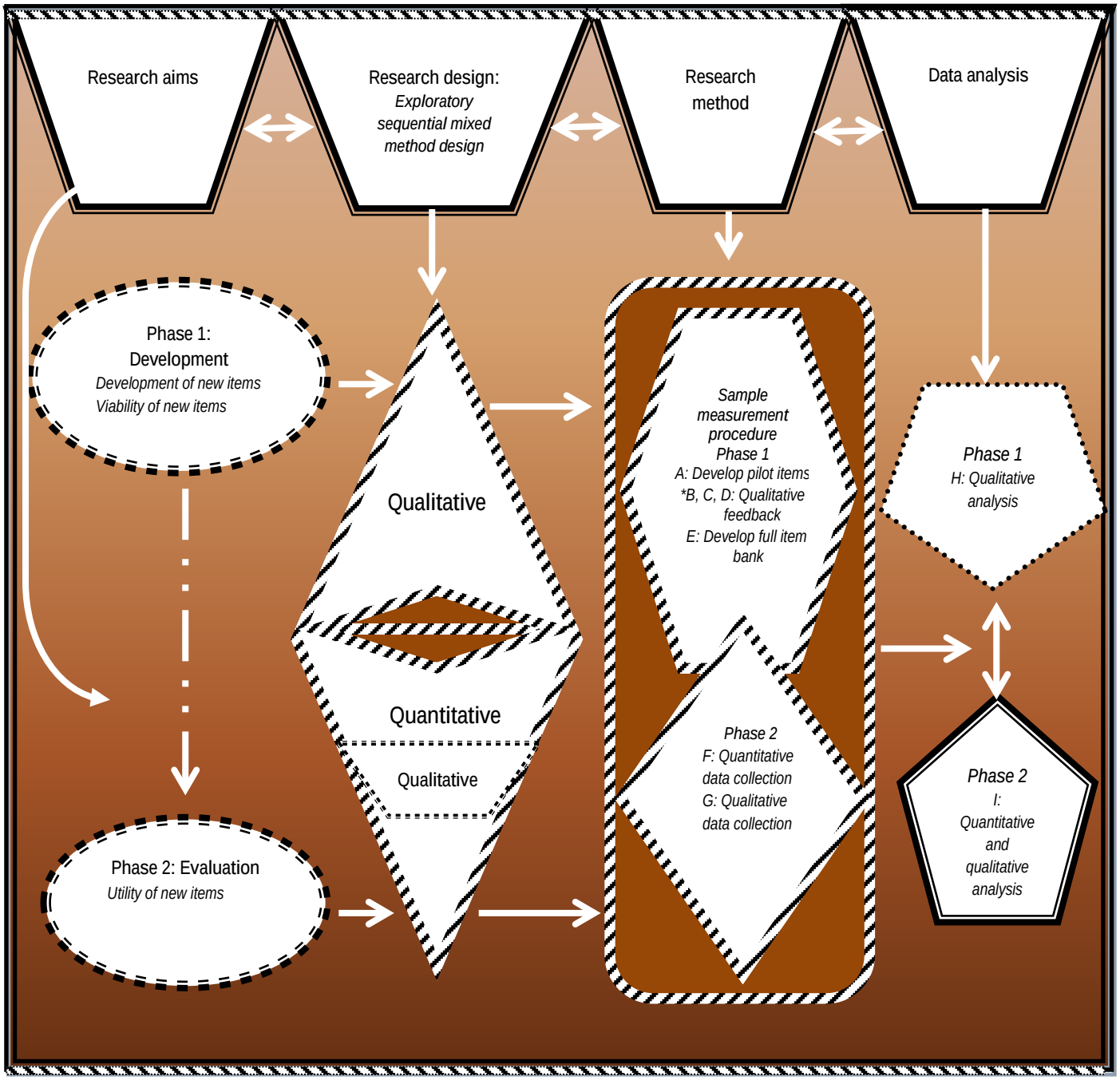
4.1 INTRODUCTION

Reliability and validity are vital concepts in relation to research results being shown to be credible and authentic. According to Christensen (2001) and Mouton (2001), these concepts are the foundation of what makes research scientific – hence the importance of following specific processes in order to adhere to scientific principles. The scientific research principles comprise clearly defined variables; processes that can be replicated; and set boundaries for control (Christensen, 2001; Mouton, 2001). The process followed to ensure adherence to these scientific principles is discussed in the research design and methodology chapter for this study.

Figure 4.1 below provides an overview of the research project and outlines the different phases and processes of the research. A brief explanation of figure 4.1 is provided. Also included in the chapter is a more detailed discussion of the research aims, the research design, the sampling strategies used, a description of the research participants, the measures used to collect data, the research procedure followed and the techniques used to analyse the data. Lastly, the research questions and central hypothesis will be presented.

4.2 FOCUS OF THE RESEARCH PROJECT

The flow of and sequence in the research project is illustrated in figure 4.1 below. The main steps of the research project regarding methodology, namely the research aims, research design, research method and data analysis (Salkind, 2014) are presented at the top with two-sided arrows connecting them to indicate their interrelationship. Any decisions made in each of these steps should be aligned throughout to ensure that the research aims are addressed.



* B = qualitative - cultural expert feedback; C = qualitative colour-blind feedback; D = qualitative pilot school group feedback;

Figure 4.1. Overview of the research project

The research aims are presented on the basis of the phases of the study in which both the development and evaluation of the *new items* are addressed. The research aims linked to phase 1 were to develop *new items* and evaluate their viability; while the purpose of phase 2 was to evaluate the utility of the *new items* for measuring cognitive ability. Congruent with these aims of the research project, the research

design reflects how the two phases were addressed, where qualitative and quantitative approaches were incorporated – hence the choice of a mixed method research design, namely the exploratory sequential mixed method research design (Caruth, 2013; Creswell et al., 2011; De Lisle, 2011; Venkatesh, Brown, & Bala, 2013). Aligned to this research design, a sequentially based research method was followed, specifically entailing sampling, instruments and procedural decisions. The research method elements revolved around the aims of phase 1 [represented by A, B, C, D and E in figure 4.1] and phase 2 [represented by F and G in figure 4.1], after which the analysis of both qualitative and quantitative data was embarked on [depicted by H and I in figure 4.1]. All the elements in figure 4.1 were interconnected to ensure that general scientific principles were adhered to.

It is important that this chapter, specifically the flow and sequence illustrated in figure 4.1 should not be looked at in isolation – but has to be considered in relation to the magic circle of Trafford and Leshem (2008) as depicted in figure 1.4. According to Trafford and Leshem (2008), the research issue and research design influence one another and provide the evidence for the internal empirical consistency of the research. The actions and decisions discussed below should therefore be viewed as being guided by the research issue.

4.3 RESEARCH AIMS

As indicated above, the purpose of the research was to develop a new format of items and evaluate them for viability in terms of perceived cultural fairness as well as assessing the psychometric properties of the *new items*. African art and cultural artefacts were used as inspiration in the development of the *new items* to assess cognitive ability (fluid ability). Following the development of the *new items,* other aspects of importance in the project were to evaluate the viability of the *new items* for measuring cognitive ability in terms of the perceived culture fairness of the new items as well as assessing the psychometric properties or utility of the items. Since the development of the *new items* represents the contribution of the study, chapter 5 provides a detailed account of this.

4.3.1 Research questions

The following research questions were posed to facilitate the empirical research in this study:

- With regard to phase 1, how are the new nonverbal figural reasoning ability items constructed from inspirations of African art and cultural artefacts? This research question entailed the following subquestions:
  - ✓ How is nonverbal figural (*gf*) reasoning ability defined?
  - ✓ How are African art and cultural artefacts defined and explored as inspiration?
  - ✓ How are the inspirations from African art and cultural artefacts, combined with nonverbal figural reasoning ability assessment principles?
- With regard to phase 1 of the research, how can the viability of the *new items* be determined?
  - ✓ How well does the appearance of the items represent the African art and cultural artefacts they are inspired by?
  - ✓ Are the items appropriate for use by colour-blind persons?
  - ✓ Can the items be easily administered?
  - ✓ How is the final pool of items constructed for the full psychometric evaluation process?
- With regard to phase 2 of the research, how can the utility of the *new items* be assessed?
  - ✓ Do the results of the *new items* show acceptable levels of reliability?
  - ✓ Is there a statistically significant relationship between the total score obtained for the *new items* and another measure of general nonverbal figural reasoning using more traditional item formats (construct validity)?
  - ✓ Does the qualitative feedback from participants provide supportive evidence in terms of the face validity of the *new items*?

4.3.2 Research hypothesis

Hypothesis testing applies to quantitative research and forms part of decision making, where the design of the study and the data collected are used to test the validity of the stated hypothesis (Christensen, 2001). The central hypothesis for the study was as follows:

H_1: There is a statistically significant relationship between the total scores obtained on the *new item* formats and the learning potential scores obtained on a more traditional nonverbal figural measure of general nonverbal figural reasoning.

4.4 RESEARCH DESIGN

The research design provides the necessary path from the research questions to the research results in a manner that ensures maximum validity and minimum error in the study (Babbie & Mouton, 2010; Durrheim, 2006). The research design that catered for both the phase of the development of the *new items* (phase 1) as well as the phase of evaluation of the *new items* (phase 2) was the exploratory sequential mixed method research design with quantitative embedded in phase 2 (Caruth, 2013; Creswell et al., 2011; De Lisle, 2011; Venkatesh et al., 2013). The choice of this research design is supported by various researchers, who highlighted it as appropriate for instrument development and validation (Creswell, 2012; Creswell, Plano Clark, Gutmann, & Hanson, 2003; De Lisle, 2011; Doyle, Brady, & Byrne, 2009; Klingner & Boardman, 2011; Ngulube, Mokwatlo, & Ndwandwe, 2009; Venkatesh et al., 2013).

According to Creswell (2012) and Onwuegbuzie and Johnson (2006), this research design is based on building from one phase of the study (development of the *new items*) to another phase of the study (evaluation of the *new items*). As illustrated in figure 4.1, the qualitative and quantitative approaches are combined, each addressing different but interlinked phases of the study at different times and with different aims (Ivankova, Creswell, & Stick, 2006; Johnson & Onwuegbuzie, 2004). The exploration of the use of African art and cultural artefacts for the development of

the *new items* was based on a qualitative approach, as was the data collection used to determine the viability of the items in terms of their cultural fairness. The quantitative approach was then used to collect data used to determine the utility of the items in terms of psychometric properties. Parallel to this quantitative stage, open-ended questions were included in the quantitative answer sheet to obtain further qualitative data. An insert of an illustration depicting the qualitative focus in the quantitative approach box illustrates this in figure 4.1.

The choice of the research design was also aligned and compatible with the research paradigm (critical realism) of the study (Moerdyk, 2015; Venkatesh et al., 2013). Moerdyk (2015) acknowledged the importance of adopting the critical realism paradigm in psychological assessment as it recognises the reality of different life and socialisation experiences. According to Venkatesh et al. (2013), the critical realism paradigm allows for a combination of a variety of methods to collect and interpret data. The use of qualitative and quantitative methods was thus deemed appropriate for this study.

4.4.1 Qualitative and quantitative research designs

As significant approaches of equal status in the present study, specific designs were chosen for the qualitative and quantitative parts. Content analysis was used to identify the African art and cultural artefacts to utilise in the development of *new items* and determining – based on feedback from various subgroups – whether the items were perceived as African. According to Leedy and Ormrod (2010, p. 144), content analysis is ideal for mixed method studies and entails “a detailed and systematic examination of the contents of a particular body of material for the purpose of identifying patterns, themes, or biases”. Content analysis is not limited to text, but includes analysis of pictorial material, paintings, imagery, et cetera. (Stepchenkova & Zhan, 2012), in which patterns, themes and trends are identified and categorised (Leedy & Ormrod, 2010).

In this study, photos of art objects, traditional dresses and beadwork, flea market objects, and so on, were used as material for content analysis in order to develop the

new items [represented by A in figure 4.1]. The process followed for the development of the *new items* is discussed in chapter 5. Content analysis of feedback obtained from cultural experts [B], a colour-blind individual [C] and pilot study group [D] contributed to understanding the perceived cultural fairness, suitability and preliminary face validity and practical utility of the *new items*. Content analysis was also used to identify themes from the answers to open-ended questions asked to determine the participants' perceptions of the *new items* [H & I] – thus providing further face validity evidence.

The cross-sectional survey research design was used to conduct the second part of the study, namely to evaluate the *new items* with which data was collected [F & G], while statistical analysis of the quantitative data was used to evaluate the psychometric properties of the *new items* [I]. According to Leedy and Ormrod (2010) and Salkind (2014), the cross-sectional survey design is a descriptive research design that is used to collect data at one point in time. Cohen, Manion, and Morrison (2007) referred to cross-sectional studies as providing retrospective or prospective snapshots of a population at a given time. They further indicated that this research design makes it possible to compare different groups and variables (Cohen et al., 2007). According to Levin (2006) and Salkind (2014), the cross-sectional survey research design is simple and inexpensive, and it enables one to collect data quickly, identify correlations and develop hypotheses. Olckers (2013) also recommended the cross-sectional survey design for use in descriptive studies because of its cost effectiveness and time-saving advantage. The challenges of this design that have been highlighted are the limitations with regard to comparing groups (Levin, 2006; Salkind, 2014) and the fact that the design is considered static in its approach and therefore ineffective for change studies (Cohen et al., 2007).

4.4.2 Benefits and challenges of mixed method research designs

As with any of the approaches, there are benefits and challenges to their use. One of the benefits of mixed methods is that the design enables both the depth of qualitative research and the breadth of quantitative research to be used in one research project, thus ensuring complementary information (Teddlie & Yu, 2007; Venkatesh et al.,

2013). By acquiring complementary information, mixed methods can bridge the gap between the exploratory and explanatory formats of the two research approaches (Cronholm & Hjalmarsson, 2011; Ngulube et al., 2009), thus resulting in all-encompassing solutions to address research problems (Heyvaert, Maes, & Onghena, 2011). De Lisle (2011) acknowledged the in-depth discoveries and generalisation benefits of mixed methods. Similarly, Caruth (2013) and Venkatesh et al. (2013) cited the benefits of mixed methods as better insights and detailed exploits with possibilities of generalisation.

The challenges associated with the mixed method research designs are basically associated with how to use the strengths of both approaches optimally while minimising their weaknesses (Onwuegbuzie & Johnson, 2006). The practical challenges include the fact that it takes longer to conduct the study and the logistics involved (Creswell et al., 2003; Ivankova et al., 2006). Also, researchers conducting mixed method research should be knowledgeable in both the qualitative and the quantitative approaches (Cameron, 2011; Ivankova et al., 2006). A concern voiced by Klingner and Boardman (2011) related to the extent to which the research data is integrated. Similar benefits and challenges were experienced in the current study.

4.5 RESEARCH METHOD

In the research design, it was acknowledged that the research method would entail collecting qualitative data as an initial investigation in relation to the development phase of the study; then, based on the results of this initial investigation, quantitative and qualitative data collection and analysis would be used for further investigations of the utility of the items (Caruth, 2013; Ivankova et al., 2006; Ngulube et al., 2009; Onwuegbuzie & Combs, 2011; Onwuegbuzie & Johnson, 2006; Venkatesh et al., 2013). In order to elaborate on the research method, sampling, data collection methods[4] and procedures are discussed below.

[4] The broader heading of "data collection methods" has been used to encompass all the different types of data collection rather than the more traditional subsection of "measuring instruments".

4.5.1 Sampling

According to Salkind (2014), the sampling strategies that are used can have a positive or negative impact on the scientific merit of the overall research. Hence decisions on how the research participants are selected are of crucial importance in ensuring the quality of the research (Cohen et al., 2007; Onwuegbuzie & Collins, 2007; Salkind, 2014).

According to Teddlie and Yu (2007), the sampling strategies for mixed method designs should take into consideration both the qualitative and quantitative methods of the research. This would imply a combination of strategies such as the use of probability and nonprobability sampling (Leedy & Ormrod, 2010; Salkind, 2014). Probability sampling strategies are founded on randomness, where members of the target population have an equal chance of being selected for the sample, while in nonprobability sampling strategies, this is not the case (Cohen et al., 2007; Leedy & Ormrod, 2010; Salkind, 2014). In the current study, the researcher decided to use nonprobability sampling methods for both phases of the study. Although this might not be the ideal for test development where representative samples are important to ensure the possibility of generalisation (P. Kline, 2013), previous research studies that used mixed methods indicated a majority of nonrandom sampling (Hultsch, MacDonald, Hunter, Maitland, & Dixon, 2002; Onwuegbuzie & Collins, 2007; Yang, Wang, & Su, 2006).

The sampling design used for this study was the sequential mixed method sampling design using parallel samples (Onwuegbuzie & Collins, 2007; Teddlie & Yu, 2007) because the research participants who took part in in phase 1 [A, B, C & D] were a different group from the research participants who were involved in phase 2 [F & G].

4.5.1.1 Samples for phase 1: Development of the new items

In phase 1, two sets of *new items* were developed, and for ease of reference, these are referred to as the first draft and second draft of *new items*. Both purposive and convenience sampling were used for selecting the participants who helped evaluate

the viability of the *new items* in three different ways, namely checking the cultural appropriateness of the content, checking for a possible negative impact of colour and preliminary face validity and checking for practical procedural elements such as instructions and response formats.

For the first draft of items, six African participants volunteered to participate in the first viewing and evaluation of the *new items*. This group – all African – comprised of two bead jewellery designers, a lecturer and three managers in human resources and finance positions respectively. This conveniently selected group provided their first impression views based on general knowledge and opinions. It was considered important to involve African individuals to obtain face validity evidence from such people in particular, who would be able to recognise the core African elements that were being captured by the researcher.

Purposive sampling was used to select a senior academic in the African Languages Department in a higher education institution. He was approached specifically for his expertise on culture in order to confirm, give advice on and critique the items in terms of their appropriateness as representations of recognisable African aspects in the colours and symbols used. Minor changes were made based on the feedback obtained regarding the cultural appropriateness and Africanness of the items.

Purposive sampling was again used to select a participant to evaluate and give feedback on the possible impact of colour blindness and whether colour-blind individuals could appropriately and correctly respond to the items. The participant who provided this feedback has been classified as totally colour-blind and could therefore provide the required feedback.

For the second draft of items, 29 African school learners ranging between 13 and 14 years of age and an education level between Grade 8 and 9, who were involved in a community project coordinated at their school, were selected as a convenience sample. They were chosen to provide provisional face validity feedback on the *new items* and to pilot whether the instructions and response method and format used for the administration of the items were clear and understandable.

4.5.1.2 Sample for phase 2: Evaluation of new items

For phase 2, convenience sampling was used to select the participants for the evaluation of the *new items* (item analysis) with regard to their psychometric properties (based on quantitative statistical analysis) and their face validity (based on qualitative content analysis). The participants were members of a group who were undergoing sponsored basic career-related training and guidance as part of a youth support and development programme. The programme entailed training for personal development, life skills, skills development, practical work, soft skills, et cetera. A group of 946 individuals participated in this part of the study. According to Hambleton and Jones (1993), sample sizes of 200 to 500 would be ideal for the classical test theory (CTT), while 500 or more participants would be acceptable for item response theory (IRT) sample sizes. Similarly, Foxcroft (2013) and P. Kline (2013) also recommended a sample size of 400 to 500 participants from the target population as adequate for item analysis. Some South African studies on test development that used the Rasch model had sample sizes ranging from 474 to 527 participants (De Klerk, Nel, Hill & Koekemoer, 2013; Maree, Maree & Collins, 2008). The sample as tabulated in table 4.1 is described by age, gender, home language, educational level and province.

As indicated in table 4.1, the sample could be considered to be homogenous in terms of educational level as most of the participants had a Grade 12 level of education/qualification. There was also a good balance between male and female representation. However, the convenient sampling method used limited the prospects of generalisation.

Table 4.1

Description of Sample (N = 946)

Variables	Description				
Age	Ranges between 18 and 36				
Gender	Female	Male	Missing		
	50% (473)	49% (468)	1% (5)		
Home language	Xhosa	Afrikaans	Zulu	Other	Missing
	38% (362)	16% (150)	11% (107)	34% (319)	1%(8)
Education	Grade 12	Grade 11	Grade 10	Other	Missing
	68% (645)	22% (214)	6% (52)	3% (27)	1%(8)
Province	Gauteng	Western Cape	Eastern Cape	Free State	Missing
	40% (375)	31% (298)	19% (178)	9.5% (90)	0.5%(5)

4.5.2 Data collection methods

In this section, the focus is on the various data collection methods used for the two phases of the study.

4.5.2.1 Data collection for phase 1: Development of the new items

Qualitative information was used for the development phase of the *new items*. Photographs of flea market items, paintings, clothing material, beadwork, decorations, et cetera, were collected as part of this phase. Onwuegbuzie, Leech, and Collins (2010) acknowledged the importance of the use of photographs in data collection and described it as innovative. Based on the photographs, patterns, colours and various symbols were identified and extracted for use in the *new items*. Forty items (first draft) were developed for the first draft to explore the viability of the symbols and colours used for the *new items* in terms of their cultural appropriateness. The items and request for feedback were sent to the eight participants via email – six responded and providing feedback.

A separate individual interview session was arranged with an African cultural expert. The purpose of the study and his required contribution were explained to him. Using a one-on-one interview, he was shown the 40 items (first draft) on a PowerPoint slideshow. He was then requested to consider each of the items and give feedback on its appropriateness in representing African elements in terms of the colours, symbols and/or patterns used. Based on his knowledge of the different cultural groups, and their art and cultural artefacts, the discussion moved to which cultural groups or artefacts the items represented in addition to and in correspondence with what the researcher had had in mind when the items had been developed.

Another source of information for this phase was open-ended questions on the answer sheet for the second draft of items that were administered to a group of 29 African high school pupils. The questions focused on their explanations of the questions, colours and item type they liked most and least. As a last question, they were asked to share any additional comments they might have about the questions.

The views and comments on the possible effect of using colour in the stimulus material were obtained from a male participant classified as totally colour-blind. This participant was shown items of different colours and the focus was on whether he could see the items and patterns clearly and find the correct answers without the colours that were used negatively affecting his choices.

It was essential to obtain all the data collected in this phase before finalising the larger group of 200 *new items* that were developed for administration during phase 2 of the study. The information obtained from the school sample was used to improve on the instructions in preparation for the administration of the final set of *new items* during phase 2.

4.5.2.2 Data collection for phase 2: Evaluation of the new items

Both qualitative and quantitative data collections methods were used during this phase, and these are discussed below.

a New items

A total of 200 new nonverbal figural format items were developed using African prints, art, decorations and beadwork patterns as the inspiration. The *new items* were developed to assess general cognitive ability ("*gf*" or fluid ability). The items were structured in such a way that figures following certain patterns or sequences were presented, and for each item, there was a missing figure in the patterns represented by a question mark. The question mark had to be replaced with the correct figure to complete the pattern (Anastasi & Urbina, 1997; De Beer, 2000, 2005, 2007, 2010; Gregory, 2007; Penrose & Raven, 1936). Four alternatives were provided from which the participant could choose the correct answer. The item formats for the *new items* included six different item types. Although included in these there were traditional item formats designed to take the rectangular (blocks) format, new item formats were also developed that used the overall shape of a triangle and a circle or wheel. Another variation in the design of the items was that the positioning of the missing element or the question mark was varied. The detailed description of the process and final items are discussed in chapter 5.

b The Learning Potential Computerised Adaptive Test (LPCAT)

The LPCAT was developed as an alternative to standard cognitive assessment, and was aimed at addressing some of the challenges of psychological testing and assessment such as fairness, item bias and reducing the duration of testing time typically associated with learning potential measures (De Beer, 2005; 2006). The LPCAT is a dynamic test which uses computerised adaptive testing (CAT) in a test-train-retest approach to measure learning potential with nonverbal figural reasoning test items based on fluid ability (De Beer, 2005; 2006; 2010). The item formats used are viewed as culture fair, with no language reliance, and they include figure series, figure analogies and pattern completion as item types (De Beer, 2005; 2007; 2010). According to De Beer (2005), the nonverbal figural item format was meant to counter the effect of language proficiency in test scores of standard cognitive assessments. The coefficient alpha internal consistency reliability scores of the LPCAT range from 0.925 to 0.987 for different groups (De Beer, 2005). Validity scores for construct and predictive validity with a group of students in a bridging programme were 0.661 and 0.313 to 0.525 respectively (De Beer, 2006; 2010).

The *new items* and the LPCAT were computer administered. There were other additional instruments for personality, career-related interest and motivation purposes which were also administered for the benefit of the participants so that they could receive some feedback for personal development and career guidance purposes. Although these instruments did not form part of the study, they had to be considered and planned for in the testing schedule.

4.5.3 Research procedure

The research procedure entails the practical arrangements and processes followed to conduct a study (Salkind, 2014). The procedural decisions for the current study were guided by the research design of exploratory sequential mixed methods (Ivankova et al., 2006; Johnson & Onwuegbuzie, 2004). These decisions are discussed below, along with the ethical considerations and procedures based on the phases of the study.

4.5.3.1 Ethical considerations

Permission was granted to conduct the study as part of the registration process as a doctoral student of the academic institution. Approval for the research was also obtained from the institutions where the data was collected for phase 1 (individuals concerned and the school administration and management) and phase 2 (government departments responsible for the learnership programme). A written consent form informing the participants about the purpose of the study, the voluntary nature of their participation, and the option of withdrawing at any time or stage of the research process without any negative consequences, was distributed before testing commenced. The consent form also explained that the results would be used for research purposes and all the participants' information would be treated confidentially.

Since standardised instruments (measuring personality, career preference, learning potential and motivation) were also used during the research, the participants were

also informed that they would receive feedback on these standardised and validated measures for personal development purposes.

4.5.3.2 Decisions on priority of the methods

According to Ivankova et al. (2006), priority refers to the emphasis given either to qualitative or quantitative or both methods in the study in terms of dominant or equal status. For an exploratory sequential mixed method study, a decision has to be made on whether the dominant status is given to the qualitative method, or to the quantitative method, or whether to give both methods equal status (Doyle et al., 2009; Ivankova et al., 2006; Onwuegbuzie & Johnson, 2004). Doyle et al. (2009) recommended a quantitative dominant status for instrument development, while Ivankova et al. (2006) argued that the decision should depend on the objectives of the study and what the researcher chooses to emphasise.

For the current study, the researcher opted for the qualitative and quantitative methods of the study to be given equal status. The decision was influenced by the aims of the study, namely to develop *new items* inspired by African art and cultural artefacts and to evaluate the *new items* for viability (qualitative focus) and utility (quantitative focus). The qualitative information collected in both phases of the study for the development and evaluation of the *new items* was highly significant in contributing to addressing the aims of the study. Similarly, the quantitative data collected was equally important in providing information on the quality and measurement properties of the *new items*. Neither of the methods held dominant status over the other, that is, the qualitative and quantitative methods were accorded equal status (Onwuegbuzie & Combs, 2011).

4.5.3.3 Decisions on the sequence of implementation of the methods

Implementation refers to whether the phases of the study were conducted concurrently or sequentially (Ivankova et al., 2006; Johnson & Onwuegbuzie, 2004). As indicated previously, the phases of the study were conducted chronologically, starting with qualitative data collection and analysis, followed by both quantitative

and qualitative data collection and analysis (Creswell, 2012; Onwuegbuzie & Combs, 2011; Venkatesh et al., 2013). Phase 1 involved the development of the *new items* and was started by exploring the African art and cultural artefacts in order to gain a better understanding of what the items should entail. The *new items* were then checked for viability by obtaining feedback from a small group of *African* volunteer participants as well as from a cultural expert, a colour-blind person and a small pilot sample for initial face validity and practical administrative detail checking of the instructions, response options and answer formats. Once the final pool of *new items* had been developed, phase 2 was initiated to evaluate the utility of the *new items* in terms of their psychometric properties (based on quantitative data gathered) and face validity data (collected through the open-ended questions).

4.5.3.4 Decisions on the integration of the methods

According to Ivankova et al. (2006), integration refers to the stage in the research process during which the methods are integrated. For exploratory sequential mixed methods, integration occurs after the first phase, because the outcomes in that phase are what inform the second phase, thus referred to as intermediary integration (Ivankova et al., 2006). In the current study, because the development of the *new items* in phase 1 (qualitative method) led to the evaluation of the *new items* in phase 2 (quantitative method), the integration was intermediary. In addition, phase 2 also had a qualitative element, which means the integration also occurred at the end with the final collation of quantitative and qualitative results.

4.5.3.5 Procedure for phase 1: Development of the new items

A creative process of exploration of African art and cultural artefacts was initiated to make it possible to combine the art and science of item development (Gregory, 2007). Typical steps in the development of items entail identifying the purpose of the items, deciding on the item and response format, writing items, reviewing items, arranging and finalising the number of items and instructions, administering items and item analysis (Anastasi & Urbina, 1997; Foxcroft, 2004, 2013; Gregory, 2007;

Kanjee, 2006; Moerdyk, 2015). As mentioned earlier, the development process of the items will be discussed in chapter 5.

4.5.3.6 Procedure for phase 2: Evaluation of the new items

Since measurement requires precision, accuracy and consistency, it is crucial that the quality of the items should be investigated as early as possible during the development stages (Mueller, Bullock, & Leierer, 2010). The data collection for this phase was conducted over a two-week period with a total of six instruments being administered to 946 participants. The *new items* and the LPCAT required the use of computers. The *new items* were presented in a PowerPoint presentation format, and a paper-and-pencil answer sheet was used to capture the participants' responses. Additional testing for personality, career preference and motivation also had to be scheduled into the test administration arrangements and included both paper-and-pencil and computerised assessments.

4.6 DATA ANALYSIS

Various data analysis techniques had to be used for the information collected throughout this study. Since the research included both qualitative and quantitative data, the techniques used for data analysis had to include both qualitative and quantitative methods. Aligned to the research design, the data analysis took into consideration the two phases of the study – hence the use of sequentially mixed analysis techniques (Onwuegbuzie & Combs, 2011).

4.6.1 Data analysis for phase 1: Development of *new items*

Qualitative information was gathered during the item development phase. The content analysis technique was therefore used (see point [H] in figure 4.1) for data analysis. According to Babbie (1989) and Leedy and Ormrod (2010), content analysis involves coding and categorising information in order to give meaning to and understanding of information. Babbie (1989) specifically referred to the coding of manifest content, which would include the surface content of oral and written

communication (e.g. specific words); and the directly visible and identifying characteristics of artefacts and objects (e.g. patterns and colours used in paintings). Content analysis was therefore applicable to the study in determining the different patterns and colours for the *new items* and to highlight and incorporate for use relevant comments and feedback from the participants regarding the culture-related elements (culture fairness and face validity), the possible effect of colours on colour-blind participants and practical instructions and administration considerations.

The process used to work through the interviews and questionnaire data for the evaluation of the *new items* in terms of their viability included organising the information into sentences and specific words. These were then highlighted and grouped into categories or themes, and integrated in order to address the research questions on the viability of the *new items* (Leedy & Ormrod, 2010).

4.6.2 Data analysis for phase 2: Evaluation of items

For the evaluation of the *new items* in terms of their utility, the data was analysed quantitatively using descriptive and inferential statistics; and the Rasch model was used for item analysis (see point [I] in figure 4.1). The SSPS program, version 22 (SPSS, 2013) and Winsteps version 3.71.0 (Linacre, 2011b) were used for statistical analysis. Additional qualitative data gathered during this phase was also analysed by means of content analysis for further evidence of face validity (see point [I] in figure 4.1).

4.6.2.1 Descriptive statistics

Descriptive statistics were used to describe the composition of the samples and to provide the presentation of the frequency distribution tables, percentages, means, standard deviation, graphs, et cetera (Welman, Kruger, & Mitchell, 2009). For determining the relationship between the total scores of the *new items* and the LPCAT, the correlation coefficient was used in which the magnitude, direction and strength of the coefficient were used for interpretation (Leedy & Ormrod, 2010).

4.6.2.2 Classical test theory (CTT) item analysis

According to Urbina (2004), CTT is used in the development and evaluation of psychological tests focusing on the total score of the test. At item level, the information from CTT item analysis indicates the easy and difficult items (Fan, 1998). This is referred to as the item difficulty (*p*-value), where the proportion of the participants who responded to the item correctly is determined (Fan, 1998; Hambleton & Jones, 1993; Urbina, 2004). This would be important for the *new items* as there would be an indication of the level of difficulty of the different item format types. The ideal in test development is to include items that cover a reasonably wide range in terms of difficulty level, with the bulk of items close to the average range of difficulty (Gregory, 2007). CTT also provides information on item discrimination, which is determined by the correlation between the item response and overall total on all the items (Gregory, 2007).

4.6.2.3 Rasch analysis

The Rasch model was developed by George Rasch (1961), a Danish mathematician, based on the assumption of the relationship between the ability level of the participants and the difficulty level of the item (Bond & Fox, 2007; Lantano, 2010; Rasch, 1979). The relationship is such that when the ability level of the participants increases, the probability of answering the items correctly also increases (Mueller et al., 2010). For measurement, a common scale was created for Rasch analysis on which the estimation of ability level and item difficulty can be indicated. This is referred to as the logit scale (Bond & Fox, 2007). The Rasch model (or Rasch measurement) is appropriate for test development and validation (Maree, Maree, & Collins, 2008, Mueller et al., 2010; Wright & Stone, 1979) – hence the researcher's decision to use it in the current study.

According to McCamey (2014), the Rasch model is similar in many ways to the one-parameter logistic model (1PL). Rasch analysis is appropriate for item development because items can be designed to fit the model expectations from the outset (De Bruin, Hill, Henn, & Muller, 2013; Edwards & Alcock, 2010; Maree et al., 2008; McCamey, 2014; Tennant & Conaghan, 2007). Fundamental to the requirements of

Rasch measurement, Bond and Fox (2007) highlighted the importance of checking unidimensionality (one dominant trait being measured) and local independence (items being independent of each other). Linacre (2005; 2009) referred to unidimensionality as a necessary condition for Rasch analysis, in which the items are expected to measure one single latent trait. The other assumption of local independence entails ensuring that items do not influence one another – which means the items do not correlate with each other (Linacre, 2009; McCamey, 2014). When these assumptions are satisfied, the items can be deemed "stable and not dependent on the sample being assessed" (Chachamovich, Fleck, Trentini, & Power, 2008, p. 311).

Another key requirement for Rasch analysis is that of data fitting the model (Bond & Fox, 2007; Thompson & Barnard, 2009; Wright & Stone, 1999). In order to determine the extent to which the data fits the model, fit statistics are used. Fit statistics entail both the infit and outfit statistics, which are mean-square summary statistics that indicate potentially problematic items (Bond & Fox, 2007; Greene & Frantom, 2002; Wright & Stone, 1999; Wu & Adams, 2007). The infit mean square (irregular responses influenced by on-target responses) is the distance between the person position and item difficulty, and the outfit mean square (irregular responses influenced by off-target responses) is the unweighted mean square order (Bond & Fox, 2007). Wright and Stone (1999) described the fit statistics as a tool to monitor responses for quality control and validation of the items. In situations where there are misfit statistics, these are referred to as over-fitting and under-fitting items and people (Bond & Fox, 2007; Wu & Adams, 2007). According to Mueller et al. (2010), fit statistics confirm the assumption of unidimensionality and the consistency of the participants' response to each item. However, the Wu and Adams' (2007) cautionary assertion on fit statistics should be taken into consideration. They cautioned that fit statistics should not be accepted or acted upon without critical consideration. They highlighted the use of reliability and item discrimination as additional sources of decision making and investigating the sources of misfit as an alternative to gain a better understanding of problematic items (Wu & Adams, 2007). According to Thompson and Barnard (2009), misfit is due to items that are of poor quality or

instances where the items can be good but do not form part of the set of items being analysed.

In item analysis, the quality of items is measured using item difficulty (the proportion of people who answer the item correctly) and item discrimination or the degree to which items discriminate (the degree to which items can differentiate correctly) on the construct being tested (Anastasi & Urbina, 1997; De Kock et al., 2013; Foxcroft, 2013). According to Bardaglio, Settanni, Marasso, & Musella (2012), DIF investigations highlight items that show indications of possible biases, and this was significant for the outcomes of the current study.

4.7 CHAPTER SUMMARY

In this chapter, the process followed in this research project was described. In the discussions, the research design, sample, data collection methods, data analysis, instruments used and the analysis techniques applied for both the development and evaluation phases of this research were explained. Regarding the quotation at the beginning of the chapter, the process followed in the research was clearly described, and an illustration of the process was included in figure 4.1 in order to provide a visual representation. As noted earlier, in chapter 5, the process followed for developing the *new items* will be described in detail.

CHAPTER 5

DEVELOPMENT OF THE *NEW ITEMS*

We can't solve problems by using the same kind of thinking we used when we created them – Albert Einstein (1879-1955)

5.1 INTRODUCTION

Discrimination is part of the positive qualities and challenges of psychological testing. As Cascio (1987) noted, a well-designed test battery should discriminate between those who have more of the trait measured by the tests and those who have less of the trait. However, discrimination that is not related to the construct being measured and is thus unfair, has also been one of the reasons tests have been mistrusted (Claassen, 1997; Kanjee, 2006; Nzimande, 1995). This is but one example of the challenges inherent in psychological tests in the midst of the evidence that psychological testing – and in particular cognitive assessment – is the most criticised practice in modern psychology (Gregory, 2007; Pedrajita & Talisayon, 2009). Continuous research and development are therefore required to ensure that the criticisms are continuously addressed creatively and scientifically. The need for the current study emanated from such criticisms – hence the development of the *new items* inspired by African art and cultural artefacts for multicultural cognitive assessment.

This chapter provides an overview of the exploratory process of developing the *new items*. The motivation for the development of the *new items* is highlighted and the steps followed to develop the *new items* are discussed.

5.2 MOTIVATION FOR DEVELOPING THE *NEW ITEMS*

Although South African history is characterised by a story of exclusion in psychological testing and assessment, great strides have been made to change the story to one of fairness, equity and inclusion. The story shifted significantly with the

political and legislative changes post 1994 (Foxcroft, Roodt, & Abrahams, 2013a). The introduction of and adherence to the requirements of the Employment Equity Act of 1998 (EEA) could be viewed as a new chapter in the story. This has been consequential to the increased prospects and opportunities for research on developing new instruments, fairness and bias investigations, validation of existing instruments and test adaptation processes (De Beer, 2000; Foxcroft, 2004; Foxcroft & Aston, 2006; Paterson & Uys, 2005; Rothmann & Cilliers, 2007; Van de Vijver & Rothmann, 2004). Maree (2010) commended the progress made in the sphere of developing and standardising assessment instruments in South Africa. The development of the *new items* in this research could make a further contribution, and to some extent add to the progress.

Different voices (Claassen, 1997; Nzimande, 1995; Theron, 2007) have emphasised the importance of taking into consideration social, economic and educational differences when testing multilingual and multicultural groups in the South African context. Maree (2010) cited similar considerations as being of importance globally. This is supported by the call made by Vasquez (2012), who highlighted the reduction of discrimination, the promotion of diversity and addressing educational inequalities through the use of psychology as an applied science, as some of her focus areas for her tenure as the President of the American Psychological Association (APA). For Foxcroft and Roodt (2013a), these considerations all contribute to the continuous challenge and importance of fair psychological assessment in South Africa. The development of the *new items* offers different voices and alternative options for cognitive assessment to be considered in an attempt to address particular issues regarding familiarity with the stimulus material and item content and test anxiety in the fairness challenge.

The issues and challenges raised with regard to psychological tests are certainly not new. Penrose and Raven (1936), for example, noted the urgent need for new designs of intelligence tests that would eliminate the effect of educational background differences on test performance. During that time, the debates were on perceptual tests and the argument was on the development and standardisation of nonverbal tests (Penrose & Raven, 1936). The same old problems and challenges of

the 1930s continue to surface as problems and challenges of the present day. Claassen (1997) and Foxcroft (2004) commented on cognitive tests, with concerns about content that is focused on crystallised abilities, which tends to favour individuals who come from backgrounds with access to more and better quality education and exposure to higher socioeconomic conditions. As suggested by De Beer (2005) and Schaap and Vermeulen (2008), nonverbal figural tests can be regarded as a fair cognitive assessment alternative in multicultural contexts. However, Wicherts, Dolan, Carlson, and Van der Maas (2010) raised similar concerns about nonverbal cognitive tests. They referred to the challenges participants might be faced with when responding to the unfamiliar geometric shapes of nonverbal figural cognitive tests such as the Standard Progressive Matrices (Wicherts et al., 2010). This challenge of unfamiliar geometric shapes was also identified by critics who attributed it to the Eurocentric origins of the assessments (Maree, 2010; Wicherts et al., 2010). In his reflections, Maree (2010) raised questions about whether a balance could be found between the Eurocentric and Afrocentric perspectives in psychological assessment practice in South Africa. The development of the *new items* could therefore make a contribution in terms of the original content of items that is inspired by indigenous material and is thus more familiar – possibly lowering test anxiety.

Notwithstanding the above motivations, the underlying drive for the development of the *new items* was also the curiosity of the researchers. According to Leedy and Ormrod (2010, p. 3), an “inquisitive mind” is the origin of the research process. In the current study, the curiosity was “what if” a way could be found to use African artefacts as inspirations in the development of items measuring general nonverbal figural reasoning ability. The curiosity developed into a research question of how such items could be developed and whether such items would be viable, whether they would meet the core psychometric requirements and whether they would be perceived as more fair.

5.3 PLANNING AND WRITING THE *NEW ITEMS*

The development of the *new items* was a manifestation of the combination of art and science (Ambrose & Anstey, 2010; Gregory, 2007; Reckase, 1996). The process entailed the exploration of African art and cultural artefacts to solidify the ideas and stimulate the inspirations, while upholding the scientific requirements to anchor and substantiate the process with empirical evidence. In developing the items, the typical steps to developing a measure were adapted from various psychological textbooks (Anastasi & Urbina, 1997; Foxcroft, 2013; Gregory, 2007; Kanjee, 2006; Moerdyk, 2015) in order to structure the process. The steps are discussed below.

5.3.1 Identifying the purpose of the *new items*

According to Durrheim and Painter (2006), this is the step of conceptualising where the domain of interest is defined. This means that an explanation of the elements or constructs to be measured should be provided (Ambrose & Anstey, 2010). Hinkin (1995) emphasised the importance of this step as that of providing clear boundaries for determining content and construct validity. The accuracy of the measurement can only be confirmed if the elements or characteristics of the traits or constructs have been clearly clarified (Ambrose & Anstey, 2010; Hinkin, 1995). For this study, the key elements of the new format items were the African art and cultural artefacts which were used as inspiration and the intended purpose of measuring general nonverbal figural reasoning ability based on the fluid ability construct defined by Cattell (1963).

According to Cattell (1963) and Raven (2000), general fluid ability or eductive ability entails adaptation to new situations and abstract reasoning ability. Kvist and Gustafsson (2008, p. 423) described fluid intelligence as the “capacity to solve novel, complex problems, using operations such as inductive and deductive reasoning, concept formation, and classification”. The Learning Potential Computerised Adaptive Test (LPCAT) is an example of a test using nonverbal figural items to measure the construct of fluid ability (De Beer, 2005). The use of such nonverbal figural items has been identified as appropriate in multicultural contexts because it reduces the negative impact of language on test results (De Beer, 2005; Penrose &

Raven, 1936). Therefore, similar to other matrices tests (Penrose & Raven, 1936; Raven, 2000), for the new format items, geometric figures were used where a series of figures was presented with one entry missing. These items entailed identifying relationships, similarities and differences between shapes and patterns, and recognising visual sequences and relationships between objects (Gregory, 2007; Kunda, McGreggor, & Goel, 2010; Penrose & Raven, 1936).

5.3.2 Sourcing ideas and inspirations

Although for the purpose of this study the word “Africanising” was used simplistically to depict the African art and cultural artefacts that were used as inspiration in the development of the *new items*, “Africanising” or “Africanisation” is a term that has been studied and defined in various disciplines. Louw (2009, p. 63) described it as the “embracing of our African heritage”. Franke and Esmenjaud (2008) referred to it as the process of increasing the extent and quality of African influence, while Msila (2009) explained it as the process of reinstalling the African culture. According to Eglash and Odumosu (2005), the African identity is not necessarily homogenous; however, the subtle similarities across multiple cultural groups can be recognised. Hence the inspirations used for the *new items* were not limited to any ethnic or cultural group, but any artefacts that could be deemed African were used.

Collecting photographs helped to provide visual evidence in the process of developing the *new items*. According to Onwuegbuzie et al. (2010), photographs are recognised as an innovative data collection strategy in qualitative research. The photographs used for this study were taken in flea markets and curio shops, at traditional weddings and tourism destinations, and from internet sites with African art and fashion. It is acknowledged that the “self” becomes an instrument in qualitative research (Babbie & Mouton, 2000) – hence the process of collecting these photographs was a subjective one. The choice of objects captured was dependent on what the photographers deemed African, interesting and beautiful (Cruickshank & Mason, 2003; Donaldson, 2001; Peters & Mergen, 1977).

In terms of the ethical considerations highlighted by Pauwels (2008) and Ray and Smith (2011), issues such as anonymity, the intrusiveness of the camera and informed consent were noted. For this study, no facial images were used to ensure anonymity, and in those instances where permission was required, for example, in shops or flea markets, a brief explanation was offered. Since the focus was on the patterns and colours of the artefacts, brand names, logos or aspects of the photographs that could be sensitive were cropped out of the images.

The artefacts used in this study included African prints, paintings on walls, decorations and beadwork, as depicted by some of the example photographs presented in figure 5.1 below. Nettleton (2010) referred to similar examples as indigenous crafts. In his studies on ethnomathematics, Gerdes (2001) successfully used similar cultural artefacts of decorative art such as paintings on walls, engravings and ornaments.

The symbolic representations found in the artefacts were not that different to the geometric figures found in nonverbal figural items. According to Jones (2002, p. 126), “many cultural artefacts involve geometric principles”. This is similar to observations made by researchers in ethnomathematics, where designs in cultural artefacts were found to be easily aligned to geometric shapes and figures used in mathematics (Davison, 2007; Eglash & Odumosu, 2005; Rosa & Orey, 2010). The African artefacts had designs of triangles, rectangles, circles, et cetera, that were shaped in the same way as the geometric figures found in the traditional item formats of tests such as the LPCAT. The core characteristics of the items were therefore not changed.

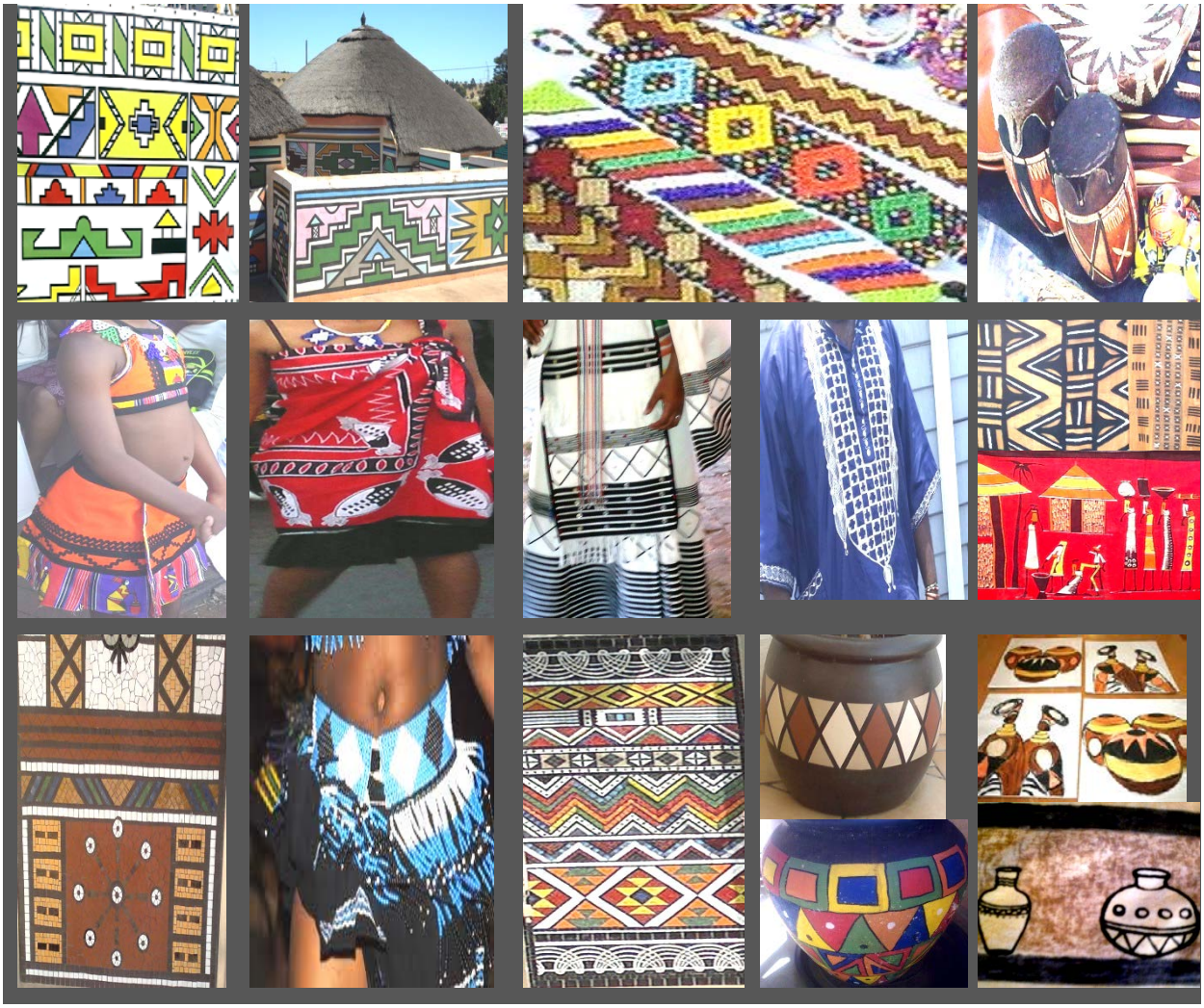


Figure 5.1. Examples of African inspirations used for item development

5.3.3 Creating the *new items*

Creativity entails generating new ideas, changing or finding alternatives to the status quo, striving for uniqueness and pioneering new approaches or ways of doing (Amabile, 1998; Amabile, Barsade, Mueller, & Staw, 2005; Hirschman, 1980; Sternberg, 2006). According to Sternberg (2004), creativity is the ability to produce work that is both novel (or original and unexposed) and appropriate (or useful and adaptive). In the current study, the creativity involved in the development of the new items was to balance the use of the African artefacts as inspirations with the core elements and requirements for items for the measurement of nonverbal figural reasoning. This was a challenge, but also experienced as an exciting and highly enjoyable part of the project. Experiencing these positive emotions was

acknowledged in the literature, which is to be expected when creative activities are involved (Amabile, 1998; Amabile et al., 2005).

As the process of drafting the items was initiated, it was necessary to ensure that the geometrical shapes and patterns of the artefacts remained intact and visible rather than using the original cultural artefacts such as houses (huts), baskets and calabashes, shields, drums, dresses, et cetera. As noted by De Beer (2005), any pictorial usage of cultural artefacts could be viewed as culturally loaded. Some of the selected colours, shapes and figures that were identified and extracted from the photographs and used in the development of the *new items,* are illustrated by the examples presented in figure 5.2 below.

The *new items* were structured as typical geometric analogy problems in which various figures were presented to follow a certain pattern or sequence; and for each item, there would be a missing figure represented by a question mark (Anastasi & Urbina, 1997; De Beer, 2000; 2005; Gregory, 2007; Penrose & Raven, 1936). The question mark represented the task (question) to which the respondents had to find the correct answer (replacement). The possible answer to the missing figure would need to be selected from four alternatives – similar to the way in which multiple-choice questions are structured. The correct answer for each item is based on recognising the pattern or sequence of each item. Penrose and Raven (1936) referred to this as the relationship that had to be identified and listed the types of relationships as being based on similarity, opposition and addition.

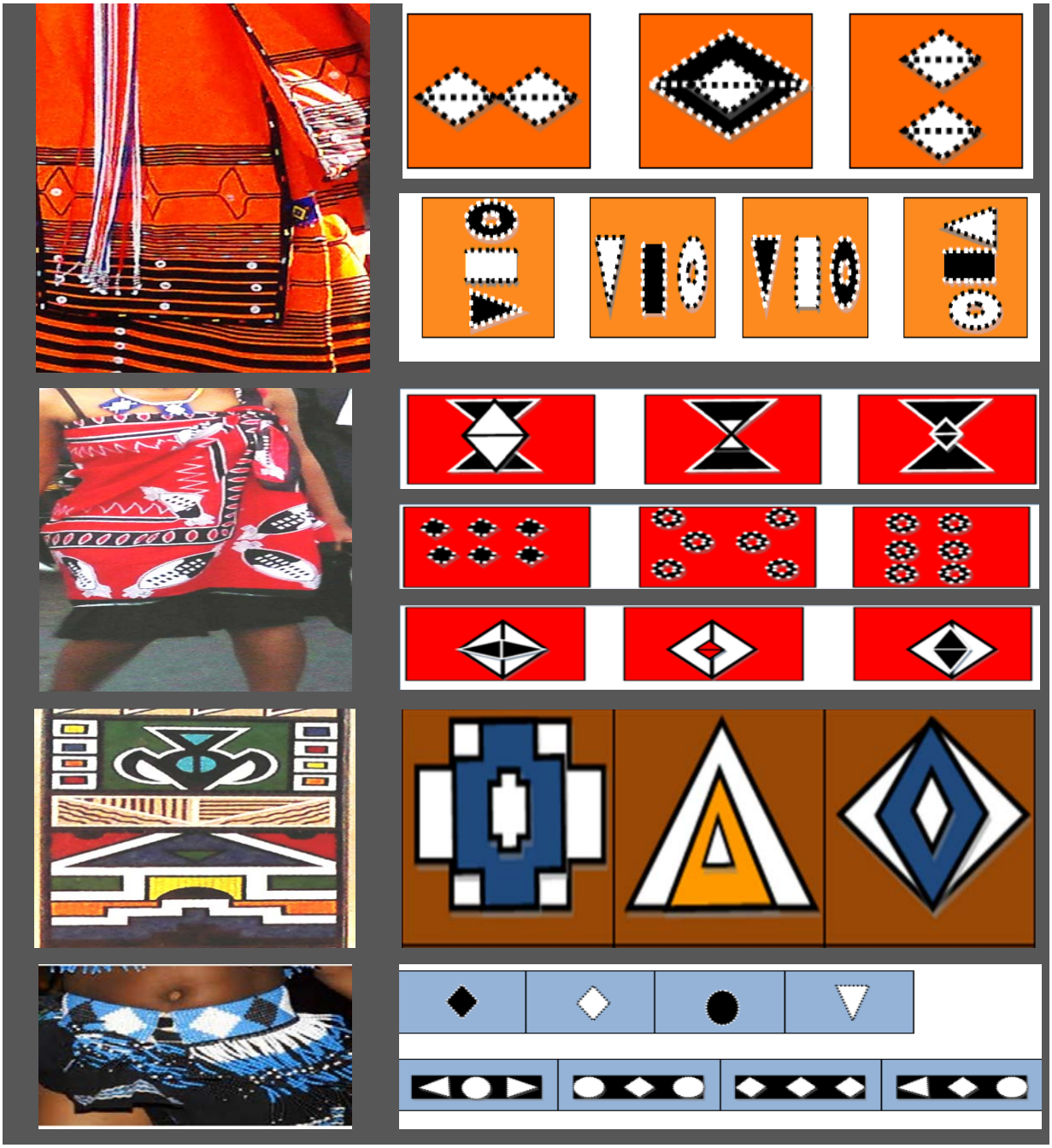


Figure 5.2. Examples of how African inspirations were used

Some of the commonly used item formats have figures presented in rectangular (block) patterns (Penrose & Raven, 1936). In addition to this traditional style, the *new items* also used circles (wheels) and triangles as the base to the patterns. The question mark positions were also changed for different questions to ascertain if this would impact on the item difficulty – therefore, rather than always having the question mark placed at the end of the sequence, the question mark was placed in

different positions in pattern sequence. The other obvious new elements were that of the colour and designs of the *new items*.

5.3.4 First draft of *new items*

A group of 40 items was initially developed using standard Windows office computer software. This first draft of items was emailed to six volunteer African participants to obtain preliminary face validity feedback on the appropriateness of the symbols used for items with regard to their identifiable representation of African elements. The participants were not expected to answer the items, but to merely look at them and comment about their thoughts and views on the items.

An expert on African cultural who is an academic in the African Languages Department in higher education was also shown the new format items during a one-on-one interview and discussion session. The focus was again on whether the symbols and colours used in the items captured what could be considered typical African elements and characteristics.

5.3.5 Second draft of the *new items*

Based on the first draft feedback, additional items were developed taking into consideration the comments provided. These items were specifically structured to capture the required elements of nonverbal figural reasoning items. The second pilot testing of the *new items* that was conducted comprised 24 items, and African school learners were used to obtain feedback. The participants were requested to provide answers to the items and to respond to the open-ended questions at the end of the answer sheet. The open-ended questions were about what questions the participants liked or disliked, and why. They were also asked to choose the colours used in the items in terms of the colours they liked the most and least. The different item formats (blocks, circles/wheels and triangles) were also considered, and the participants were asked which of the three formats they liked or disliked, and why. Lastly, an open space was provided for general comments.

The goal of this second pilot test was to determine the clarity of instructions and practical utility of the response formats in a paper-and-pencil administration for which the items were printed in colour. The instructions were given as a PowerPoint presentation. As part of this presentation, the participants were informed of the purpose of the session and that the results would only be used for gathering research data. They were also familiarised with the items by providing and explaining a set of examples for each item type. Since no time restrictions were set for the *new items,* the participants could take their time working through the items, and fortunately the class period that was used for this was the last one of the day, so extra time was available if needed.

The new items were also reviewed for elements that might be challenging for colour-blind individuals. This process also involved a one-on-one session, where a colour-blind participant was shown the items and had to comment on the appropriateness of the colours used for persons who had similar challenges. According to Rigden (1999), colour blindness is an important challenge to consider for any computer designs as it can impact on the performance of the participants. She recommended that colour use should be for decorative purposes rather than being part of the clue of the question (Ridgen, 1999). This was the way in which colour was used for the *new items*.

5.3.6 Final pool of *new items*

The final pool of *new items* was developed taking into account all the feedback from the previous drafts. Six different item types were decided on – these are discussed below. Although traditional item formats were designed to take the rectangular (blocks) format, new item formats were also developed that used the shapes of the triangles and circles (wheels). The items developed for this round totalled 200, with a minimum of at least 30 items of each type.

The general guidelines agreed upon for writing the *new items* were as follows:

- All items had to be nonverbal figural in format – the basic figures and shapes used in the *new items* included rectangles (blocks), triangles, circles, arrows, lines, hexagons, et cetera.
- The item structure had to be multiple-choice questions, in which four response alternatives were used.
- The core item characteristics of a figural series or analogy or matrix pattern for items had to be maintained – the focus was thus on the pattern sequence and transformations rather than what colour or design or symbols had been used in the *new items*.
- The primary colours to be used in the *new items* were the five selected colours of red, blue, orange/yellow, green and brown – with black and white as the basic contrast or additional colours.
- Some items could be repeated, but with changes in the position of the question mark.

The different item formats are illustrated below with example items.

5.3.6.1 Type 1 items

Type 1: A block is divided into five squares with figures presented to form or follow a series pattern or sequence. One of the squares has a question mark which should be replaced by choosing the correct alternative to replace the question mark from the four alternatives provided below the figure series.

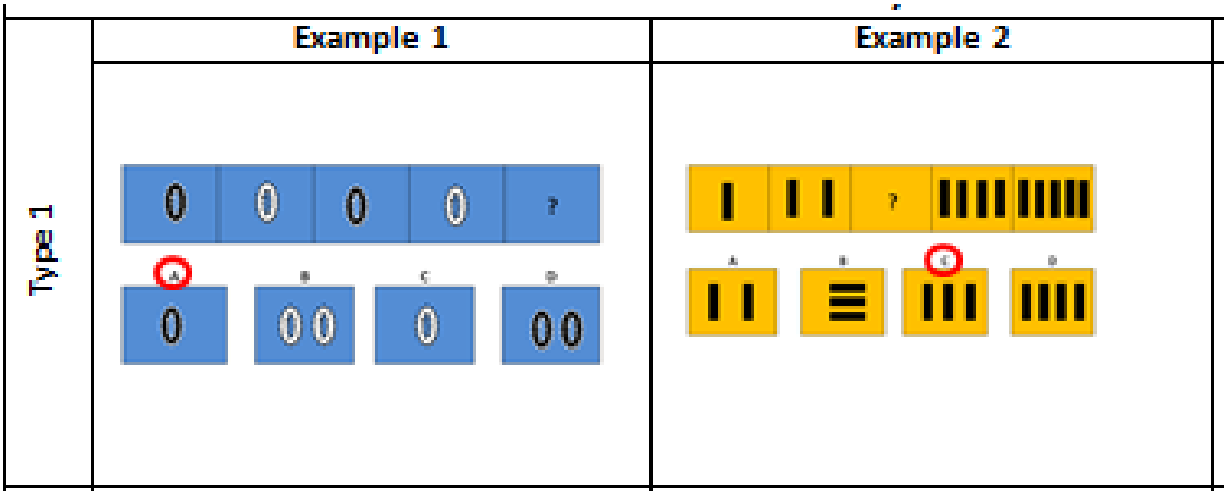


Figure 5.3. Example of type 1 items

5.3.6.2 Type 2 items

Type 2: A block is divided into four squares with figures in two rows and two columns. For each row and column, the two figures form a pattern or sequence. One of the squares has a question mark which should be replaced by choosing from the four alternatives given below the figure series, the correct option to replace the question mark.

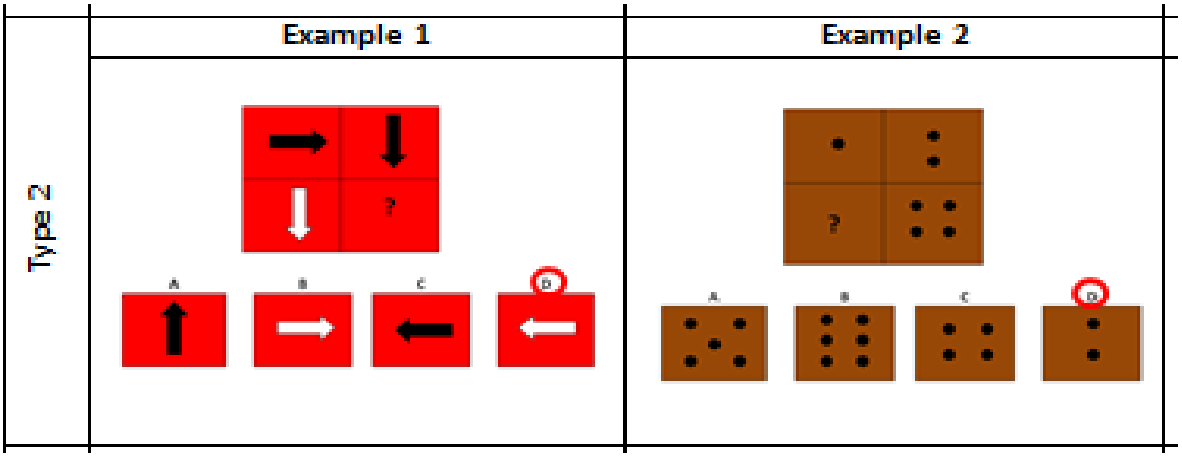


Figure 5.4. Example of type 2 items

5.3.6.3 Type 3 items

Type 3: Two pairs of blocks are presented with figures that form a pattern or sequence. The pattern between the figures can be determined for each pair, or it can be cross-checked by corresponding the first and the third figure from each pair or the second and fourth figures in order to complete the one with a question mark. One of the squares has a question mark, which should be replaced by choosing from the four alternatives given below the figure series, the correct option to replace the question mark.

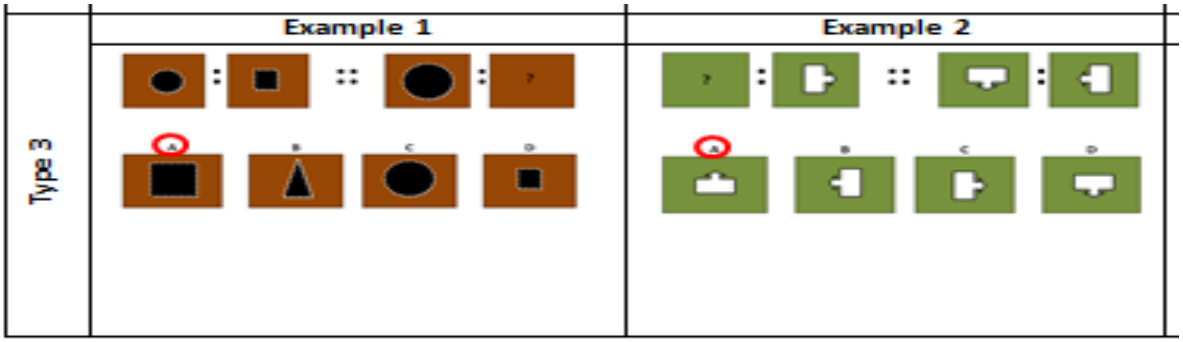


Figure 5.5. Example of type 3 items

5.3.6.4 Type 4 items

Type 4: A block is divided into nine squares with figures in three rows and three columns. For each row and column, the three figures form a pattern or sequence. One of the squares has a question mark, which should be replaced by choosing from the four alternatives given below the figure series, the correct option to replace the question mark.

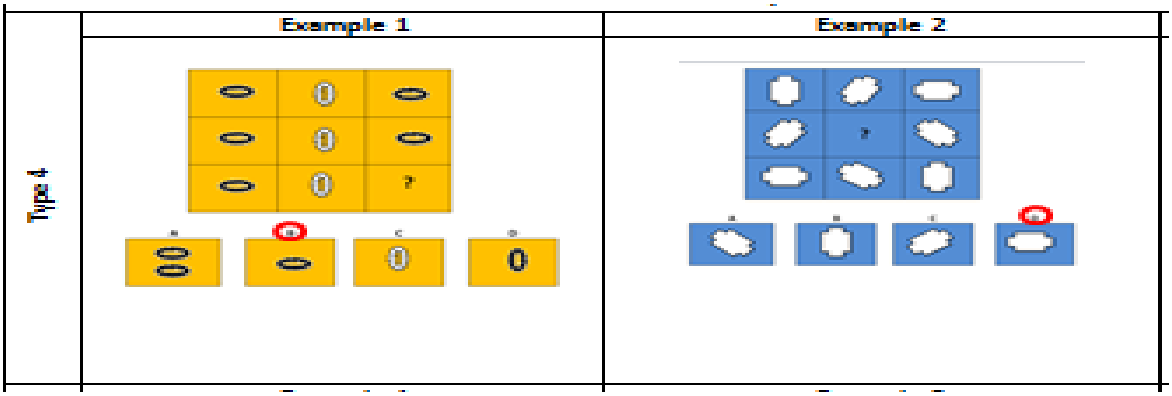


Figure 5.6. Example of type 4 items

5.3.6.5 Type 5 items

Type 5: A triangle is divided into six parts from top to bottom, with figures that form or follow a pattern or sequence with one figure missing. One of the spaces has a question mark which should be replaced by choosing from the four alternatives given below the figure series, the correct option to replace the question mark.

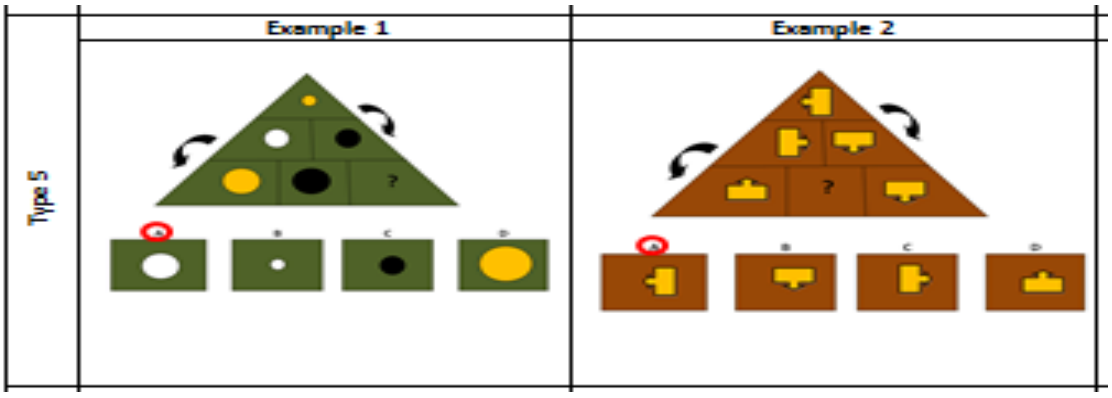


Figure 5.7. Example of type 5 items

5.3.6.6 Type 6 items

Type 6: A circle is divided into six parts with figures that form or follow a certain pattern or sequence, with one figure missing. One of the squares has a question mark which should be replaced by choosing from the four alternatives given below the figure series, the correct option to replace the question mark.

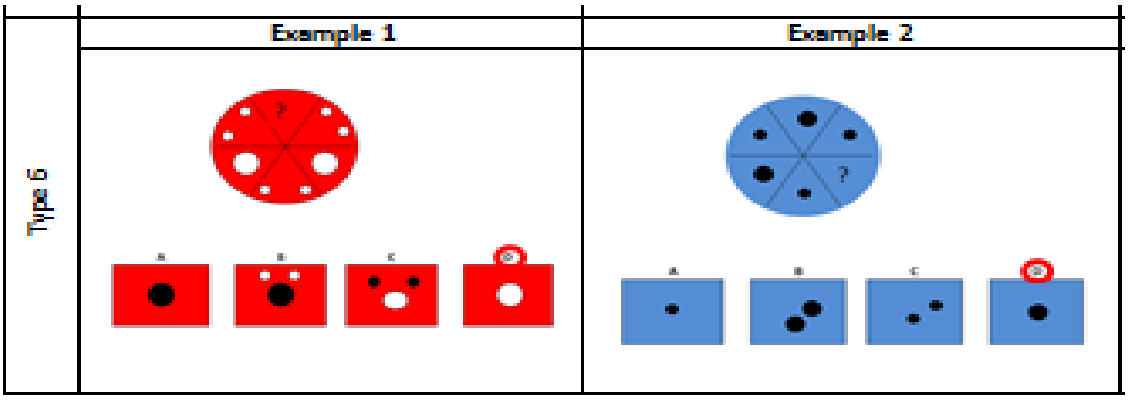


Figure 5.8. Example of type 6 items

5.3.7 Reviewing the *new items*

The aim of the study was to develop *new items* inspired by African art and cultural artefacts and evaluate them to ascertain their viability in terms of perceived fairness and their utility in terms of their psychometric properties. Other than the qualitative feedback that was used for the first two drafts of the *new items;* the review of the final item pool (as mentioned in the previous chapter) mainly focused on the quantitative analysis of the *new items*. Although the qualitative feedback was used for face validity purposes, the quantitative analysis included descriptive results for all the *new items* and for each item type, such as minimum and maximum scores, mean, standard deviation, difficulty levels, reliability and validity.

5.4 CHAPTER SUMMARY

In this chapter, the development process followed for the *new items* was described. In the discussion, the motivation for developing the *new items*, and the planning and

writing of the *new items* were highlighted. Examples of the different item types were also provided. According to the quotation at the beginning of the chapter, the challenges that were highlighted in section 5.2 could be better addressed with creative alternatives for instruments to be explored. The next chapter focuses on the results from both the qualitative and quantitative data collected.

CHAPTER 6

RESULTS

"The only man who behaves sensibly is my tailor; he takes my measurements anew every time he sees me, while all the rest go on with their old measurements and expect me to fit them" - George Bernard Shaw

6.1 INTRODUCTION

The field of psychological testing and assessment is both fascinating and resilient. Despite the mistrust and the many challenges relating to psychological tests and their use, the field continues to grow because the need to assess human attributes is ongoing (Weiner, 2013). The assessment of human attributes dates as far back as biblical times where references were made to Gideon who "observed how his soldiers drank water from a river so he could select those who remained on the alert" (Foxcroft, Roodt, & Abrahams, 2013a, p. 10). As highlighted in chapter 1, the curiosity about and intrigue of observing such actions and developing some form of profile about the soldier's alertness are fascinating.

Accurately and scientifically reporting on assessment processes and interventions is crucial to the continuity of the field. The notion of "publish or perish" referred to by Effendi and Hamber (2006, p. 112) thus emphasises the importance of disseminating research results. Although used in a different context, the message is applicable to the current research and to the field of psychological testing and assessment. After working through all the research steps and processes presented in the previous chapters, it would be futile not to report the results. According to Weiner (2013), publication of new information on the utility and benefits of the use of psychological assessment would contribute to the continuous growth of the field – hence the importance of the results chapter to the thesis, and to the field of psychological testing and assessment.

The results chapter is used to present the results based on the two phases of the research study, starting with the development phase and then the evaluation phase. Tables and graphs are included to further illustrate the results.

6.2 PHASE 1 RESULTS: DEVELOPMENT OF THE *NEW ITEMS*

The aims of this phase were to develop the *new items* from inspirations of African art and cultural artefacts in order to measure nonverbal figural (fluid or *gf*) reasoning ability and to evaluate the viability of these items in terms of their perceived cultural fairness. The results discussed in this section are based on the qualitative data obtained during the drafting of the *new items,* starting with the first draft of items in which email responses were obtained, followed by interviews and discussions with an expert of African culture and a colour-blind person, as well as comments from the first pilot administration of the *new items* (as indicated by [A], [B], [C], [D] and [H] of figure 4.1 – see chapter 4).

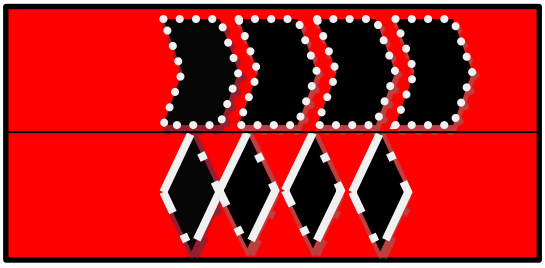
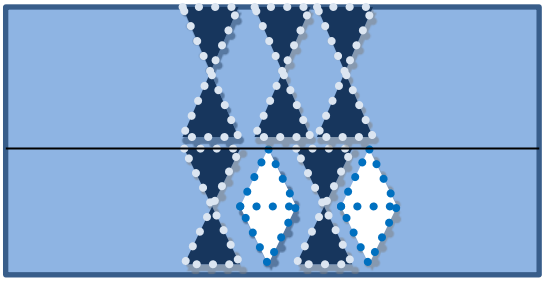
For any item development process, face validity is vital because first impressions and perceptions are likely to play a role in the acceptance of the final product (Anastasi & Urbina, 1997). The research question addressed by this data was how well the appearance of the *new items* represents the African art and cultural artefacts that inspired them. The participants viewed the *new items* as positively capturing the Africanness of the art and artefacts. They used words such as Afrocentric, traditional decoration and ethnic colours, and these comments were deemed to indicate the appropriateness of the *new items* in representing the inspirations. In table 6.1 (below), a few examples of the words used in relation to the symbols presented to the participants confirm that the appearance of the new items was perceived as having African elements, and that the patterns, symbols and colours chosen were also viewed as representing African art and cultural artefacts. This feedback was highly encouraging as it confirmed that the symbols that were created had captured the origins of the artefacts appropriately.

The African cultural expert also confirmed the symbols as representing different African cultural groups. He could identify the colours and patterns as belonging to specific ethnic groups (e.g. Ndebele, Zulu, Swati, etc.). However, he had a concern about whether all the groups would be represented equally in the designs. The explanation and motivation for developing the *new items* based on the artefacts, were not for the items to specifically represent each grouping equally – but rather to

create enough relevance and familiarity for anyone to see the *new items* as being African inspired.

Table 6.1

Examples of comments on the appropriateness of the symbols

Symbols	Comments
	❖ *Zulu hat colouring* ❖ *Theme is African, maybe amaMpondo* ❖ *More of jewellery beads that are African* ❖ *Zulu colours*
	❖ *Ethnic colours – very nice and African* ❖ *Ethnic colours, modern African theme and Afrocentric* ❖ *Culture hut* ❖ *Touch of modern African theme*
	❖ *Shangaan or Venda – not sure, but it is African* ❖ *Thinking of Venda or so theme* ❖ *Ndebele paintings*
	❖ *[5]Seshoeshoe print – very African. Sotho to be exact* ❖ *Print material* ❖ *Material of seshoeshoe with patches* ❖ *I like the pattern – very familiar*

[5] Material or dress mostly associated with Basotho women (Pheto-Moeti, 2005)

Regarding the colour-blind participant, the selected colour combinations (blue, red, orange/yellow, green and brown combined with the basic colours of black and white) were not problematic for colour blindness, and the comments were that the patterns in the items were visible and identifiable and thus easily answerable.

The second set of items also received positive feedback from the second sample group of high school learners. On questions of which item formats they liked the most or the least, and why, the reasons provided were mainly based on the colour preferences and the level of difficulty (challenge). In the general comments section, participants raised various issues which were categorised in terms of the colour and symbols used; increased motivation and interest; and connection to surroundings. The participants seemed to like the use of colour and the different symbols in the *new items* – for example "… interesting mix of shapes, colours and patterns …"; "… exciting, familiarity of some patterns …"; "… looking forward to go to the next page to see what's in store" were some of the phrases used. Participants indicated interest in testing – for example, "I think learners will love taking tests that have colour and maybe they won't be scared or confused to answer the questions"; and another "… really enjoyed your questions it was quite fun and enjoyable it really helped a lot, please next time bring more of this with 50 or 100 questions …". It would seem that concerns that had previously been raised on the unfamiliarity with test material and content used in psychological tests (Van de Vijver & Rothmann, 2004), had to some degree been attended to. For example, comments such as "… reminds me of my mommy's dress", and another who mentioned that "I like the traditional stuff decoration" could be viewed as positive in addressing the unfamiliarity issue. One participant cited the reason for liking an item least as "the arrows are not cultural material although brown is" – further providing evidence of the positive reaction to what could be considered more culturally appropriate content.

The feedback from this phase was considered in relation to the reflections by Maree (2010) where he highlighted the need for change in instrument development from the Eurocentric themed instruments that were perceived as disadvantaging the majority of the rural and township sector of South Africa (Maree, 2010). The comments that referred to the African aspects of the symbols, patterns and colours, were therefore

viewed as encouraging for continuing with the development of the *new items.* According to Kgosana (2012), in order to improve acceptance of tests and face validity, test content should be directly related to the surroundings and experiences of the participants. The feedback provided positive support regarding the viability of the *new items* in terms of the responses to the use of symbols from the African art and cultural artefacts appropriately, thus contributing positively to face validity.

6.3 PHASE 2 RESULTS: EVALUATION OF THE *NEW ITEMS*

The aim of this phase was to evaluate the utility of the *new items* in terms of their psychometric properties. The results discussed in this section are based on quantitative and qualitative data obtained from the administration of the full item bank of 200 of the *new items* (as indicated by [F], [G] and [I] in figure 4.1). One should note at this stage that one of the items (item 171) was removed after finding that the options given were not correctly formulated, which meant that the question did not have one correct answer. Hence further analysis was done on 199 items.

6.3.1 Descriptive statistics

The descriptive statistics based on the *new items* and the scores from each *new item* type are provided in table 6.2. The values of the number of items for each *new item type*, minimum and maximum values, mean, standard deviation, skewness and kurtosis are tabulated. The mean value (average) represents the central tendency of the items for the responses, while the standard deviation represents the measure of variability of the scores (Salkind, 2014). A low standard deviation indicates scores that are closely placed around the mean, while a larger standard deviation means the scores are more widely distributed (Salkind, 2014). Totals of correctly answered questions were calculated for each of the *new item* types and for the overall total score of the *new items*.

According to Čisar and Čisar (2010) and Roodt (2013a), the skewness value indicates the symmetry of the distribution which can be normally distributed or skewed positively or negatively. Positively skewed distributions are deemed to show

items that are more difficult, while negatively skewed distributions are deemed to indicate items that are easier (Austin, 2013; Roodt, 2013a). All the *new item* types indicated in table 6.2 were skewed negatively, which means the items were generally easier, and many of the respondents answered the items correctly. The relatively high formal level of qualification of the sample group – compared to that of the general South African population – should be taken into account in this regard. Kurtosis provides information on the peakedness (peaked or flat) of the distribution (Austin, 2013). All the values presented in table 6.2 are negative, which indicates a flat distribution, meaning the frequencies of scores are similarly spread. Considering the values of both the kurtosis and skewness, all these values are between -1.0 and 1.0, which means the distribution could be considered normal (Čisar & Čisar, 2010; Roodt, 2013a). However, Čisar and Čisar (2010) cautioned on drawing conclusions about normalcy based on skewness and kurtosis alone, as these cannot provide a definitive interpretation. Instead, they recommended that additional tests such as the Shapiro-Wilk or Kolmogorov-Smirnov could be done (Čisar & Čisar, 2010).

Table 6.2

Descriptive statistics for the various new item types (N = 946)

Measure	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	All items
Items	35	34	36	31	33	30	199
Minimum	4	2	4	2	1	1	29
Maximum	35	34	36	31	33	30	189
Mean	20.07	19.32	22.92	17.53	17.99	19.45	117.45
Std. dev.	7.27	5.68	6.59	5.83	5.89	5.55	33.62
Skewness	-0.24	-0.45	-0.55	-0.12	-0.43	-0.68	-0.43
Kurtosis	-0.90	-0.28	-0.58	-0.62	-0.24	-0.30	-0.58

6.3.2 Classical test theory (CTT): Item difficulty (*p*-values)

Foxcroft (2013) referred to the indices of CTT item analysis as including the item difficulty (*p*) and discriminating power. The item difficulty value is determined by the

proportion of respondents who answer an item correctly, where the higher the percentages of respondents who answer the questions or items correctly, the easier the items are (Anastasi & Urbina, 1997; Foxcroft, 2013). For the current study, the *p*-values for all the *new items* are listed in appendix A.

A mean item difficulty around 0.5 is often recommended because it would indicate that items are within a difficulty range where most participants should be able to answer them correctly (Anastasi & Urbina, 1997; Gregory, 2007). However, Gregory (2007) further added specific considerations for four-option multiple-choice items, where he recommended an estimate of 0.63 as the ideal level of item difficulty, with indices ranging between 0.3 and 0.7. Therefore any items with indices below 0.3 are very difficult items, while those with indices greater than 0.7 are very easy items. From the list of *p*-values, 17 (9%) items out of the 199 had indices lower than 0.3, while 49 (25%) items had indices greater than 0.7. The majority of items (133 or 66%) had difficulty values between 0.3 and 0.7. The item difficulty indices ranged from 0.06 (6.1% – as a percentage or proportion of correct answers to a question) to 0.91 (91% – as a percentage or proportion of correct answers to a question). In table 6.3, the bottom five items (deemed the easiest) as well as the top five items (deemed the most difficult) are listed. Overall, the item difficulty of the *new items* provided satisfactory results.

Table 6.3

Five most and least difficult items

Most difficult items		**Least difficult items**	
Item & (item type)	*p-values*	*Item & (item type)*	*p-values*
165 (2)	0.06	29 (6)	0.91
61 (5)	0.06	26 (6)	0.91
11 (4)	0.13	54 (6)	0.90
98 (3)	0.14	3 (3)	0.87
57 (3)	0.14	12 (4)	0.87

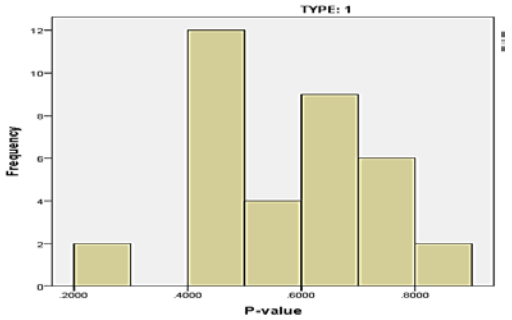
The means for the *p*-values for the different *new item* types and the total items are reported in table 6.4 below, and each of the graphical representations of the *p*-values is illustrated in figure 6.1. The same graphical representations provided in figure 6.1 are included in appendix B. As indicated in the table for all the various *new item* types, the percentage of correct responses was higher than 50%, which means the items were easier. The mean item difficulty values ranged between 0.53 (for *new item* type 5) and 0.68 (for *new item* type 6), and these could thus be viewed as acceptable. However, Foxcroft (2013) cautioned that the values are derived from frequencies of responses and may therefore be specific to that particular sample. For a more extensive analysis, it would be necessary to also determine the reasons for such values. As indicated earlier, the high formal level of qualification of the sample group – compared to the general South African population – should be taken into account. The *p*-values may be lower for a more generally representative sample of South African adults.

Table 6.4

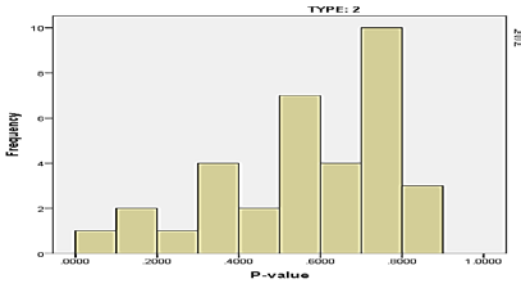
P-values for the various new item types

	Type 1	**Type 2**	**Type 3**	**Type 4**	**Type 5**	**Type 6**	**All items**
N	35	34	36	31	33	30	199
Mean	0.58	0.57	0.64	0.57	0.53	0.68	0.60
Std. dev.	0.15	0.22	0.20	0.16	0.20	0.15	0.19

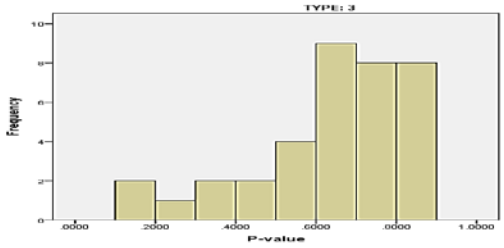
New item type 1



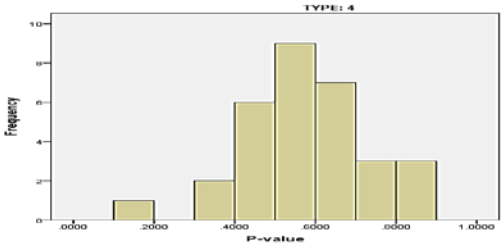
New item type 2



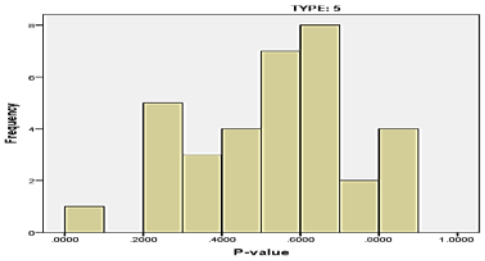
New item type 3



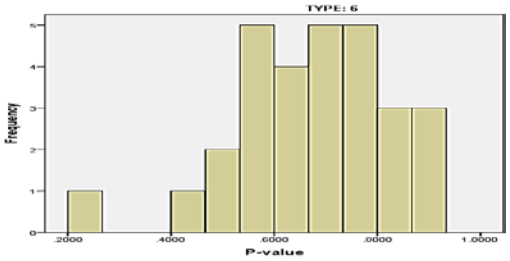
New item type 4



New item type 5



New item type 6



All items

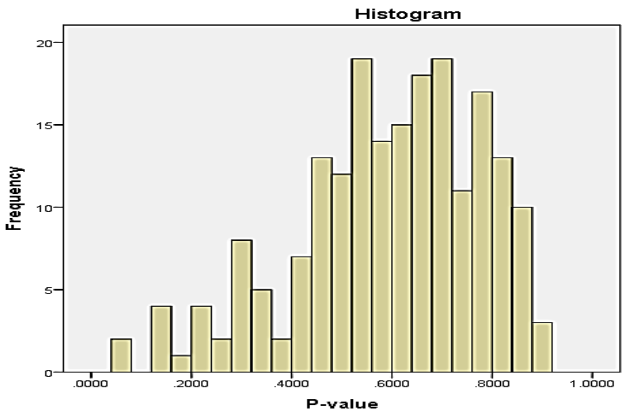


Figure 6.1. Graphical representation of the distribution of p-values

Parallel to the CTT analysis of the *new items*, Rasch analysis of the *new items* was also performed to evaluate the psychometric properties (utility) of the *new items*.

6.3.3 Rasch analysis

Rasch analysis for dichotomous responses was used to analyse the responses to the *new items*. According to Edwards and Alcock (2010), Rasch analysis makes it possible to rank, on the same logit scale, items by their difficulty and participants by their ability, based on their responses to questions. Other measurement (psychometric) factors reported on are the person separation reliability, item separation reliability, item infit and outfit statistics, and the differential functioning of the items. The presentation takes into account each of the six *new item* types and the total group of all *new items*.

6.3.3.1 Unidimensionality

One should bear in mind that factor analysis and principal component analysis are similar in terms of their aim of evaluating the dimensionality of the data set, but different in terms of their application (Bond & Fox, 2007; Sick, 2011). Various researchers have investigated the advantages and disadvantages of using one over the other of these techniques or using both (Bond & Fox, 2007; Coleman, 2006; Krishnan, 2011; Linacre, 1998; Sick, 2011). These were discussed in chapter 4. According to Clark (2007), using Rasch analysis during the pilot stage of the development of a test ensures that the balance between appropriate content (item difficulty) for the appropriate people (person ability) is achieved.

According to Bond and Fox (2007) and Wright and Stone (1999), the analysis of fit is the difference (residual) between the actual observed score (empirical data) and the expected score (based on the Rasch model's theoretical predictions). It is assumed that the Rasch measurement dimension should account for more variance in the data than any other of the dimensions (Bond & Fox, 2007). The strongest secondary dimension is referred to as the first contrast, and the following secondary dimensions as the second, third, fourth and fifth contrast, respectively (Bond & Fox, 2007).

As illustrated in table 6.5, the percentage of variance explained by measures (26.4%) is almost similar to the 26.6% indicated for empirical variance. Furthermore,

the values from the first contrast had the strength of 2.5%; while the second, third, fourth and the fifth had 1.3%, 1.1%, 1.1% and 0.9%, respectively. There was a large amount of unexplained variance unaccounted for by the model with the percentage of 73.4%, but this does not necessarily mean multidimensionality (Miyata, 2007). The values of the secondary contrasts (1st to 5th contrasts) are smaller than the percentage of variance explained by the measures, which means the model dimension accounts for more variance than all the other dimensions – thus confirming the unidimensionality requirement of the items (Miyata, 2007).

Table 6.5

Standardised residual variance of all items

		Empirical		**Model**
Total variance in observations	271.26	100%		100%
Raw variance explained by measures	72.26	26.6%		26.4%
Raw variance explained by persons	26.65	9.8%		9.7%
Raw variance explained by items	45.61	16.8%		16.6%
Unexplained variance (total)	199.00	73.4%	100%	73.6%
Unexplained variance in 1st contrast	6.73	2.5%	3.4%	
Unexplained variance in 2nd contrast	3.47	1.3%	1.7%	
Unexplained variance in 3rd contrast	3.03	1.1%	1.5%	
Unexplained variance in 4th contrast	2.89	1.1%	1.5%	
Unexplained variance in 5th contrast	2.42	0.9%	1.2%	

6.3.3.2 Local independence

Local independence as a requirement of the Rasch model is to ensure that item responses are determined only by the person's ability (Chachamovich et al., 2008;

Embretson & Reise, 2000; Linacre, 2009). Residual item correlations are computed for each pair of items, and the lower the correlation coefficient (ideally a coefficient of 0) the more independent the items are (Chachamovich et al., 2008; Embretson & Reise, 2000). In the current study, the correlation of residuals for all items was used to identify the pairs of items that were locally dependent, and the correlation coefficients equal to or higher than 0.3 were considered indicators of local dependence (Chachamovich et al., 2008; Linacre, 2009). Seven pairs of items were found to be within the range of local dependence. Most of these items were the ones intentionally included by the item developers, using the same items, but with the colours and answer positions changed.

Table 6.6

Residual item correlations

Correlation	Entry number item	Entry number item	Type of items
0.36	92	93	1 & 1
0.35	165	57	2 & 2
0.34	161	166	2 & 2
0.32	44	86	5 & 5
0.31	49	41	5 & 5
0.31	49	31	5 & 5
0.30	177	186	1 & 1

6.3.3.3 Person separation reliability

According to Moerdyk (2015), internal consistency reliability is a vital property of good items as it indicates the homogeneity of the items. As explained by Bond and Fox (2007), person separation reliability is used to determine the extent to which the person values would be reproduced if similar items were administered to the same

sample. The index is equivalent to Cronbach alpha reliability and uses the logits between -1 and +1 – hence the higher the value of the reliability index, the higher the replicability (Bond & Fox, 2007; Linacre, 2005; Mueller, Bullock & Leierer, 2010).

Furthermore, information on the overall indication of the fit of the persons to the model (MNSQ) is presented in the tables below. The recommendation for deciding on the misfit of items or persons is given for samples between 500 and 1 000, as MNSQ values less than 0.5 and greater than 1.2 (Bond & Fox, 2007; Linacre, 2012). However, the present study used the ranges of MNSQ values "less than 0.7 and greater than 1.30" as per the recommendation of Bond and Fox (2007, p. 310). For each of the *new item* types and all items, the raw scores, count, measure, model error, infit and outfit mean squares (MNSQ), and the standardised transformation of the mean squares (ZSTD) are presented below.

a New item type 1

The person reliability of the *new item* type 1, as indicated in table 6.7 below, is 0.86. Based on the values for assessing the model fit, the mean square (MNSQ) has values of 0.99 and 1.00, thus indicating that the persons show a good fit to the model.

Table 6.7

Summary statistics of measured persons for new item type 1

	Raw score	**Count**	**Measure**	**Model error**	**Infit**		**Outfit**	
Type 1 (N = 905)					**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**
Mean	20.07	35	0.42	0.41	0.99	0.0	1.00	-0.1
S.D.	7.27	0.9	1.12	0.08	0,13	0.8	0.27	0.9

Real RMSE = 0.43, Adj. SD = 1.04, Separation = 2.41, Person reliability = 0.85, Model RMSE = 0.42, Adj. SD = 1.04, Separation = 2.47, **Person reliability = 0.86**
Cronbach alpha (KR-20) Person raw score reliability **= 0.88**

b New item type 2

In table 6.8 below, the person reliability of the *new item* type 2 is indicated as 0.80. The mean square (MNSQ) has values of 0.98 and 1.13, thus showing that the persons show a good fit to the model.

Table 6.8

Summary statistics of measured persons for new item type 2

	Raw score	Count	Measure	Model error	Infit		Outfit	
Type 2 (N = 906)					**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**
Mean	19.32	34	0.35	0.43	0.98	0.0	1.13	0.1
S.D.	5.68	0.8	0.97	0.05	0.21	1.1	0.87	1.3

Real RMSE = 0.44, Adj. SD = 0.86, Separation = 1.94, Person reliability = 0.79, Model RMSE = 0.43, Adj. SD = 0.87, Separation = 2.03, **Person reliability = 0.80**
Cronbach alpha (KR-20) Person raw score reliability **= 0.82**

c New item type 3

The person reliability of the *new item* type 3 as indicated in table 6.9 below is 0.84. Based on the values for assessing the model fit, the mean square (MNSQ) has values of 0.97 and 1.08, thus indicating that the persons show a good fit to the model.

Table 6.9

Summary statistics of measured persons for new item type 3

	Raw score	Count	Measure	Model error	Infit		Outfit	
Type 3 (N = 905)					**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**
Mean	22.92	36	0.77	0.43	0.97	0.0	1.08	0.2
S.D.	6.59	0.8	1.09	0.08	0.20	1.0	0.61	1.3

Real RMSE = 0.45, Adj. SD = 0.99, Separation = 2.19, Person reliability = 0.83
Model RMSE = 0.44, Adj. SD = 1.00, Separation = 2.27, **Person reliability = 0.84**
Cronbach alpha (KR-20) Person raw score reliability **= 0.86**

d New item type 4

The person reliability for the *new item* type 4 as indicated in table 6.10 below is 0.82. Based on the values for assessing the model fit, the mean square (MNSQ) has values of 1.00 and 1.01, thus indicating that the persons show a good fit to the model.

Table 6.10

Summary statistics of measured persons for new item type 4

	Raw score	Count	Measure	Model error	Infit		Outfit	
Type 4 (N = 905)					**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**
Mean	17.53	31	0.37	0.43	1.00	0.0	1.01	0.0
S.D.	5.83	0.8	1.03	0.08	0.15	0.9	0.31	1.0

Real RMSE = 0.45, Adj. SD = 0.92, Separation = 2.04, Person reliability = 0.81
Model RMSE = 0.44, Adj. SD = 0.93, Separation = 2.12, **Person reliability = 0.82**
Cronbach alpha (KR-20) Person raw score reliability **= 0.83**

e *New item type 5*

The person reliability of the *new item* type 5 is indicated as 0.80 in table 6.11 below. The mean square (MNSQ) values of 1.00 and 1.10 also included in the table indicate a good fit to the model.

Table 6.11

Summary statistics of measured persons for new item type 5

	Raw score	Count	Measure	Model error	Infit		Outfit	
Type 5 (N = 906)					**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**
Mean	17.99	33	0.14	0.42	1.00	0.0	1.10	0.0
S.D.	5.89	0.9	0.97	0.05	0.19	1.0	0.80	1.2

Real RMSE = 0.44, Adj. SD = 0.86, Separation = 1.94, Person reliability = 0.79, Model RMSE = 0.43, Adj. SD = 0.87, Separation = 2.03, **Person reliability = 0.80**
Cronbach alpha (KR-20) Person raw score reliability **= 0.82**

f *New item type 6*

The person reliability of 0.81 for the *new item* type 6 is shown in the table 6.12 below. Based on the values for assessing the model fit, the mean square (MNSQ) has values of 0.99 and 1.00, thus indicating that the persons show a good fit to the model.

Table 6.12

Summary statistics of measured persons for new item type 6

	Raw score	Count	Measure	Model error	Infit		Outfit	
Type 6 (N = 903)					**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**
Mean	19.45	30	1.02	0.49	0.99	0.1	1.00	0.1
S.D.	5.5	0.7	1.15	0,12	0.16	0.8	0.46	1.0

Real RMSE = 0.52, Adj. SD = 1.03, Separation = 2.00, Person reliability = 0.80

Model RMSE = 0.50, Adj. SD = 1.04, Separation = 2.00, **Person reliability = 0.81**

Cronbach alpha (KR-20) Person raw score reliability **= 0.85**

g All items

The person reliability for all the *new items* as indicated in table 6.13 below is 0.96. Based on the values for assessing the model fit, the mean square (MNSQ) has values of 0.99 and 1.04, thus indicating that the persons show a good fit to the model.

Table 6.13

Summary statistics of measured persons for all new items

	Raw score	Count	Measure	Model error	Infit		Outfit	
Total items (N = 906)					**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**
Mean	117.45	199	0.48	0.17	0.99	0.0	1.04	0.2
S.D.	33.62	4.1	0.91	0.02	0.11	1.4	0.34	1.9

Real RMSE = 0.18, Adj. SD = 0.90, Separation = 5.10, Person reliability = 0.96
Model RMSE = 0.17, Adj. SD = 0.90, Separation = 5.10, **Person reliability = 0.96**
Cronbach alpha (KR-20) Person raw score reliability **= 0.97**

The overall indication from the tables 6.5 to 6.13 above is that the values of the person reliability for the *new items* as a total group and the different *new item* types range from 0.80 to 0.96. Acceptable values as per the ranges recommended by Anastasi and Urbina (1997) for reliability coefficients are 0.80 and 0.90. De Beer (2010) indicated ranges between 0.926 and 0.978 for the LPCAT, which has items that measure in a similar way to the new items. For the *new item* types, the values show a slightly lower reliability, which can be attributed to the small number of items per type, whereas the reliability for all the new items together is acceptable at 0.96 (Linacre, 2012).

6.3.3.4 Item separation reliability

In contrast to person separation reliability, item separation reliability is used to determine the extent to which the item parameter values would be reproduced if the same items were administered to another sample (Bond & Fox, 2007). According to Mayes et al. (2015), the item separation and reliability indices reflect the reproducibility of item characteristics with the typical desired levels of 0.9.

Information on the overall indication of the fit of the items to the model (MNSQ) is also been presented in tables 6.14 to 6.20 below. As in the case of person reliability, the recommendation for deciding on the misfit of items that was used for the present study was the ranges of MNSQ values less than 0.7 and greater than 1.30 as per the recommendation of Bond and Fox (2007, p. 310). For each of the *new item* types and all the items together, the raw scores, count, measure, model error, infit and outfit mean squares (MNSQ), and the standardised transformation of the mean squares (ZSTD) are included.

a New item type 1

The item reliability of the *new item* type 1 as indicated in table 6.14 below is 0.99. The fit statistics indicate that the values for the infit and outfit mean square (MNSQ) are 1.00 and 1.00, and the standardised fit statistics (ZSTD) have values of -0.3 and -0.2, thus suggesting that the items show a good fit to the model.

Table 6.14

Summary statistics of measured items for new item type 1

	Raw score	**Count**	**Measure**	**Model error**	**Infit**		**Outfit**	
Type 1 (n = 35)					**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**
Mean	518.5	898.7	0.00	0.08	1.00	-0.3	1.00	-0.2
S.D.	130.0	2.4	0.79	0.01	0.14	3.5	0.19	3.0
Real RMSE = 0.08, Adj. SD = 0.79, Separation = 9.71, Item reliability = .0.99								
Model RMSE = 0.08, Adj. SD = -.79, Separation = 9.74, Item reliability = 0.99								

b New item type 2

The item reliability indicated in table 6.15 below is 1.00. Based on the values for assessing the model fit, the mean square (MNSQ) values are 0.99 and 1.12, and the ZSTD values -0.3 and 0.4, thus indicating a good fit to the model for the *new item* type 2.

Table 6.15

Summary statistics of measured items for new item type 2

	Raw score	Count	Measure	Model error	Infit		Outfit	
Type 2 (n = 34)					**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**
Mean	514.9	898.8	0.00	0.08	0.99	-0.3	1.12	0.4
S.D.	194.9	2.4	1.24	0.01	0.12	3.1	0.57	4.1
Real RMSE = 0.09, Adj. SD = 1.23, Separation = 14.24, Item reliability = 1.00								
Model RMSE = 0.08, Adj. SD = 1.24, Separation = 14.57, Item reliability = 1.00								

c New item type 3

In table 6.16 below, the item reliability is shown as 0.99. Based on the values for assessing the model fit, the MNSQ values are 0.99 and 1.08, and the ZSTD values -0.2 and 0.2, thus indicating that the items show a good fit to the model.

Table 6.16

Summary statistics of measured items for new item type 3

	Raw score	Count	Measure	Model error	Infit		Outfit	
Type 3 (n = 36)					**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**
Mean	575.8	898.6	.00	0.09	0.99	-0.2	1.08	0.2
S.D.	172.6	2.8	1.12	0.01	0.12	2.9	0.52	3.6
Real RMSE = 0.09, Adj. SD = 1.11, Separation = 12.70, Item reliability = 0.99								
Model RMSE = 0.09, Adj. SD = 1.11, Separation = 13.00, Item reliability = 0.99								

d *New item type 4*

The item reliability of the *new item* type 4 is 0.99, as indicated in table 6.17 below. The values for the MNSQ are 1.00 and 1.00, and the values for the ZSTD 0.0 and 0.1, thus in terms of fit statistics indicating a good fit to the model.

Table 6.17

Summary statistics of measured items for new item type 4

	Raw score	Count	Measure	Model error	Infit		Outfit	
Type 4 (n = 31)					**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**
Mean	511.3	898.7	0.00	0.08	1.00	0.0	1.00	0.1
S.D.	141.7	3.4	0.88	0.01	0.09	2.9	0.17	2.7
Real RMSE = 0.08, Adj. SD = 0.88, Separation = 10.90, Item reliability = 0.99								
Model RMSE = 0.08, Adj. SD = 0.88, Separation = 11.09, Item reliability = 0.99								

e *New item type 5*

The item reliability of the *new item* type 5 as indicated in table 6.18 below is 0.81. Based on the values for assessing the model fit, the values for the MNSQ are 0.99 and 1.10, and for the ZSTD, -0.2 and 0.6, thus indicating that the items show a good fit to the model.

Table 6.18

Summary statistics of measured items for new item type 5

	Raw score	Count	Measure	Model error	Infit		Outfit	
Type 5 (n = 33)					**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**
Mean	478.6	898.0	0.00	0.08	0.99	-0.2	1.10	0.6
S.D.	180.4	1.9	1.12	0.01	0.11	2.9	0.46	3.8

Real RMSE =0.44, Adj. SD = 0.87, Separation = 1.97, Item reliability = 0.79
Model RMSE = 0.42, Adj. SD = 0.88, Separation = 2.07, Item reliability = 0.81

f *New item type 6*

The item reliability is indicated in table 6.19 as 0.99. The fit statistics values show the MNSQ values as 0.99 and 1.00, and the ZSTD values as 0.0 and 0.1, thus suggesting a good fit to the model for the type 6 *new items*.

Table 6.19

Summary statistics of measured items for new item type 6

	Raw score	Count	Measure	Model error	Infit		Outfit	
Type 6 (n = 30)					**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**
Mean	606.3	895.9	0.00	0.09	0.99	0.0	1.00	0.1
S.D.	129.7	3.5	0.91	0.01	0.11	2.7	.24	2.9
Real RMSE = 0.53, Adj. SD = 1.05, Separation = 10.24, Item reliability = 0.99								
Model RMSE = 0.51, Adj. SD = 1.06, Separation = 10.39, Item reliability = 0.99								

g All items

The item reliability of all the *new items* is indicated in Table 6.20 below as 0.99. For assessing the model fit, the MNSQ has values of 0.99 and 1.04, and the ZSTD values of -0.1 and 0.2, thus indicating that the items show a good fit to the model.

Table 6.20

Summary statistics of measured items for all new items

	Raw score	Count	Measure	Model error	Infit		Outfit	
All items (n = 199)					**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**
Mean	534.7	899.2	0.00	0.08	0.99	-0.1	1.04	0.2
S.D.	167.1	2.8	1.03	0.01	0.11	3.2	0.34	3.7
Real RMSE = 0.08, Adj. SD = 1.02, Separation = 12.31, Item reliability = 0.99								
Model RMSE = 0.08, Adj. SD = 1.02, Separation = 12.56, Item reliability = 0.99								

High values of item reliability indicate how consistent the items are and whether the item parameters can be obtained in the same way from another sample of participants (Bond & Fox, 2007). As indicated in tables 6.14 to 6.20, excluding table 6.18 for *new item* type 5, all the other calculations have values ranging from 0.99 to 1.00. All the *new item* separation indices have sufficiently high values, except for *new item* type 5, which has a lower separation of less than 3 (separation = 1.97) and item reliability of less than 0.8 (item reliability = 0.79) (Linacre, 2012).

6.3.3.5 Item-person map

According to Bond and Fox (2007), the item-person map displays the logit scale down the centre of the map (in equal intervals), with the respondents located on the left-hand side of the map, according to their standing on the latent trait, and the items located on the right-hand side of the map according to their level of measurement of the latent trait. As indicated in the graphs, the person ability and item difficulty increase in a vertically upward direction towards the top of the graph, and there is a decrease in the vertically downward direction towards the bottom of the graph. Lantano (2010) referred to the top and bottom of the graph as ceiling and floor effects to indicate a discrepancy when the items were not sufficient for the person abilities. The maps depicted in Figure 6.2 show the graphical results for each new item type and for all the new items. The same graphical illustration provided in Figure 6.2 is included in appendix C, presenting each graph in a larger more visible format.

a New item type 1

In Figure 6.2, the *new item* type 1, there seems to be a ceiling effect because there do not appear to be enough items to align with the person abilities at that level. Therefore, overall, there seems to be a higher ability distribution than the item difficulty distribution. Items 19 and 59 stand out at the top, but more items at that difficulty level would be required.

b New item type 2

The *new item* type 2 item and person map in Figure 6.2 indicates a spread of items at varying difficulty levels, but many gaps that do not align with person abilities. Item

165 stands out as very difficult with items 47 and 72 ranking next as difficult items. Items 68, 161, and 166 are at the bottom of the graph, indicating very easy items as they are below the person ability levels.

c New item type 3

The *new item* type 3 map illustrates a few outliers at both ends of the graph with items 57 and 98 indicating difficulty levels that are not well aligned with the rest of the person abilities, while items 3, 7, 66, and 81 are outliers at the bottom of the graph indicating that that they are easy items. A few gaps are also evident in the graph, indicating a need for more items to address those specific ability levels.

d New item type 4

Item 11 is indicated as a difficult item in Figure 6.2, while items 2, 8, 12 and 62 are indicated as easy items. Both the ceiling and floor effects were observed as there are gaps where person abilities are not addressed.

e New item type 5

This type of items has a number of items that are not aligned with person abilities, with item 61 standing out on its own on the top of the graph. Although there is some clustering in the middle, ranging between -1 and +1, there are gaps that would require the addition of items for this item type.

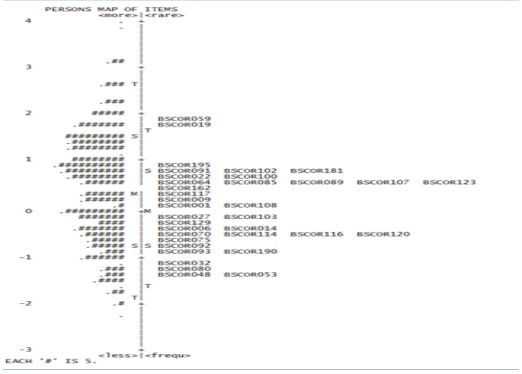
f New item type 6

Except for item 170, which seems to be addressing the higher ability level, most of the person abilities are not addressed by the items. Also, items 26, 29 and 54 were indicated as very easy as they are fairly far below the person ability levels of this sample. The *new item* type 6 observations illustrate both the ceiling and floor effects.

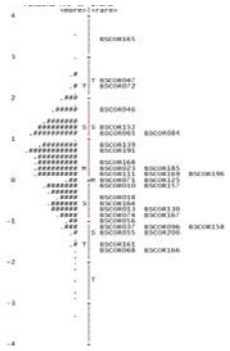
g All items

When looking at the new items as a whole, the map in Figure 6.2 shows a good distribution of difficulties and abilities. Again, the same easy and difficult items could be observed with 26, 29 and 54 indicated as the easiest items, while items 61 and 165 are shown as the most difficult items. This corresponds to the CTT indications provided by the *p*-values in Table 6.3.

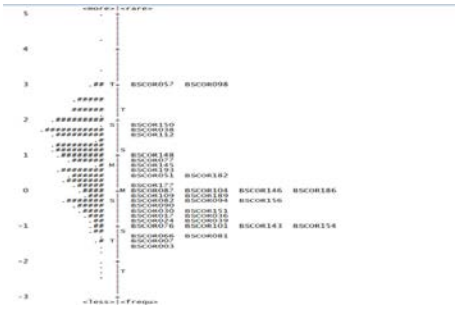
***New item* type 1**



***New item* type 2**

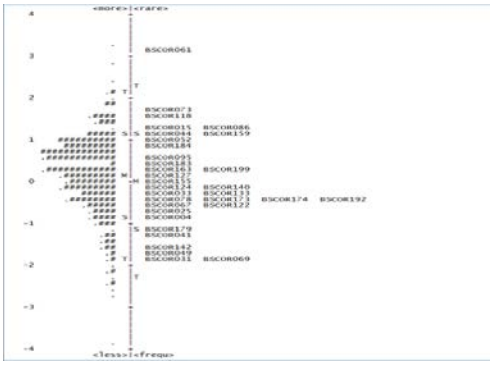


***New item* type 3**

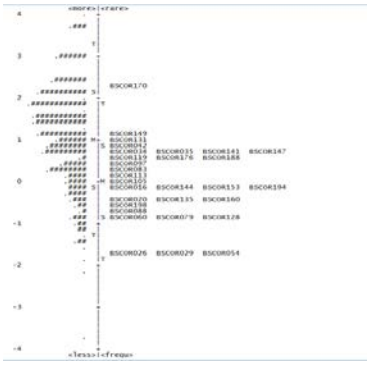


***New item* type 4**

***New item* type 5**



***New item* type 6**



All items

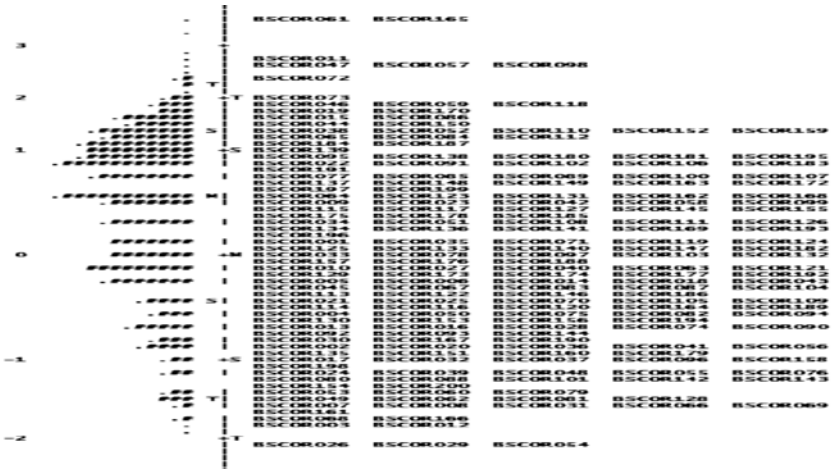


Figure 6.2. Person-item map for the new items

In the item-person map for all the items, one can observe items that are from the different item types that have the same difficulty levels that can possibly represent the same content areas, thus resulting in redundancy (Lantano, 2010). These item-person maps illustrate a good ability distribution that is within the range of the difficulty levels of the items, but, overall, the *new items* seem to be slightly easier for the participants.

6.3.3.6 Item infit and outfit

According to Linacre (2012), the infit and outfit statistics are key to how well individual items meet the assumption of unidimensionality. These are used to identify items that do not fit the model which, if included, could impact on the validity and reliability of the instrument (Bond & Fox, 2007). Fit statistics use unstandardised mean square (MNSQ) and standardised (ZSTD) forms as the main indicators (Bond & Fox, 2007; Linacre, 2012). According to Wright and Stone (1999), these misfitting items indicate the discrepancies between the model and observed data for items and persons. The interpretation of fit is based on overfit (outfit) values and underfit (infit) values, which are misfits because of possible poor items or items not being relevant to that specific set of items (Thompson & Barnard, 2009). Miyata (2007) explained infit and outfit items as unexpected responses that are indicated when people incorrectly respond to items which are estimated to be close to their abilities (infit) or when the people with low ability correctly respond to very difficult items (outfit). The presentation of the infit and outfit statistics in Tables 6.21 to 6.26 highlights the misfitting items – the full tables are in appendix D.

a New item type 1

For *new item* type 1, as shown in the Table 6.21, most of the 35 items were a good fit, except for three. Items 59 and 195 had outfit mean square values above 1.30, while item 93 had an outfit mean square value below 0.75. This indicates unexpected responses of persons to the items (Linacre, 2012).

Table 6.21

Summary of measured items: new item type 1

Item (n = 35)	**Infit**		**Outfit**	
	MNSQ	**ZSTD**	**MNSQ**	**ZSTD**
59	1.22	4.6	**1.46**	4.8
93	0.79	-5.5	**0.68**	-4.8
195	1.25	7.7	**1.30**	5.8

b New item type 2

New item type 2 items are indicated in the Table 6.22. Six out of the 34 items, namely 46, 47, 65, 72, 84 and 165, had outfit mean square values above 1.30, while three items, 55, 161 and 166, had outfit mean square values below 0.75. This indicates unexpected responses of persons to the items (Linacre, 2012).

Table 6.22

Summary of measured items: new item type 2

Item (n = 34)	**Infit**		**Outfit**	
	MNSQ	**ZSTD**	**MNSQ**	**ZSTD**
46	1.10	2.4	**1.55**	6.3
47	1.16	2.3	**2.41**	8.9
55	0.82	-4.0	**0.68**	-4.4
65	1.30	8.7	**1.58**	9.6
72	1.13	2.2	**1.85**	6.6
84	1.22	6.8	**1.33**	5.8
161	0.84	-2.6	**0.66**	-3.8
165	1.11	0.9	**3.69**	8.7
166	0.84	-2.5	**0.61**	-4.1

c New item type 3

As indicated in Table 6.23, five of the 36 *new item* type 3 items, namely 38, 51, 57, 98 and 150, had outfit mean square values above 1.30, while items 101, 143 and 151 had outfit mean square values below 0.75 This indicated unexpected responses of persons to the items (Linacre, 2012).

Table 6.23

Summary of measured items: new item type 3

Item (n = 36)	**Infit**		**Outfit**	
	MNSQ	**ZSTD**	**MNSQ**	**ZSTD**
38	1.17	4.6	**1.42**	6.0
51	1.26	7.9	**1.39**	8.1
57	1.30	4.2	**3.48**	9.9
98	1.20	2.9	**2.55**	8.5
101	0.85	-3.1	**0.70**	-3.5
143	0.86	-2.8	**0.71**	-3.2
150	1.12	3.3	**1.43**	5.8
151	0.83	-4.2	**0.72**	-4.0

d New item type 4

New item type 4 items are indicated in the Table 6.24. Two of the 31 items, namely 11 and 172, had outfit mean square values above 1.30, while two others, namely 12 and 62, had outfit mean square values below 0.75. This indicates unexpected responses of persons to the items (Linacre, 2012).

Table 6.24

Summary of measured items: new item type 4

Item (n = 31)	**Infit**		**Outfit**	
	MNSQ	**ZSTD**	**MNSQ**	**ZSTD**
11	1.12	1.6	**1.47**	3.3
12	0.88	-1.7	**0.70**	-2.5
62	0.85	-2.8	**0.67**	-3.4
172	1.24	8.4	**1.38**	8.4

e New item type 5

The unexpected responses of persons to items for *new item* type 5 items are indicated in the Table 6.25. Items 15, 44, 52, 61, 78, 86, and 159 had outfit mean square values above 1.30, while three items, namely 41, 49 and 142, had outfit mean square values below 0.75.

Table 6.25

Summary of measured items: new item type 5

Item (n = 33)	**Infit**		**Outfit**	
	MNSQ	**ZSTD**	**MNSQ**	**ZSTD**
15	1.07	2.0	**1.43**	5.9
41	0.79	-4.8	**0.68**	-4.9
44	1.09	2.5	**1.42**	6.0
49	0.82	-3.2	**0.69**	-3.5
52	1.10	3.0	**1.36**	5.6
61	1.10	0.9	**3.38**	8.0
78	1.23	7.1	**1.35**	7.2
86	1.10	2.8	**1.58**	7.8
142	0.78	-4.5	**0.65**	-4.7
159	1.13	3.9	**1.37**	5.6

f New item type 6

For the new item type 6 items indicated in the Table 6.26, only one of the 30 items, namely 170, had an outfit mean square value above 1.30, while four items, namely 79, 128, 144 and 198, had outfit mean square values below 0.75. This indicates unexpected responses of persons to the items (Linacre, 2012).

Table 6.26

Summary of measured items: new item type 6

Item (n = 30)	Infit		Outfit	
	MNSQ	**ZSTD**	**MNSQ**	**ZSTD**
79	0.78	-4.3	**0.65**	-3.6
128	0.83	-3.1	**0.68**	-3.1
144	0.83	-4.5	**0.73**	-4.0
170	1.26	5.9	**2.00**	9.8
198	0.85	-3.1	**0.74**	-3.0

6.3.4 Differential item analysis (DIF)

Based on the group representation as indicated in the description of the sample in Table 4.1, gender (female = 50%; male = 49% and 1% = missing values) was chosen for the DIF analysis of, as opposed to the other groupings of home language, education and province. The reason for selecting the latter subgroups was that they were not generally representative of the national biographical indicators. The gender DIF analysis was done by comparing the item difficulty measures of each gender, thus computing the differences between item difficulty measures for males and females (Linacre, 2012). Linacre (2012) highlighted the importance of the DIF contrast value and asserted that to indicate a meaningful difference, the value should be greater than 0.5 logits.

6.3.4.1 New item type 1

The analysis of DIF for the 35 new items of type 1 did not indicate any DIF contrast value greater than 0.5 logits, which meant there were no items of this type that showed item bias for the gender groups. The graph in Figure 6.3 shows similar scale functions for both gender groups. Items such as 6, 181 and 195 indicated some differences, but not enough to merit evidence of bias.

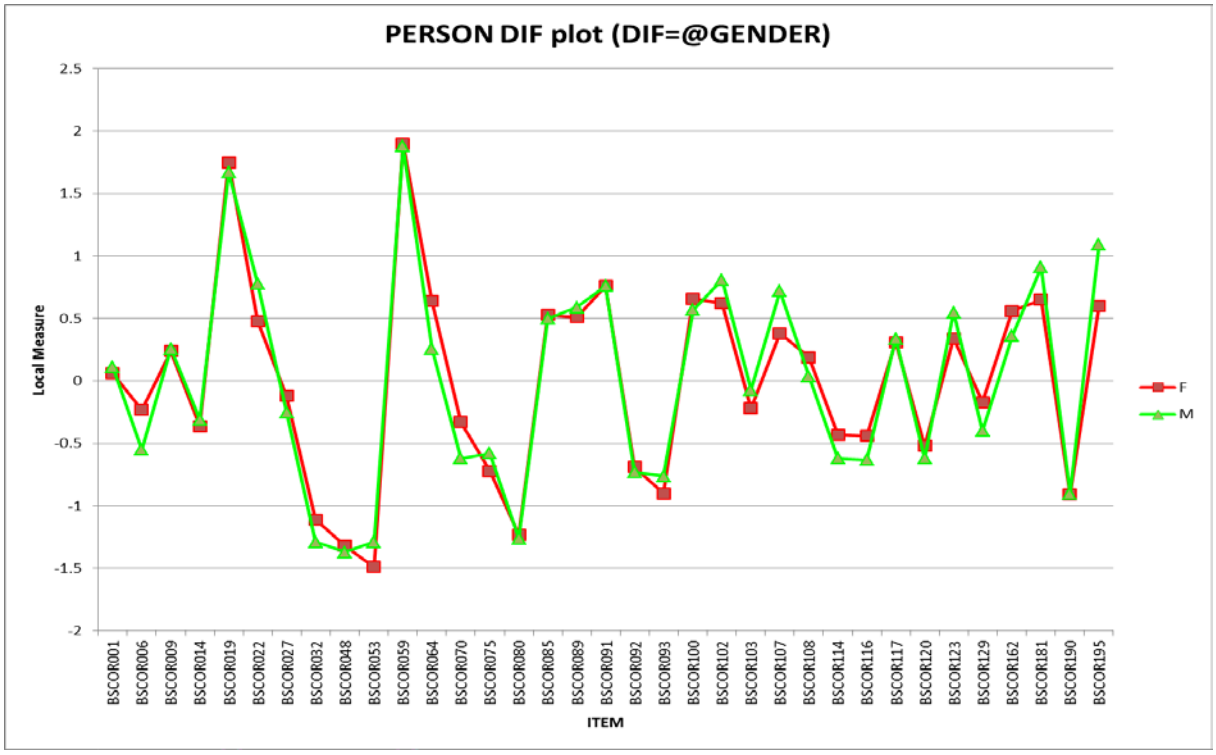


Figure 6.3. Gender DIF for new item type 1

6.3.4.2 New item type 2

Since all the items for the 34 *new items* type 2 had a DIF contrast value less than 0.5 logits, there does not appear to be any biased items between males and females for the 36 new items of type 2. It is evident from the figure below that females may have found item 72 easier than males, while items 161, 166 and 169 may have been more difficult for females.

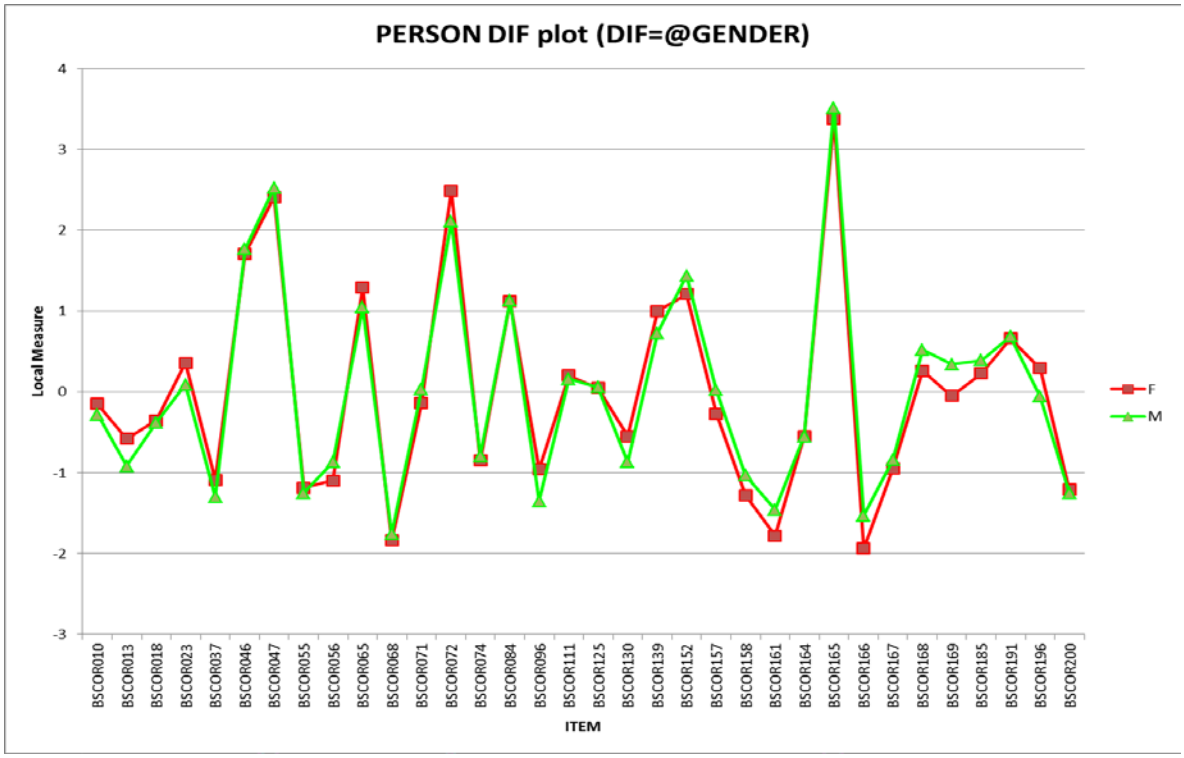


Figure 6.4. Gender DIF for new item type 2

6.3.4.3 New item type 3

This type of items also did not indicate any significant DIF contrast values, and in relation to the Figure 6.5, the scale pattern was similar for both gender groups for the 36 *new items* type 3. Two items that had visible discrepancies were 51 and 98, where the females found them easier than the males, although the DIF contrast values did not exceed 0.5 logits to indicate bias.

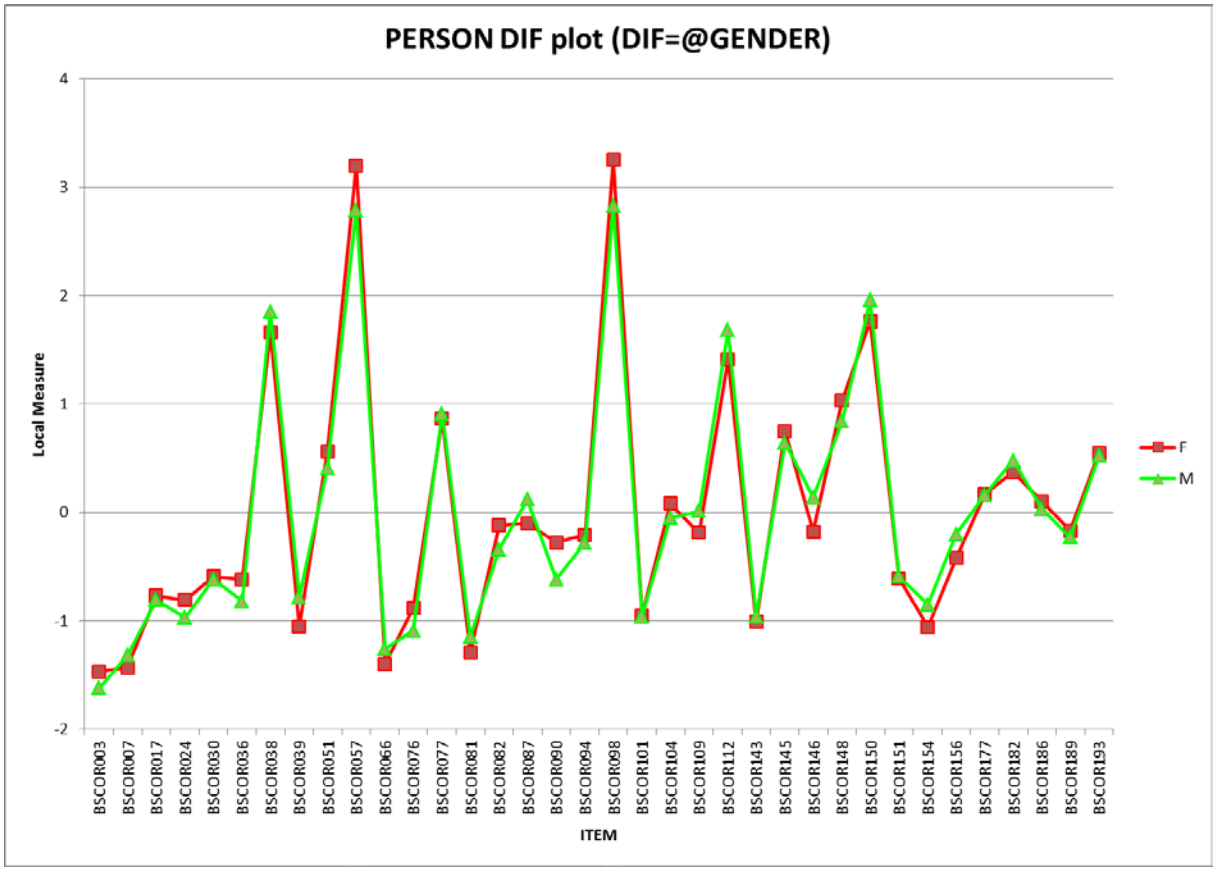


Figure 6.5. Gender DIF for new item type 3

6.3.4.4 New item type 4

The DIF contrast was larger than 0.5 logits for two of the 31 *new item type 4*, namely item 12 (DIF contrast = 0.51) and 43 (DIF contrast = 0.67). The males seemed to find these two items more difficult to answer than the females. Although the two items showed item bias, the rest of the items generally followed a similar pattern throughout, thus not showing any evidence of item bias.

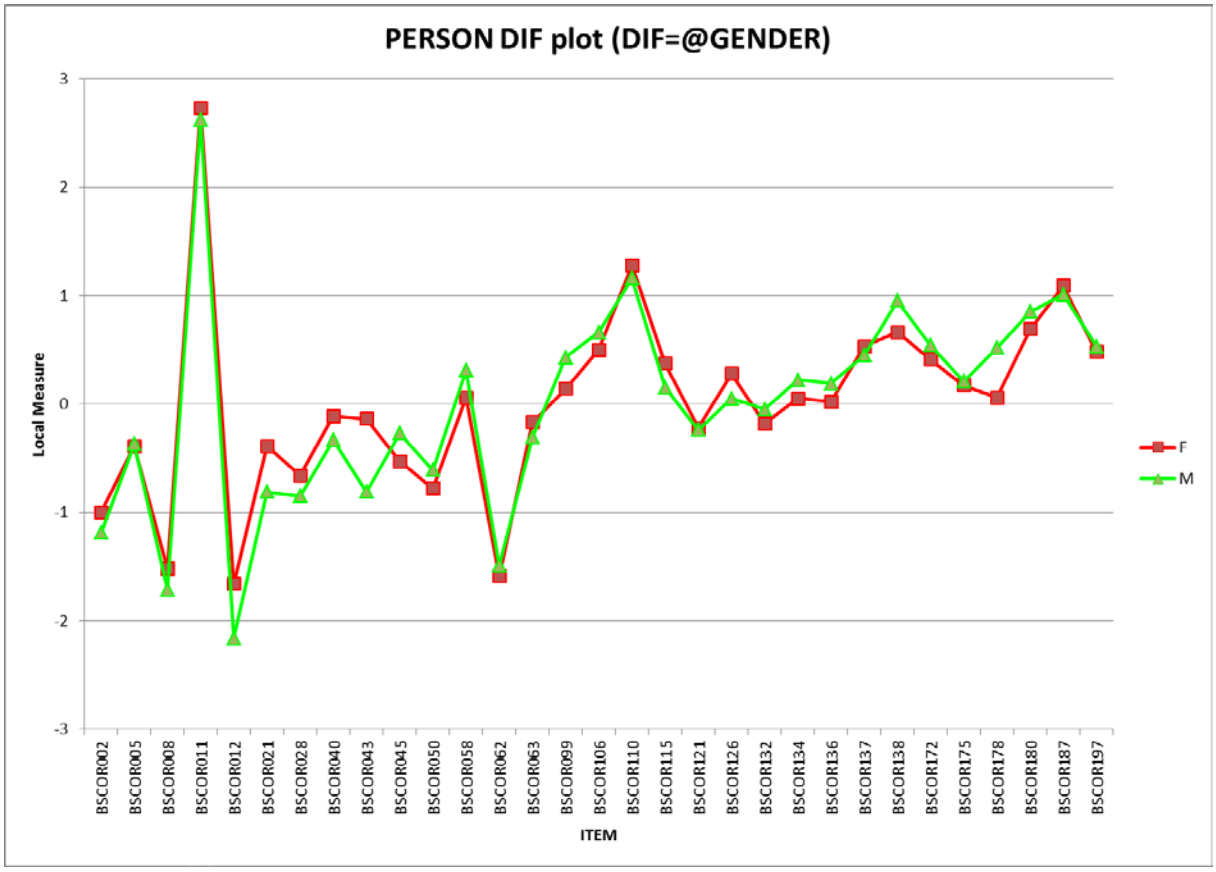


Figure 6.6. Gender DIF for new item type 4

6.3.4.5 New item type 5

Except for a few items, the graph depicted in Figure 6.7 below shows that the item scale patterns were noticeably similar for females and males. Some of the 33 *new items* type 5, namely 127, 133, 159 and 173 indicated a number of discrepancies, but not significant evidence from the DIF contrast values.

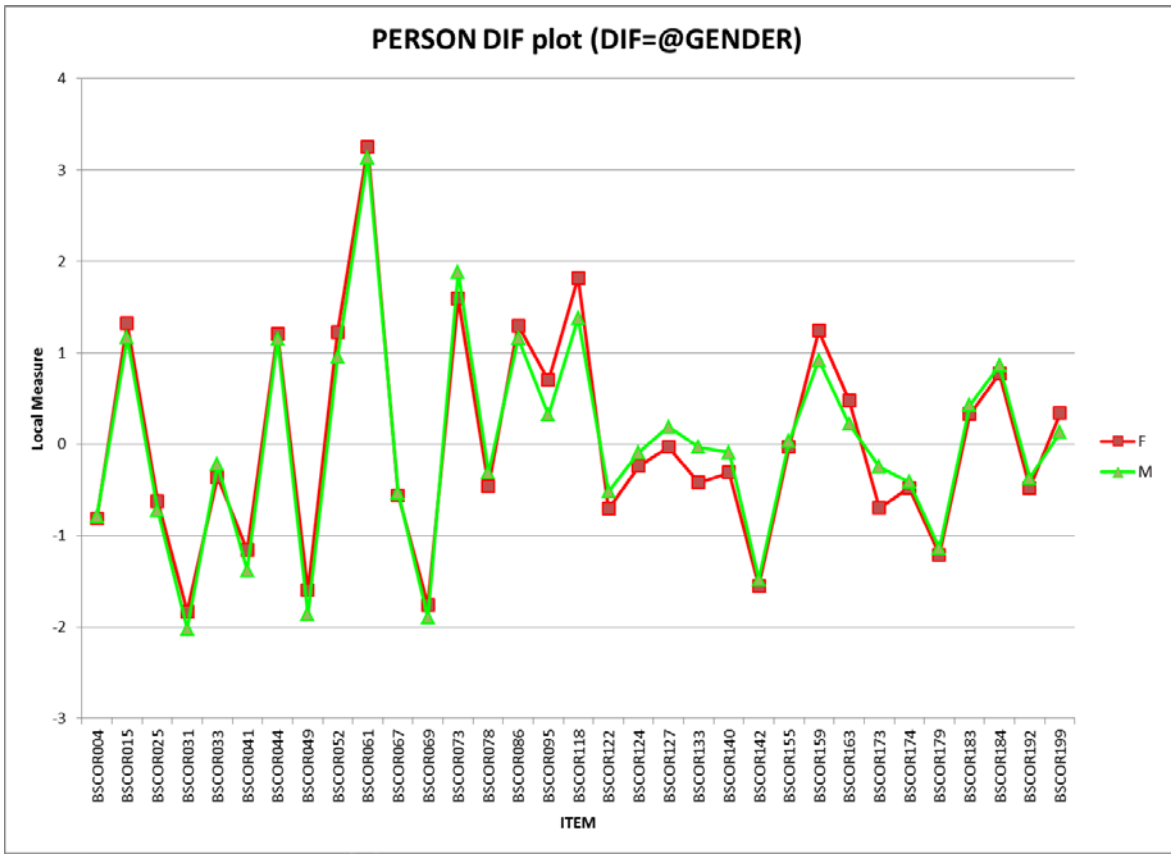


Figure 6.7. Gender DIF for new item type 5

6.3.4.6 New item type 6

The patterns of the item positions for the 30 *new items* type 6 are shown below. Most of them functioned in the same way for both gender groups. However, items 26 (DIF contrast = 0.84), 29 (DIF contrast = 0.71), 34 (DIF contrast = 0.58) and 149 (DIF contrast = -0.58) indicated a DIF contrast value greater than 0.5 logits, thus indicating bias for the gender groups.

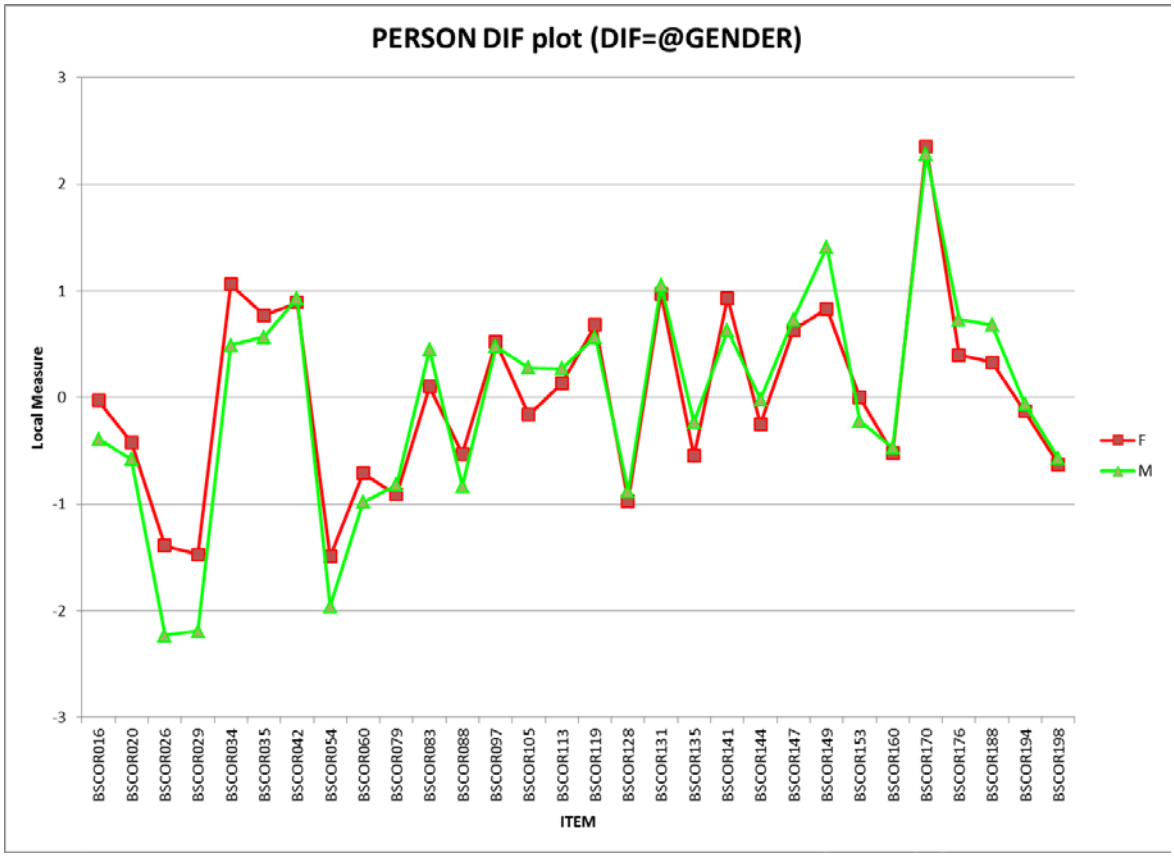


Figure 6.8. Gender DIF for new item type 6

Except for a few items highlighted above, most of the items did not show evidence of bias as they appeared to function in the same way for both genders. This means the responses to items are not dependent on what gender the participants are, but on their ability.

6.3.5 Correlation for construct identification

One of the ways of evaluating construct validity involves determining the relationship between two measures which both evaluate the same theoretical construct (Roodt, 2013c). For the current study, both the pre-test and post-test scores of the Learning Potential Computerised Adaptive Test (LPCAT) were used to determine the correlation with the scores of the six item types and the total score on all items for the new items. As indicated in Table 6.27 below, the correlations between the six new item types and the LPCAT results ranged from 0.522 to 0.592, while the correlation coefficients of the total items were 0.616 and 0.712, respectively, with the pre-test and post-test scores of the LPCAT. These values are deemed satisfactory because Moerdyk (2015) reported the acceptable levels of construct validity to be greater than 0.5.

6.3.6 Qualitative feedback results

Similar to the questions that were posed during phase 1 of the pilot item administration, the participants who completed the total group of new items were asked questions on which *new item* types they liked the most or the least, and why. The reasons cited were mainly based on the level of difficulty (challenge). What differed from the previous comments, was that the feedback under the general comments section, did not focus on the appearance of the *new items* in terms of their African representations – the comments related more to the content of the *new items* and how easy or difficult they seemed. With regard to nonverbal figural reasoning ability, the phrases used to describe it are the ability to understand complex concepts, analyse new information and solve problems using visual reasoning, identify relationships, similarities and differences between shapes and patterns, and recognise visual sequences and relationships between objects (Gregory, 2007; Lohman, 2005). These were similar to some of the phrases of the participants in the current study, that is: “make me think out of the box”; “find different ways of reaching solution”; “gives mind its rightful duty to think”; “be able to work with patterns”; and “take big problems and turn them into easy way”.

Table 6.27

Correlation results

		LPCAT_PRE	LPCAT_POST
LPCAT_PRE	Pearson correlation	1	.919**
	Sig. (2-tailed)		.000
	N	814	814
LPCAT_POST	Pearson correlation	.919**	1
	Sig. (2-tailed)	.000	
	N	814	814
TYPE 1	Pearson correlation	.592**	.685**
	Sig. (2-tailed)	.000	.000
	N	802	802
TYPE 2	Pearson correlation	.572**	.651**
	Sig. (2-tailed)	.000	.000
	N	802	802
TYPE 3	Pearson correlation	.563**	.653**
	Sig. (2-tailed)	.000	.000
	N	802	802
TYPE 4	Pearson correlation	.577**	.665**
	Sig. (2-tailed)	.000	.000
	N	802	802
TYPE 5	Pearson correlation	.522**	.600**
	Sig. (2-tailed)	.000	.000
	N	802	802
TYPE 6	Pearson correlation	.532**	.621**
	Sig. (2-tailed)	.000	.000
	N	802	802
TOTAL ITEMS	Pearson correlation	.616**	.712**
	Sig. (2-tailed)	.000	.000
	N	802	802

**Correlation is significant at the 0.01 level (2-tailed).

The participants' comments also captured the explanation by Kvist and Gustafsson (2008) where they described fluid intelligence as the capacity to solve novel and complex problems. The participants referred to patterns as "tricky", "mind puzzling", "twisting my head in many directions to finally come up with the answer", "exercised my brain by challenging and difficult questions" and "very mind opening and challenging".

Another comment of interest for one participant was "Green – like green grass; blue – is the sky; orange – like lemon; red – like blood; brown – like bread, brown bread". Although it was initially difficult to place this comment in the context of testing and assessment, it seems to have captured what Rosa and Orey (2010, p. 25) referred to in their study of ethnomathematics, in which because of ethnomathematics, students were motivated as they started to "recognise mathematics as part of their everyday life". They noted this as having led to better performance and deeper understanding of the subject (Rosa & Orey, 2010).

From the comments received from this sample, which were mostly positive, although they did indicate the challenging nature of the items, the results add value to the quantitative results, which indicated positive results and acceptable psychometric properties in terms of reliability and construct validity.

6.4 DISCUSSION

As noted by Meiring et al. (2005), the onus is on the psychologists to prove adherence to the regulations of the Employment Equity Act regarding reliability, validity, fairness and unbiased tests. Hence the purpose of the study was to develop and evaluate the viability and utility of the *new items* for cognitive assessment in the South African context.

The first phase of the research was devoted to the development of the new items, which were based on inspiration from African art and cultural artefacts and intended to measure nonverbal figural reasoning ability. Another aspect of importance in this

phase was to evaluate the viability of the *new items* for measuring cognitive ability in terms of their perceived culture fairness. The comments were positive and reflected possibilities in terms of how participants were likely to accept instruments of this kind.

In phase 2, the *new items* were evaluated for their utility in terms of their psychometric properties, meaning reliability and face and construct validity. CTT was used to determine the *p*-values, and the Rasch model to verify the unidimensionality of the new items, local independence and reliability. The construct validity of the *new items* was determined on the basis of their relationship with a similar previously standardised measure of the same construct.

Rasch analysis showed an overall indication of good fit to the model. The item separation reliability of all *new items* and *new item* types mostly showed fairly high values, thus indicating a relative order of item difficulty. Also, the high reproducibility of the test items was consistent along the estimated continuum. The results provided positive results for both the viability and utility of the new items – in support of further research and the refinement of these *new item* formats in a standardised full measure of nonverbal figural (fluid or *gf*) ability.

6.5 CHAPTER SUMMARY

True to the quotation at the beginning of the chapter, with Rasch analysis, the measurements have to fit the model – not the other way around. The results indicated above were shown to fit the model. The qualitative data was also encouraging because it indicated positive and appreciative comments from the participants. The questions formulated in chapter 4 were addressed. Chapter 7 deals with the conclusions, limitations and recommendations, based on the results of the study.

CHAPTER 7

DISCUSSION, CONCLUSIONS, LIMITATIONS AND RECOMMENDATIONS

It always *seems* impossible until it's done

Nelson Mandela

7.1 INTRODUCTION

The doctoral journey started with the ending in mind – in which the bigger picture was captured in chapter 1. The question of interest at this stage is whether the end is as it was envisioned in the beginning. This chapter is therefore used as a reflective mirror for the research, in which the purpose of the journey and how it unfolded are highlighted, and whether it could have been done differently. This is followed by a discussion of the contribution of the research study. Recommendations based on the results of the study are also made.

7.2 THE MAGIC CIRCLE MODEL

According to Trafford and Leshem (2008), the use of the magic circle model (figure 7.1 below) is a visual representation of the doctoral research journey, which depicts the interconnected and circular process of the doctorate research project. The cycle moves clockwise from the top towards the right starting from the gap in current knowledge and ending with the contribution to existing knowledge. Salkind (2014, p. 67) referred to a research process that "starts with a question and ends with asking new questions." This is true of research, because as a contribution is made to existing knowledge, recommendations for further research are suggested, thus revolving continuously within the magic circle.

Although the process is circular, the interconnections between the various parts can also move diagonally, as shown in figure 7.1 below. Here the research focus is linked to the research design; the research statement and questions are linked to the factual and interpretive conclusions respectively; and the conceptual framework is linked to the conceptual conclusions. Each of these will be discussed in this chapter.

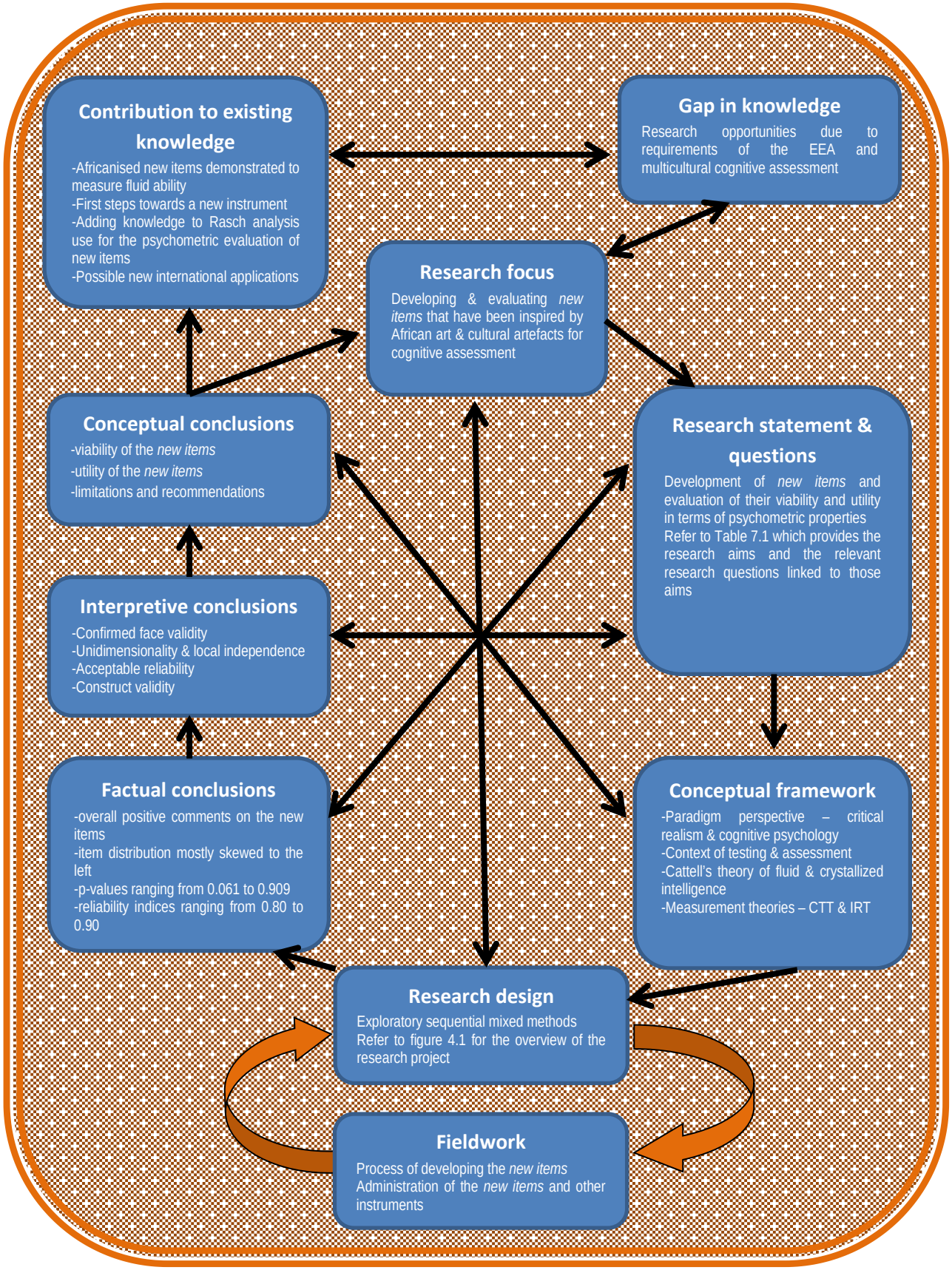


Figure 7.1. The magic circle (adapted from Trafford & Leshem, 2008, p. 170)

7.2.1 Gap in current knowledge

According to Trafford and Leshem (2008), a clear understanding of what the gap in current knowledge is, is essential. This was discussed in chapters 1 and 5 of the thesis. In chapter 1 (see section 1.2) the background to and rationale for the research were clarified, while in chapter 5 (see section 5.2) the discussion was specifically focused on explaining why there was a need to develop the *new items*. The motivation for the research included addressing the requirements of the Employment Equity Act (EEA), which promote adherence to reliable, valid, fair and unbiased tests (Meiring et al., 2005; Rothmann & Cilliers, 2007; Theron, 2007). The issue of the differences in test performance due to different socioeconomic backgrounds, schooling, rural versus urban differences and so on, was also raised and highlighting the need for the study. Furthermore, the continuous challenges of addressing fairness in psychological assessment and investigating the issues of language in testing and assessment were discussed. The debates around multicultural testing and assessment; transformation of tests and testing practices; and the need for creativity in dealing with testing and assessment challenges were some of the issues emphasised as the motivation for the study. The need for culture-fair instruments that can be used for different cultural and language groups is urgent, and the research on these issues needs to continue (Paterson & Uys, 2005; Rothmann & Cilliers, 2007; Schaap & Vermeulen, 2008).

Davison (2007, p. 143) used the quotation, “Mathematics is culture-free, but its contexts are not”, and this can be adapted to read as follows: “Testing and assessment should be culture-fair, but its context is not”. The context of testing and assessment in which the project was conducted was described in chapter 1. Figure 1.2, which depicted the macro, meso and micro levels of the context of testing and assessment, was discussed to show the interconnectedness of the context as significant. The context was described as encompassing both historical and current issues in which various research opportunities can be conceived with regard to culture, social, economic, political, legislative and transformative research issues. The complexities of addressing diversity and past imbalances were acknowledged as ongoing. On the basis of this context, a research opportunity was identified, and

on the strength of that, the problem statement, central research question and the aims of the study were formulated. These were dealt with in chapters 1 and 4.

7.2.2. Research focus, statement and questions

The research focus can be summarised according to the topic of the research thesis which is the development and evaluation of *new items* for multicultural cognitive assessment. The research statement is supposed to highlight the focus of the research with regard to what the researcher intends investigating (Trafford & Leshem, 2008). In this study, the two areas of interest were how the new nonverbal figural reasoning ability items were developed and how they were evaluated.

The general and specific aims of the study were outlined in chapter 1 and the research questions set out in chapter 4. The general aim of this research was to develop and evaluate the viability of *new items* inspired by African art and cultural artefacts and also to evaluate the utility of these *new items* in measuring general nonverbal figural reasoning ability. For ease of reference the aims were categorised in two phases of the research project, namely the development of the *new items* and the evaluation of the *new items*. In table 7.1 below, the aims in relation to the research questions that were used to address them are outlined.

Table 7.1

The aims and research questions addressed in the study

Aims	Research questions	Chapter
To develop items inspired by African art and cultural artefacts to measure general nonverbal figural reasoning ability	How is nonverbal figural reasoning ability defined?	Chapter 2
	How are African art and cultural artefacts defined and explored as inspiration?	Chapter 5
	How are the inspirations from African arts and cultural artefacts combined with nonverbal figural reasoning ability assessment principles?	
To evaluate the viability of the items in terms of their appropriateness in symbolising African art and cultural artefacts, specifically to determine the face and content validity of the items from a cultural perspective	How well does the appearance of the items represent the African art and cultural artefacts they are inspired by?	Chapter 5
	Are the items appropriate to use for colour blind persons?	
	Can the items be easily administered?	
	How is the final pool items constructed for the full psychometric evaluation process?	
To evaluate the utility of the items in terms of their psychometric properties	Do the results of the *new items* show acceptable levels of reliability?	Chapter 4
	Does gender impact on the responses to the *new items*?	
	Is there a statistically significant relationship between the total score obtained from the *new items* and another measure of general nonverbal figural reasoning using more traditional item formats (construct validity)?	
	Does the qualitative feedback from participants provide supportive evidence in terms of the face validity of the new items?	

The above aims were specific to the empirical study, but in order to proceed with the research, the theoretical (literature review) aims had to be addressed – hence the importance of chapters 1, 2 and 3 for the conceptual framework. The theoretical aims were as follows: (1) to conceptualise intelligence in order to provide an understanding of the theories, principles and debates regarding fluid reasoning ability measurement; and (2) to review the processes of test development and evaluation. For this study, appropriately addressing the theoretical aims was also aa vital consideration. Clear understanding of the construct of interest (general fluid (*gf*) ability – nonverbal figural reasoning ability) was essential for the *new items* to be more easily operationalised as items.

7.2.3 Conceptual framework

This part of the research process depicted in the magic circle focuses on a literature review in order to prepare the theoretical and conceptual basis for the study (Trafford & Leshem, 2008). The paradigm perspective was discussed to identify the framework for and boundaries within which the research would be conducted. The cognitive psychology paradigm was chosen to provide subject-related understanding for the study, while critical realism was used as the research paradigm. The cognitive psychology paradigm was important for providing the theoretical boundaries for cognitive ability and its assessment. The critical realism paradigm was discussed in chapter 1 where the domains of critical realism (real, actual and empirical) were explained.

Industrial psychology was identified as the field of study, specifically the psychometrics subfield for its systems and techniques necessary for the development and evaluation of the *new items*. Specific theories and models on multicultural cognitive assessment, item development and item evaluation were selected for use. Fluid intelligence (*gf*) was used as a theoretical base for the development of the African art and artefacts inspired *new items,* as the items were meant to measure nonverbal figural reasoning ability (Gregory, 2007). The steps used for the development of the *new items* were adapted from the generic test development process; and measurement theories such as classical test theory (CTT)

and the Rasch analysis model were used for the analytical purposes of the study. Chapter 1 also focused on the identification and definition of concepts and constructs such as intelligence, reliability and validity, fairness and bias, and item difficulty.

The aim of chapters 2 and 3 was to present the theoretical and research-based explanations and debates surrounding multicultural cognitive assessment and the criteria used to evaluate such assessments. In chapter 2, an overview of the measurement of cognitive functioning was provided which highlighted the early developments both internationally and locally. Intelligence was defined and a number of theories that provide a foundational understanding of nonverbal reasoning were discussed. Issues in multicultural ability testing such as culture, socioeconomic status, language and cognitive styles were highlighted. The role of acculturation in assessment was explained, whereby the different cultures interact continuously and changes are effected. The last section in chapter 2 focused on nonverbal figural reasoning tests which were developed as an alternative to language-based tests. This was necessary to gain an understanding of the construct the *new items* were intended to measure.

In chapter 3, the focus was on the criteria used to evaluate measures, which are guided by the EEA requirements. Before the discussion of the requisites of the Act, an overview of measurement was provided which highlighted the levels of measurement, measurement errors and measurement in practice. Reliability, validity, fairness and bias were explained, noting their use in practice and their relationship with one another and other concepts. Measurement theories such as CTT and IRT were discussed.

This information, as noted provided the theoretical base for the study and was essential for drawing conclusions.

7.2.4 Research design and fieldwork

The selected research design chosen for the study was the exploratory sequential mixed method design which appropriately catered for the development and

evaluation phases of the study. Both chapters 1 and 4 were used to provide information on the research design and methodology followed in the study. An interesting challenge in writing the thesis was to ensure that the processes and procedures followed in each of the two phases were reported on appropriately. Decisions on sampling were important for obtaining relevant and appropriate data from each participant or group of participants who provided information, either for drafting the *new items* or from the group of participants who responded to the final pool of items. Research procedure decisions to accommodate the mixed method research design were reported on in an effort to ensure the internal validity of the study. As illustrated by the double arrows in figure 7.1, the research design was intertwined with the research focus – in other words, whatever decisions, processes and procedures followed were all aimed at addressing the research focus. The link between the research focus and research design and fieldwork were discussed in chapters 4 and 5. The depiction of the overview of the research project in figure 4.1 (see chapter 4) provides a clear picture of the interconnectedness of these elements.

The exploratory sequential mixed method research design and the qualitative and quantitative processes used during the fieldwork to gather data were explicitly aligned towards addressing the research issue, so that the research can be deemed to have an acceptable internal empirical consistency (Trafford & Leshem, 2008).

7.2.5 Conclusions

As noted by Trafford and Leshem (2008), the factual, interpretive and conceptual conclusions are diagonally linked to the research statement, research questions and the conceptual framework respectively. According to Trafford and Leshem (2008), the factual conclusions should be seen as the descriptive micro level conclusions that are provided to indicate the findings of the research in terms of the results from the data that was gathered and analysed. This was followed by the presentation of the macro level conclusions to indicate the interpretation of those facts in relation to answering the research questions (Trafford & Leshem, 2008), based on the variables represented by the data obtained. Lastly, the conceptual conclusions for the study were drawn. Trafford and Leshem (2008) referred to this as a high meta-level of

thinking that entails a critical overview of the whole research with a focus on the relevance of the research findings.

7.2.5.1 Factual conclusions

According to Trafford and Leshem (2008), this section pertains to the research statement and should provide descriptive facts about the conclusions drawn from the data. No interpretation is expected at this stage.

a Phase 1 conclusions: Development of the new items

The first phase of the research was devoted to the development of the *new items* which were based on inspiration from African art and cultural artefacts and was intended to measure nonverbal figural (fluid or *gf*) reasoning ability.

Of significance in this phase was to evaluate the viability of the *new items* based on inspiration from African art and cultural artefacts for measuring cognitive ability in terms of their perceived culture fairness. The comments were positive with participants highlighting the similarities of the symbols used in the items to certain cultural artefacts as intended. The feedback also reflected how participants would welcome (accept) items of this kind. The cultural expert confirmed the cultural appropriateness of the *new items* and that they appropriately reflected the African art and cultural artefacts that had inspired them. The colour-blind person confirmed that the colours and patterns used in the items had prevented him from being able choose the correct responses to the *new items*.

Following the above feedback, the development of the *new items* was successfully achieved with 200 *new items* finalised for administration. As indicated earlier, 200 *new items* were developed, and in accordance with the sequential mixed method research design (Creswell, 2012; Onwuegbuzie & Johnson, 2006), these were used for further data collection in phase 2 of this study. The evaluation during this phase confirmed the viability of the *new items*.

b Phase 2 conclusions: Evaluation of new items

In phase 2, the 200 *new items* were evaluated for their utility in terms of their psychometric properties, as presented in chapter 6.

The **descriptive statistics** looked at the minimum and maximum values, means, standard deviations, skewness, and kurtosis of the distribution of the *new items*. The distribution of the *new items* was mostly skewed to the left, which meant the items were more likely to be easy.

This possible easiness of the *new items* was also indicated in the **item difficulty** values (*p*-values), which were determined as part of the CTT analysis. A high p-value indicates a high proportion or percentage of correct answers, whereas a low p-value indicates a low proportion or percentage of correct answers to the particular question. Anastasi and Urbina (1997) recommended that the mean item difficulty should be around 0.5; while Gregory (2007) suggested a specific range of 0.3 to 0.7. In this study, the *p*-values ranged between 0.06 (most difficult item) and 0.91 (easiest item). When grouping the *new items* according to item types, the mean p-values ranged between 0.53 (*new item* type 5) and 0.68 (*new item* type 6). The overall average item difficulty was 0.60. One should bear in mind that the average formal level of qualification of the sample group (grade 12 level) in this study was higher than that of the general South African population.

Unidimensionality and local independence, both of which are the basic assumptions of the Rasch model (Bond & Fox, 2007; Linacre, 2009) were found to be satisfactory. The analysis of the *new items* using principal component analysis indicated that only one latent trait was being measured (unidimensionality). The local independence of the *new items* was examined using the correlation of residuals of the new items. This was found to be acceptable (correlation coefficient less than 0.3) for most of the pairs of *new items*, except for seven pairs (items 92 and 93; 165 and 57; 161 and 166; 44 and 86; 49 and 41; 49 and 31; 177 and 186), which were found to some degree to be dependent on each other.

Reliability was determined using the Rasch analysis procedures which provide the person reliability index, the Cronbach alpha coefficient (α) and item reliability index. These are expressed in values ranging from 0.0 to 1.0 (Bond & Fox, 2007). A summary of the reliability indices is presented in table 7.2 below, where the reliability indices are shown to yield high reliability ranging between 0.80 and 1.00.

Table 7.2

Summary of reliability indices

New item type	Person reliability	Cronbach's alpha	Item reliability
Type 1 (n = 35)	0.86	0.88	0.99
Type 2 (n = 34)	0.80	0.82	1.00
Type 3 (n = 36)	0.84	0.86	0.99
Type 4 (n = 31)	0.82	0.83	0.99
Type 5 (n = 33)	0.80	0.82	0.81
Type 6 (n = 30)	0.81	0.85	0.99
All items (n = 199)	0.96	0.97	0.99

The **fit statistics** include the determination of the infit and outfit indices which express the extent to which the observed performance matches the expected performance (Bond & Fox, 2007). Items that were highlighted as misfits were three out of 35 items of *new item* type 1 (items 59, 93 and 195); nine out of 34 items for *new item* type 2 (items 46, 47, 55, 65, 72, 84, 161, 165 and 166); eight out of 36 items for *new item* type 3 (items 38, 51, 57, 98, 101, 143, 150 and 151); four out of 31 items for *new item* type 4 (items 11, 12, 62 and 172); ten out of 33 items for *new item* type 5 (items 15, 41, 44, 49, 52, 61, 78, 86, 142 and 159); and five out of 30 items for new item type 6 (items 79, 128, 144, 170 and 198). Notwithstanding the above, overall, the *new items* indicated a good fit to the model.

Differential item functioning (DIF) analysis recognised the values of the DIF contrast that were greater than 0.5 logits as indicating a meaningful difference (item bias). A few of the *new items* for *new item* types 4 (items 12 and 43) and 6 (items 26, 29, 34 and 149) exhibited evidence of bias as the DIF contrast for these items was greater than 0.5 logits. For the rest of the items, no significant DIF contrast was exhibited and the scale patterns were similar for both gender groups.

Construct validity was investigated by determining the correlation between the total score from the *new items* and another measure of general nonverbal figural reasoning using more traditional item formats (the Learning Potential Computerised Adaptive Test [LPCAT]). The correlation coefficients for the various scores of the *new item* types and the total score for all *new items* and total score of the pre- and post-tests of the LPCAT were found to be positive and moderate to strong in strength. The correlation values for the *new item* types and the LPCAT pre-test ranged from 0.522 to 0.592 while the correlation coefficients for the new item types and the LPCAT post-test ranged from 0.600 to 0.685. The correlation coefficients for the total score of all items and the pre- and post-tests were 0.616 and 0.712 respectively.

7.2.5.2 Interpretive conclusions

The research was structured in two phases which were based on three main aims (see table 7.1). The aims were the basis for all the research questions and processes that were addressed and followed in this research. According to Tafford and Leshem (2008), the interpretive conclusions are supposed to show the internal theoretical consistency of the research.

a Phase 1 conclusions: Development of the new items

The aims of phase 1 were to develop items from inspirations of African art and cultural artefacts to measure nonverbal figural reasoning ability and to evaluate the viability of these items in terms of their perceived cultural fairness.

The feedback confirmed that the symbols used in the new items successfully captured the Africanness of the art and cultural artefacts. The familiarity of the content used in the *new items* was highlighted in the feedback in which participants could relate the symbols to their lives – this was seen as encouraging to further develop the *new items*. This study can be related to the studies of ethnomathematics. Although it is acknowledged that general fluid ability testing is not the same as the learned ethnomathematics, the principles used in developing the content of the items can be viewed as the same. In their research of ethnomathematics, Lipka and Adams (2004) and Wong and Lipka (2011) found that introducing familiar everyday content to the tasks resulted in the students feeling connected to the subject and perceiving the experience as increasing accessibility. Hence the usage of cultural artefacts such as bead work, painting and other creative designs for mathematics tasks of geometry (Eglash, Krishnamoorthy, Sanchez & Woodbridge, 2011; Davison, 2007; Wong & Lipka, 2011) is similar to the use of African cultural art and artefacts as inspirations that were used for the development of the *new items* in the current study. This also means the *new items* addressed the reflections of Maree (2010), who advocated the inclusion of more Afrocentric content as opposed to the Eurocentric content, which is perceived to be used in the majority of tests in South Africa. According to the feedback in this study, the *new items* appeared to take the African perspective into account.

The cultural expert employed in this study had concerns about whether all the cultural groups would be represented equally in the designs of the *new items.* Eglash et al. (2011) acknowledged the impossible task of capturing the African culture in its entirety, but were clear that overall, the creations and designs used in their study of fractal simulations were sufficiently and appropriately representative.

It is evident that the validation process in this study was continuous and central to the development and evaluation of the *new items* (Zumba, 1999; Zumbo et al., 2002). At the outset of phase 1, the development process began by clarifying the purpose of the *new items* and throughout the item development phase, the focus was on ensuring that the *new items* adhered to the expectations of that purpose. The

results indicated positive support for the face validity and multicultural viability of the *new items.*

The exploratory activities of developing the *new items* provided vital information that confirmed the viability of the *new items* in terms of their appropriateness in representing the African art and cultural artefacts that were used as inspiration, thus endorsing the face validity of the *new items*. Based on this positive feedback, the full bank of 200 *new items* was developed and administered to a larger sample of 946 participants.

b Phase 2 conclusions: Evaluation of new items

The aim of this phase was to evaluate the utility of the *new items* in terms of their psychometric properties, which included reliability, validity and bias.

The Rasch model bases the probability of success of a person on an item as depending on the **item difficulty** and person ability (Bond & Fox, 2007; Edwards & Alcock, 2010; Wu & Adams, 2007). In this study, the item difficulty was first determined using the *p*-values of the CTT – after adopting the boundaries provided by Gregory (2007) of 0.3 to 0.7 as acceptable. Based on this, the *p*-values indicated a few items on the extreme ends of easy and difficult. The *new item* type that indicated more difficulty than the other types was *new item* type 5. This is one of the new item formats depicted by the triangle (chapter 5, figure 5.7, for an example of this *new item* type). The item difficulty is also addressed by the item and person map which also indicated some of the *new items* as outliers either at the top or at the bottom of the graphs. One should note the gaps which are evident in the graphs (see figure 6.2 of the person item map), which could imply a need for more items that would cater for higher levels of person ability (Lantano, 2010; Mueller et al., 2010).

Unidimensionality and local independence as assumptions of the Rasch model (Bond & Fox, 2007; Linacre, 2009) were found to be satisfactory. It should be noted that unidimensionality was investigated in various analyses in this study, including principal component analysis, fit statistics and the item- person maps. In all of these, it was found that the *new items* measure one latent trait.

The **reliability** estimates as determined by the person reliability index, the Cronbach alpha coefficient and the item reliability index, indicated satisfactory reliability levels. This means that if the same sample were to be given similar items, then similar results would be obtained. Also, if the items were given to a similar sample, then similar results could be expected (Bond & Fox, 2007; Mueller et al., 2010). According to Mueller et al. (2010), the reliability of all *new items* together and each of the *new item* types respectively generally showed fairly high values, therefore indicating a relative order of item difficulty and the high reproducibility of the test items was consistent along the estimated continuum. The results provided positive support for both the viability and utility of the new items in support of further research and refinement of the *new items*.

The Rasch **fit statistics** analysis showed an overall indication of good fit to the model, which means that the new items did meet the assumption of unidimensionality (Bond & Fox, 2007; Linacre, 2012; Mueller et al., 2010). The infit and outfit statistics indicated relatively few items and persons which were found to show misfit. A case in point would a person (based on his/her ability) who is expected to answer an item correctly but, does not, and vice versa. Hence such items could be seen as being too predictable or too unpredictable – both of which would be outside the fit of the Rasch model (Bond & Fox, 2007). Although these items were highlighted and reported as showing misfit, thus flagged as problematic, for the purposes of the current study no further analysis was done. However, in future studies, the impact of these items could be investigated further.

Gender related **differential item functioning** (DIF) was investigated and except for six items that were found to exhibit bias, the other 193 items did not indicate any significant differences between the gender groups.

Construct validity which was determined using the Learning Potential Computerised Adaptive Test (LPCAT), was found to be satisfactory. The new items were thus confirmed to measure fluid ability using nonverbal figural reasoning ability items. These results therefore confirmed the utility of the *new items* in terms of reliability and validity.

The **qualitative feedback** included at this stage also confirmed the face validity of the new items in terms of the content of the questions from the larger group of participants responding to the full set of items.

7.2.5.3 Conceptual conclusions

With reference to figure 7.1, it is clear that all the arrows end up contributing to the conceptual conclusions this section thus consolidates the whole research project. The presentation of these conclusions will therefore take the reflective journey around the magic circle with reference to the research problem; conceptual framework; research design and conclusions.

a Research problem

When defining the problem in chapter 1, various arguments were presented to highlight the gaps in the current knowledge and to describe the nature of the research problem. Some of these included the acknowledgement of the history of assessing cognitive functioning, which goes back centuries, and contentious issues such as the effects of language and socioeconomic background on test performance, the use of test results, the culture fairness of the tests and so on, that have been ongoing for years (Claassen, 1997; Nzimande, 1995; Van de Vijver & Rothmann, 2004). It was because of such issues that multicultural assessment moved forward, and the development of new instruments continued to be an option for addressing the challenges (Claassen, 1997; Maree, 2010; Van de Vijver & Rothmann, 2004). For the multicultural South African context where socioeconomic and educational differences are still prevalent, the use of nonverbal figural content for testing cognitive ability was regarded as a more culture-fair testing format (De Beer, 2005; Paterson & Uys, 2005). Other researchers were even more specific by recommending the use of Afrocentric content, as opposed to Eurocentric content, that is often used in tests (Maree, 2010), and investigations of different methods or item formats for the South African context (Foxcroft, 2004; Maree, 2010; Paterson & Uys, 2005; Rothmann & Cilliers, 2007). The main aim of the present study was to develop and evaluate the new items that had been developed from African art and cultural artefacts inspiration, for the assessment of general fluid ability, specifically

nonverbal figural reasoning ability. The *new items* had new item types which did not confine the questions to the traditional item formats of figure series, figure analogies, 2x2 and 3x3 matrices, but added the use of circles and triangles.

The promulgation of the EEA (Government Gazette, 1998) was also highlighted as one of the recent turning points in the South African psychological testing and assessment field. Its introduction increased research opportunities for validation studies of existing instruments, the development of new instruments and improvements in testing practices (Van de Vijver & Rothmann, 2004). The results of the research by Paterson and Uys (2005) highlighted the requirement for new instruments to specifically indicate evidence of predictive validity, cross-cultural fairness, relevance and reliability. Similar sentiments were expressed in the research by Rothmann and Cilliers (2007) as they had also recommended research on reliability and validity, as well as the equivalence and bias of the assessment tools. The current study did take the EEA into consideration because the viability and utility of the items were investigated in terms of the psychometric properties, specifically using the Rasch model to analyse the data collected. According to Ding (2014), use of this model can provide construct related evidence and sample independent deductions which are important for the South African context.

In her presidential address during her tenure as president of the American Psychological Association, Vasquez (2012) committed to the following five focus areas: reducing discrimination and enhancing diversity; addressing educational disparities; focusing on applying science and practice to the advancement of society; addressing the challenges of the changing demographics and applying knowledge to address the grand challenges of society. These emphasise the similarities between the issues of interest internationally and those of South Africa. As mentioned earlier, South Africa is a country moving from a discriminatory era to democracy – hence the focus areas she mentioned in an international context also correspond to South African challenges regarding diversity, socioeconomic and educational differences, rural versus urban people differences, et cetera (Claassen, 1997; De Beer, 2004; Kanjee, 2006; Nzimande, 1995; Theron, 2007). If the challenges are the same internationally, then the innovations and recommendations for this study could be

adopted in other international context, thus also highlighting the possible international applicability of this study.

b Conceptual framework

The challenges and issues of concern for multicultural cognitive assessment cut across time as shown in the discussions of the periods between 1890 to the present day (see chapter 2, section 2.2) with reference to the earlier test developments and the challenges of validation that were experienced (Anastasi & Urbina, 1997). Early test developers such as Binet and Simon were already sorting items based on difficulty levels and based on how well the test taker progresses to the more difficult items (Binet, 1916; Moerdyk, 2015). According to De Beer (2006; 2007), such processes heralded the advent of IRT. The current study used Rasch analysis, which recognises item difficulty as a key element to be used together with person ability to determine the probability answering the item correctly.

As the use of tests progressed, issues of culture, socioeconomic status, language, educational background, et cetera, became matters of concern in the field of testing and assessment (Gregory, 2007). The purpose of the innovations effected by the Army Alpha and Beta tests, resulting in the inclusion of nonverbal content items, was to address the language, translation and illiteracy challenges (Anastasi & Urbina, 1997; Gregory, 2007). One should bear in mind that the questions and issues of concern raised during those early years still apply today, particularly in the diverse South African context. The diversity of the South African context has a history of its own that was marred by discrimination – but the turning point came after the 1994 democratic elections. According to Schaap and Vermeulen (2008), with socio-political circumstances changing, trends in testing also had to change and cross cultural issues received more attention. Furthermore, after the promulgation of the EEA, it was clear that the testing and assessment field had to be reinvented. More importance was placed on research that empirically investigated test bias, fair use of tests for all cultural groups and validation studies of existing tests for different culture and language groups (Foxcroft, 2004; Foxcroft & Aston, 2006; Paterson & Uys, 2005; Theron, 2007; Van de Vijver & Rothmann, 2004). The current study addressed all these issues through the development and evaluation of the *new items*. The use

of African art and cultural artefacts has been shown to allay some of the anxieties created by traditional tests, and is a unique contribution of the current study as such items have never been developed or used before.

Addressing the requirements of the EEA means investigating reliability, validity, fairness and bias. To that end, one needs to start with knowing what trait or construct is being measured. Defining general fluid intelligence was thus deemed important for this study. Cattell (1963), on whose theory the general fluid ability measurement is based, described general fluid ability as an adaptation to new situations, problem solving, pattern recognition and abstract reasoning. These are the kind of measurements that are generally deemed to be culture fair as they do not include language or previously learned information as part of the items (Lohman, 2005). The *new items* were developed using this understanding as a guide to defining the construct while using the African art and artefacts as inspiration.

The criteria of evaluation of a measure are based on the EEA requirements – these are the psychometric properties (reliability, validity, fairness and bias) that were investigated in the study. The results indicated that these were acceptable, thus practically addressing issues of adherence to standards for psychometric properties, culture-fairness in cognitive assessment (Claassen, 1997; Foxcroft, 2004; Rothmann & Cilliers, 2007) and pleas for more Afrocentric content (Beets & Le Grange, 2005; Maree, 2010; Rothmann & Cilliers, 2007).

c Research design

The overview of the research project was illustrated in figure 4.1 which indicated a clear alignment of the research focus and research design. As noted by Trafford and Leshem (2008), this is where the internal empirical consistency is evidenced.

d Conclusions

As per the aims of the study, items that are meant to measure nonverbal figural reasoning ability were developed using African art and cultural artefacts as inspiration. The *new items* were analysed on the basis of the combination of CTT and the Rasch model to determine the psychometric properties. The *new items* are

based on the Cattell theory of fluid intelligence (*gf*) which focuses on abilities that are not based on previous learning. Hence the *new items* address culture fairness because they will not be limited by language. The reliability and validity of the new items were found to be acceptable. The gender related DIF analysis showed that both gender groups generally had similar response patterns.

7.2.6 Limitations of the study

Although the results were encouraging in their indication of the viability and utility of the *new items*, there were several limitations in this study.

7.2.6.1 Phase 1 limitations: Development of the new items

The samples used for phase 1 were small, which is in line with the sampling strategies of the qualitative method (Salkind, 2014). However, the views of only one colour-blind person could be regarded as a limitation. Nevertheless, the person who participated in the study is totally colour-blind, and his views were therefore of considerable value. Soliciting the views of people with different levels of colour blindness could, however, provide additional evidence from a colour-blind perspective.

7.2.6.2 Phase 2 limitations: Evaluation of the new items

The convenience sample used for phase 2 was large enough – between the acceptable range of 500 to 1000 recommended by Bond and Fox (2007). However, the limitation was that the sample was neither representative of the national population, nor randomly selected. According to Kgosana (2012), this is a common limitation in many studies, thus leading to results with limited generalisability.

According to Moerdyk (2015), the extent to which measures predict performance is important. This was a limitation for this study as there was no criterion data (such as examination results) available to evaluate the concurrent or predictive validity of the *new items*.

The determination of content validity was not done as meticulously as explained by Roodt (2013c), where a panel of experts is required to evaluate each item. The content of the *new items* was based on a literature review and a subject expert in the area of test development. Furthermore, Wu and Adams (2007) argued that the Rasch model provides information for both content and construct validity through its unidimensionality assumption. However, as noted earlier, soliciting more views and comments would broaden the spectrum of information to reflect on. More extensive content validity evaluation could still be done when this study is extended, possibly in a bid to construct a new measure.

Although valuable information was obtained from the qualitative questions included in the questionnaires of both phase 1 and phase 2, more could have been done to ensure optimal use of this source of data. For example, regarding the questions enquiring about the items the participants liked most or least, the responses were generally based on the difficulty of the items. Having worked with the feedback in phase 1, the open-ended questions for phase 2 could have been phrased differently to gain better and more extensive information and explanations.

7.2.7 Contribution to existing knowledge

Augustyn and Cillié (2008) challenged the quality of academic research, specifically in the field of industrial psychology, and questioned the contributions that such research makes to both the science and practice of the field. This study addressed both of these. Scientifically, because the processes followed to develop and evaluate the *new items* add value to the body of knowledge of psychometrics. Practically, the *new items* are envisaged for use in in exploring how best to test people within a multicultural context.

Augustyn and Cillié (2008) further called for research that is needs-driven or problem-oriented, in which real and practical issues would be addressed. Similarly, Vasquez (2012) highlighted that science and practice must be applied for the advancement of society, which means using psychological knowledge to address the challenges of a particular society. The issues of fairness and bias are real in the

diverse South African context therefore the creation of the new items provided the opportunity of exploring other possibilities of item content and item format types; and the results indicated the viability and utility of such creations.

The novelty of practically using the African art and cultural artefacts in conjunction with the scientific process to develop items that test general fluid ability is the first study of its kind in South Africa. This study has shown the vast resources available in the social context that can be creatively used for the benefit of the industrial and psychological field of study. Through the development of these *new items*, the study demonstrated practically the manifestation of combining art and science (Ambrose & Anstey, 2010; Gregory, 2007).

The viability and utility of the *new items* adds value to the debates of Africanisation that have been raised in various fields of study (Andersson, 2010; Beets & Le Grange, 2005; Louw, 2009; Msila, 2009). In this study, the concept of Africanisation was embraced in the development of the *new items* and as a theme for the research, which aligns to the philosophy of inquiry and scholarship alluded to by Andersson (2010).

Theoretically this study has also contributed to the explanation and implementation of newer statistical methods, specifically the Rasch analysis model, for assessing psychometric properties during the development of instruments. This should add to the body of knowledge on measurement theories.

7.2.8 Recommendations

Despite all the limitations, the *new items* were shown to be viable and usable. The feedback confirmed that the *new items* represent African art and cultural artefacts and that they have the required psychometric properties for measuring nonverbal figural reasoning ability. Further development and investigations building on the current research towards a new instrument are therefore recommended.

The present study did not analyse all the data elements at its disposal – for example, the items where the question mark positions were shifted; the misfitting items as indicated by the Rasch analysis; the colour used in the items; and so on. These are some of the research questions that could be explored in future studies.

The present study used convenient sampling methods thus limiting the representativeness of the sample for generalisation of the results. Also, this limited possibilities of more extensive analysis for bias, such as different culture groups, language groups, provincial representation, educational levels and age groups.

The successful development of the *new items* opens further avenue for cognitive assessment, specifically in multicultural environments. Therefore, future research on the *new item* types such as the triangle and the circle could be further explored so that item formats can be extended beyond the figure series, figure analogies and matrices that have been traditionally used for nonverbal figural reasoning ability tests.

The *new items* already lend themselves to more computerised testing – hence further investigations are recommended on the administrative format that can be followed for these types of items. Since technology is changing all the time, the opportunities for additional testing are endless.

During the collection of the photos and exploration of the art and cultural artefacts, the similarities, some subtle, and some obvious, between African art and cultural artefacts and those of other indigenous art and artefacts around the world were noted (see figure 7.2 below). This indicates the potential opportunities for exploring indigenous art and cultural artefacts in further development of items, which could be used for international test development.

Figure 7.2. Examples of global indigenous art (Adapted from Bekwa & De Beer, 2015)

7.3 IN CLOSING

Envisioning the end from the beginning may seem too overwhelming a task, even impossible, as the quotation at the start of the chapter indicates. However, with clear aims, research boundaries and frameworks, and set strategies, the bumps along the journey do not shake the direction of the compass towards the end. The study has clearly taken the first steps in exploring culture fair instruments in a context that is not yet and may for a long time still continue not to be.

REFERENCES

Abad, F. J., Colom, R., Rebollo, I., & Escorial, S. (2004). Sex differential item functioning in the Raven's Advanced Progressive Matrices: Evidence of bias. *Personality and Individual Differences, 36*, 1459–1470.

Abedalaziz, N. (2010, August). A gender-related differential item functioning of Mathematics test items. *The International Journal of Educational and Psychological Assessment, 5*, 101–116.

Adogamhe, P. G. (2008, July). Pan-Africanism revisited: Vision and reality of African unity and development. *African Review of Integration, 2*(2), 1–34.

Allalouf, A. (2004, April). *Improving second language proficiency assessment: A differential item functioning study*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, California.

Amabile, T. M. (1997). Motivating creativity in organisations: On doing what you love and loving what you do. *California Management Review, 40*(1), 39–58.

Amabile, T. M. (1998). *How to kill creativity* (pp. 77–87). Boston, MA: Harvard Business School Publishing.

Amabile, T. M., Barsade, S. G., Mueller, J. S., & Staw, B. M. (2005). Affect and creativity at work. *Administrative science quarterly, 50*(3), 367–403.

Ambrose, D. M., & Anstey, J. R. (2010). Questionnaire development: Demystifying the process. *International Management Review, 1*(6), 84–91.

Anastasi, A. (1992, May-June). Tests and assessment: What counselors should know about the use and interpretation of psychological tests. *Journal of Counseling and Development, 70*, 610–615.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Andersson, M. (2010). Settling scores: A reading of the managerial vision of transformation at Unisa 2004-2010. *International Journal of Arts and Sciences, 3*(4), 250–301.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Thayer (Eds), *Differential item functioning* (pp. 397–418). Hillsdale, NJ: Lawrence Erlbaum.

Archer, M. (1995). *Realist social theory: The morphogenetic approach*. Cambridge: Cambridge University Press.

Ariffin, S. T., Idris, R., & Ishak, N. M. (2010). Differential item functioning in Malaysian Generic Skills Instrument (MyGSI). *Jurnal Pendidikan Malaysia, 35*(1), 1–10.

Augustyn, J. C. D., & Cillié, G. G. (2008). Theory and practice in Industrial Psychology: Quo vadis? *South African Journal of Industrial Psychology, 34*(1), 70–75. Retrieved from http://www.sajip.co.za

Austin, J. (2013). Displaying data. In C. Tredoux, & K. Durrheim (Eds.), *Numbers, hypotheses and conclusions: A course in statistics for the social sciences* (2nd ed., pp. 18–39). Cape Town: UCT Press.

Austin, D. M., & Sanders, C. (2007). Graffiti and perceptions of safety: A pilot study using photographs and survey data. *Journal of Criminal Justice and Popular Culture, 14*(4), 292–316.

Azeez, O. A. (2011). Environmental sculptures: An artist's view. *Global Journal of Human Social Science, 3*(11), 60–64.

Babbie, E. (1989). *The practice of social research* (5th ed.). Belmont, California: Wadsworth.

Babbie, E., & Mouton, J. (2010). *The practice of social research* (SA edition). Cape Town: Oxford University Press Southern Africa.

Bacharach, S. B. (1989). Organizational theories: Some criteria of evaluation. *Academy of management: The Academy of Management Review, 14*(4), 496–515.

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation. Madison, WI: University of Wisconsin.

Balluerka, N., Gorostiaga, A., Gómez-Benito, J., & Hidalgo, M. D. (2010). Use of multilevel logistic regression to identify the causes of differential item functioning. *Psicothema, 22*(4), 1018–1025.

Bardaglio, G., Settanni, M., Marasso, D., Musella, G., & Ciairano, S. (2012). The development and Rasch calibration of a scale to measure coordinative motor skills in typically developing children. *Advances in Physical Education, 2*(3), 88–94. http://dx.doi.org/10.4236/ape.2012.23016

Baumgärtner, S., Becker, C., Frank, K., Müller, B., & Quaas, M. (2008). Relating the philosophy and practice of ecological economics: The role of concepts, models, and case studies in inter- and transdisciplinary sustainability research. *Ecological Economics, 67*, 384–393. Doi: 10.1016/j.ecolecon.2008.07.018

Bazeley, P. (2004). Issues in mixing qualitative and quantitative approaches to research. In R. Buber, J., J. Gadner, & L. Richards (eds.), *Applying qualitative methods to marketing management research* (pp. 141–156). Palgrave: Macmillan.

Beaujean, A. A., & Osterlind, S. J. (2008). Using item response theory to assess Flynn effect in the national longitudinal study of youth 79 children and young adults data. *Intelligence, 36*, 455–463. Retrieved from www.sciencedirect.com

Beets, P., & Le Grange, J. (2005). 'Africanising' assessment practices: Does the notion of ubuntu hold any promises? *South Africa Journal of Higher Education, 19*, 1197–1207.

Benjamin, L. T., & Baker, D. B. (2004). *From séance to science: A history of the profession psychology in America*. Belmont, CA: Wadsworth/Thomson Learning.

Bergh, Z. (2009). Fields of study and practice areas in industrial and organisational psychology. In Z. C. Bergh, & A. L. Theron (Eds), *Psychology in the work context (*4th ed., pp. 16–30). Cape Town: Oxford University Press South Africa.

Berk, L. (2000). *Child development* (5th ed.). Boston, MA: Allyn & Bacon.

Bhaskar, R. (1978). *A realist theory of science* (2nd ed.). Hassocks, Sussex, UK: Harvester Press.

Bhaskar, R. (1989). *Reclaiming reality: A critical introduction to contemporary philosophy*. London: Verso.

Bhaskar, R. (1998). *The possibility of naturalism* (3rd ed.). London: Routledge.

Bikic, V., & Vukovic, J. (2010). Boardgames reconsidered: Mancala in the Balkans. *Issues in Ethnology and Anthropology, 5*(1), 183–209.

Binet, A., & Simon, T. (1916). *The development of intelligence in children: The Binet-Simon Scale* (No. 11). Williams & Wilkins Company.

Blair, C. (2006). How similar are fluid cognition and general intelligence? A developmental neuroscience perspective on fluid cognition as an aspect of human cognitive ability. *Behavioural and Brain Sciences, 29*, 109–160.

Bond, T. G. (2003). Validity and assessment: A Rasch measurement perspectives. *Metodología de las Ciencas del Comportamiento, 5*(2), 179–194.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New York: Routledge.

Bors, D. A., & Forrin, B. (1995). Age, speed of information processing, recall, and fluid intelligence. *Intelligence, 20*, 229–248.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071.

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology, 53*, 371–399.

Brizuela-Garcia, E. (2006). The history of Africanization and the Africanization of history. *History in Africa, 33*, 85–100. Doi: 10.1353/hia.2006.0007.

Brouwers, S. A., Van de Vijver, F. J. R., & Hemert, D. A. (2009). Variation in Raven's Progressive Matrices scores across time and place. *Learning and Individual Differences, 19*, 330–338.

Buanes, A., & Jentoft, S. (2009). Building bridges: Institutional perspectives on interdisciplinarity. *Futures, 41*, 446–454.

Cameron, R. (2011). Mixed methods: The five Ps framework. *The Electronic Journal of Business Research Methods, 9*(2), 96–108. Retrieved from www.ejbrm.com

Carroll, K. K. (2010). A genealogical analysis of the worldview framework in Africa-centered psychology. *Journal of Pan African Studies, 3*(8), 109–134.

Carspecken, P. F. (1996). *Critical ethnography in educational research: A theoretical and practical guide*. New York & London: Routledge.

Caruth, G. D. (2013, 1 August). Demystifying mixed methods research design: a review of the literature. *Mevlana International Journal of Education (MIJE), 3*(2), 112–122. http://dx.doi.org/10.13054/mije.13.35.3.2

Cascio, W. F. (1987). *Applied psychology in personnel management* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*(1), 1–22.

Chachamovich, E., Fleck, M. P., Trentini, C., & Power, M. (2008). Brazilian WHOQOL-OLD Module version: a Rasch analysis of a new instrument. *Rev Saúde Pública, 42*(2), 308–316.

Christensen, L. B. (2001). *Experimental methodology* (8th ed.). Boston, MA: Allyn & Bacon.

Čisar, P., & Čisar, S. M. (2010). Skewness and kurtosis in function of selection of network traffic distribution. *Acta Polytechnica Hungarica, 7*(2), 95–106.

Claassen, N. C. W. (1997). Cultural differences, politics and test bias in South Africa. *European Review of Applied Psychology, 47*(4), 297–307.

Clark, M. (2007). *Listening placement test development and a analysis from a Rasch perspective*. Unpublished doctoral dissertation, Manoa, Honolulu: University of Hawaii.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practices, 17*, 31–44.

Cohen, B., & Murphy, G. L. (1984). Models of concepts. *Cognitive Science, 8*, 27–58.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). New York: Routledge.

Cohen, R. J., & Swerdlik, M. E. (2002). *Psychological testing and assessment: An introduction to test and assessment* (5th ed.). Boston: McGraw-Hill.

Coleman, M. A. (2006). *Construct validity evidence based on internal structure: Exploring and comparing the use of Rasch measurement modelling and factor analysis with a measure of student motivation*. Unpublished doctoral thesis. Virginia Commonwealth University, Richmond, Virginia. Retrieved from http://scholarscompass.vcu.edu/etd

Crane, P. K., Van Belle, G., & Larson, E. B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine, 23*, 241–256.

Creswell, J. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Upper Saddle River, NJ: Pearson Education.

Creswell, J. W., & Garrett, A. L. (2008). The “movement” of mixed methods research and the role of educators. *South African Journal of Education, 28*, 321–333.

Creswell, J. W., Klassen, A. C., Plano Clark, V. L., & Smith, K. C. (2011). *Best practices for mixed methods research in the health sciences.* Commissioned by the Office of Behavioural and Social Sciences Research (OBSSR). National Institutes of Health. Retrieved from http://obssr.od.nih.gov/mixed_methods_research

Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). Advanced mixed methods research designs. In A. Tashakkori, & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioural research* (pp. 209–240). Thousand Oaks, CA: Sage.

Cronholm, S., & Hjalmarsson, A. (2011). Experiences from sequential use of mixed methods. *The Electronic Journal of Business Research Methods, 9*(2), 87–95. Retrieve from www.ejbrm.com.

Cruickshank, I., & Mason, R. (2003). Using photography in art education research: A reflexive inquiry. *International Journal of Art and Design Education, 22*(1), 5–22.

D'Ambrosio, U. (2001). What is ethnomathematics, and how can it help children in schools? *Teaching Children Mathematics, 7*(6). Reston, VA: National Council of Teachers of Mathematics.

D'Ambrosio, U. (2004). Peace, social justice and ethnomathematics. *The Montana Mathematics Enthusiast*, 25–34.

D'Ambrosio, U. (2006, May). The program ethnomathematics: A theoretical basis of the dynamics of intra-cultural encounters. *The Journal of Mathematics and Culture, 1*(1), 1–7.

Das, J. P., Naglieri, J. A., & Murphy, D. (1995). Individual differences in cognitive processes of planning: A personality variable? *Psychological Records, 45*, 355–371.

Davison, D. M. (2007). In what sense is it true to claim that mathematics is culture-free. *Mathematics in a Global Community*, 139–143.

Dawes, A. (1998). Africanisation of psychology: Identities and continents. *Psychology in Society (PINS), 23*, 4–16.

De Beer, M. (2000). *The construction and evaluation of a dynamic computerised adaptive test for the measurement of learning potential.* Unpublished D. Litt et Phil dissertation. Pretoria: University of South Africa.

De Beer, M. (2004). Use of differential item functioning (DIF) analysis for bias analysis in test construction. *SA Journal of Industrial Psychology, 30*(4). 52–58.

De Beer, M. (2005). Development of the learning potential computerised adaptive test (LPCAT). *SA Journal of Psychology, 35*(4), 717.

De Beer, M. (2006). Dynamic testing: Practical solutions to some concerns. *SA Journal of Industrial Psychology, 32*(4), 8–14.

De Beer, M. (2007). Use of CAT in dynamic testing. In D.J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*. Retrieved from www.pysch.umn.edu/psylabs/CATCentral/

De Beer, M. (2010). Longitudinal predictive validity of a learning potential test. *Journal of Psychology in Africa, 20*(2), 225–232.

De Bruin, G. P., Hill, C., Henn, C. M., & Muller, K. P. (2013). Dimensionality of the UWES-17: An item response modelling analysis. *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde, 39*(2), Art. #1148, 8 pages. http://dx.doi.org/10.4102/sajip.v39i2.1148

De Klerk, M., Nel, J. A., Hill, C., & Koekemoer, E. (2013). The development of the MACE work-family enrichment instrument. *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde, 39*(2), Art. #1147, 16 pages. http://dx.doi.org/10.4102/sajip.v39i2.1147

De Kock, F., Kanjee, A., & Foxcroft, C. (2013). Cross-cultural test adaptation, translation and tests in multiple languages. In C. Foxcroft, & G. Roodt (Eds), *Introduction to psychological assessment in the South African context* (4th ed., pp. 83–106). Cape Town: Oxford University Press Southern Africa.

De Kock, F., & Schelechter, A. (2009). Fluid intelligence and spatial reasoning as predictors of pilot training performance in the South African Air Force (SAAF). *Journal of Industrial Psychology, 35*(1), 31–38. Doi: 10.4102/sajip. v35i1.753. Retrieved from http://www.sajip.co.za

De la Rey, C., & Ipser, J. (2004). The call for relevance: South African psychology ten years into democracy. *SA Journal of Psychology, 34*(4), 544–552.

De Lisle, J. (2011). The benefits and challenges of mixing methods and methodologies: Lessons learnt from implementing qualitative led mixed methods research designs in Trinidad and Tobago. *Caribbean Curriculum, 18,* 87-120.

Ding, L. (2014). Seeking missing pieces in science concept assessment: Reevaluating the brief electricity and magnetism assessment through *Rasch analysis. Physical Review Special Topics – Physics Education Research, 10.* Doi: 10.1103/PhysRevSTPER.10.010105.

Donald, F., Thatcher, A., & Milner, K. (2014). Psychological assessment for redress in South Africa organisations: Is it just? *SA Journal of Psychology, 44*(3), 333–349.

Donaldson, P. J. (2001). Using photographs to strengthen family planning research. *International Family Planning Perspectives*, 148–151.

Doyle, L., Brady, A., & Byrne, G. (2009). An overview of mixed methods research. *Journal of Research in Nursing, 14*(2), 175–185. Doi: 10.1177/1744987108093962

Durrheim, (2006). Research design. In M. Terre Blanche, K. Durrheim, & D. Painter, (Eds), *Research in practice: Applied methods for the social sciences,* (2nd ed., pp. 33–59). Cape Town: University of Cape Town Press.

Durrheim, K., & Painter, D. (2006). Collecting quantitative data: sampling and measuring. In M. Terre Blanche, K. Durrheim, & D. Painter (Eds.), *Research in practice: Applied methods for the social sciences* (2nd ed., pp. 131–159). Cape Town: University of Cape Town Press.

Eastwood, J. G., Jalaludin, B. B., & Kemp, L. A. (2014). Realist explanatory theory building method for social epidemiology: A protocol for a mixed method multilevel study of neighbourhood context and postnatal depression. *SpringerPlus, 3*(12). Doi: 10.1186/2193-1801-3-12

Edwards, A., & Alcock, L. (2010). Using Rasch analysis to identify uncharacteristic responses to undergraduate assessments. *Teaching Mathematics and its applications, 29*, 165–175. Doi:10.1093/teamat/hrq008

Effendi, K. & Hamber, B. (2006). Publish or perish: Disseminating your research findings. In M. Terre Blanche, K. Durrheim, & D. Painter (Eds.), *Research in practice: Applied methods for the social sciences* (2nd ed., pp. 112–128). Cape Town: University of Cape Town Press.

Egbo, B. (2005). Emergent paradigm: Critical realism and transformational research in educational administration. *McGill Journal of Education, 40*(2), 267 – 284.

Eglash, R. (1997). When math worlds collide: Intention and invention of ethnomathematics. *Science, Technology and Human Values, 22*(1), 79–97.

Eglash, R., Krishnamoorthy, M., Sanchez, J., & Woodbridge, A. (2011). Fractal simulations of African design in pre-college computing education. *ACM Transaction Computing Education, 11*(3). http://doi.acm.org/10.1145/2037276.2037281

Eglash, R., & Odumosu, T. B. (2005). Fractals, complexity, and connectivity in Africa. In G. Sica (Ed.), *What mathematics from Africa?* Monza, Italy: Polimetrica International Scientific Publisher.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Einarsdóttir, J., & Rounds, J. (2009). Gender bias and construct validity in vocational interest measurement: Differential item functioning in the Strong Interest Inventory. *Journal of Vocational Behaviour, 74*, 295–307.

Eysenck, H. J. (1982). Introduction. In H. J. Eysenck (Ed.), *A model for intelligence* (pp. 1–10). New York: Springer-Verlag.

Eysenck, H. J. (1988). The concept of intelligence: useful or useless. *Intelligence, 12*, 1–16.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357–374.

Fancher, R. E. (1985). *The intelligence men: Makers of the IQ controversy.* New York: Norton.

Fourie, P. J. (2005, July). The last word: The “Africanisation” of communication studies. *Communicare, 24*(1), 171–176.

Foxcroft, C. D. (2004). Planning a psychological test in the multicultural South African context. *SA Journal of Industrial Psychology, 30*(4), 8–15.

Foxcroft, C. (2013). Developing a psychological measure. In C. Foxcroft, & G. Roodt (Eds), *Introduction to psychological assessment in the South African context* (4th ed., pp. 70–81). Cape Town: Oxford University Press Southern Africa.

Foxcroft, C. D., & Aston, S. (2006). Critically examining language bias in the South African adaptation of the WAIS-III. *SA Journal of Industrial Psychology, 32*(4), 97.

Foxcroft, C., Paterson, H., Le Roux, N., & Herbst, D. (2004). *Psychological assessment in South Africa: A need analysis. The test use patterns and needs of psychological assessment practitioners*. Final Report. HRSC. Retrieved from www.hrsc.ac.za

Foxcroft, C., & Roodt, G. (2013a). An overview of assessment: Definition and scope. In C. Foxcroft, & G. Roodt (Eds.), *Introduction to psychological assessment in the South African context* (4th ed., pp. 3–8). Cape Town: Oxford University Press Southern Africa.

Foxcroft, C. & Roodt, G. (2013b). What the future holds for psychological assessment. In C. Foxcroft, & G. Roodt (Eds), *Introduction to psychological assessment in the South African context* (4th ed., pp. 287–302). Cape Town: Oxford University Press Southern Africa.

Foxcroft, C., Roodt, G., & Abrahams, F. (2013a). Psychological assessment: A brief retrospective overview. In C. Foxcroft, & G. Roodt (Eds), *Introduction to psychological assessment in the South African context* (4th ed., pp. 9–27). Cape Town: Oxford University Press Southern Africa.

Foxcroft, C., Roodt, G., & Abrahams, F. (2013b). The practice of psychological assessment: Controlling the use of measures, competing values, and ethical practice standards. In C. Foxcroft, & G. Roodt (Eds), *Introduction to psychological assessment in the South African context* (4th ed., pp. 109–124). Cape Town: Oxford University Press Southern Africa.

Franke, B., & Esmenjaud, R. (2008). Who owns African ownership? The Africanisation of security and its limits. *SA Journal of International Affairs, 15*(2), 137–158. Doi: 10.1080/10220460802614486

Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement, 47*(4), 432–457.

Freng, S., Freng, A., & Moore, H. A. (2007). Examining American Indians' recall of cultural inclusion in school. *Journal of American Indian Education, 46*(2), 42–61.

Fulop, N., & Robert, G. (2015). *Context for successful quality improvement: Evidence review.* London: The Health Foundation.

Furr, R. M., & Bacharach, V. C. (2008). *Psychometrics: An introduction*. Thousand Oaks, CA: Sage.

Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences.* London: Macmillan.

Geranpayeh, A. (2008). Using DIF to explore item difficulty in CAE listening. *Cambridge ESOL: Research Notes, 32*(04), 16–23.

Gerdes, P. (2001). Ethnomathematics as a new research field, illustrated by studies of mathematical ideas in African history. In Saldaña, J. J. (Ed.), *Science and cultural diversity: Filling a gap in the history of science*, 11–36. Mexico City: Cuadernos de Quipu.

Gierl, M. J. (2004). Using a multidimensionality-based framework to identify and interpret the construct-related dimensions that elicit group differences. *Paper presented at the Annual Meeting of the American Educational Research Association (AERA), 12–14 April 2004, California, San Diego.*

Glickman, M. E., Seal, P., & Eisen, S. V. (2009). A non-parametric Bayesian diagnostic for detecting differential item functioning in IRT models. *Health Service Outcomes Research Method. Doi: 10.1007/s10742-009-0052-4*.

Gómez-Benito, J., Hidalgo, M. D. & Guilera, G. (2010). Bias in measurement instruments: Fair tests. *Papeles del Psicólogo, 31*(1), 75–84. Retrieved from http://www.cop.es./papeles

Gorin, J.S. (2007). Reconsidering issues in validity theory. *Educational Researcher, 36*(8), 456–462. Doi: 10.3102/0013189X07311607

Gottfredson, L. S. (1998). The general intelligence factor. *Human Intelligence*, 24–29.

Government Gazette. (1998). *Employment Equity Act 55 of 1998*. Cape Town: South African Government.

Graham, J. R. & Lily, R. S. (1984). *Psychological testing.* Upper Saddle River, NJ: Prentice-Hall.

Greene, J. C. (2006). Toward a methodology of mixed methods social inquiry. *Research in the Schools, 13*(1), 93-98.

Greene, K. E., & Frantom, C. G. (2002). *Survey development and validation with the Rasch model.* Paper presented at the International Conference on Questionnaire Development, Evaluation, and Testing, Charleston, South Carolina.

Gregory, R. J. (2007). *Psychological testing: History, principles and applications (5th ed.)*. Boston, MA: Pearson International.

Griessel, L., Jansen, J., & Stroud, L. (2013). Administering psychological assessment measures. In C. Foxcroft, & G. Roodt (Eds), *Introduction to psychological assessment in the South African context* (4th ed., pp. 125–144). Cape Town: Oxford University Press Southern Africa.

Grieve, K. W., & Foxcroft, C. (2013). Interpreting and reporting assessment results. In C. Foxcroft, & G. Roodt (Eds), *Introduction to psychological assessment in the South African context* (4th ed., pp. 257–267). Cape Town: Oxford University Press Southern Africa.

Guo, F., Rudner, L. M., & Talento-Miller, E. (2006). *Differential impact as an item bias indicator in CAT and other IRT-based tests*. Paper presented at the International Test Commission's 5th Conference, 6–8 July, Brussels, Belgium.

Hagg, G. (2010). The state and community arts centres in a society in transformation: The South African case. *International Journal of Cultural Studies, 13*(2), 163–184.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 535–556.

Harre, R. (Ed.) (1986). *The social construction of emotions.* Blackwell.

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and practices, 10*(2), 33–41.

Haslberger, A. (2005). Facets and dimensions of cross-cultural adaptation: Refining the tools. *Personnel Review, 34*(1), 85–109.

Heyvaert, M., Maes, B., & Onghena, P. (2013). Mixed methods research synthesis: Definition, framework, and potential. *Qualily & Quantity, 47*(2), 659–676. Doi: 10.1007/s11135-011-9538-6.

Hill, C., Nel, J. A., Van de Vijver, F. J. R, Meiring, D., Valchev, V. H., Adams, B. G. et al. (2013). Developing and testing items for the South African Personality Inventory (SAPI). *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde, 39*(1), Art. #1122, 13 pages. http://dx.doi.org/10.4102/sajip.v39i1.1122

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of management, 21*(5), 967–988.

Hirschman, E. C. (1980). Innovativeness, novelty seeking, and consumer creativity. *Journal of Consumer Research*, 283–295.

http://www.rehab-scales.org/rasch-measurement-model.html

Hultsch, D. F., MacDonald, S. W. S., Hunter, M. A., Maitland, S. B., & Dixon, R. A. (2002). Sampling and generalizability in developmental research: Comparison of random and convenience samples of older adults. *International Journal of Behavioral Development, 26*(4), 345–359. Doi: 10.1080/1650250143000247

Huysamen, G. K. (1996). *Psychological measurement: An introduction with South African examples*. Pretoria: Van Schaik.

Huysamen, G. K. (2006). Recent proposals to estimate transient error within the classical test theory tradition. *SA Journal of Industrial Psychology, 32*(4), 41–47.

International Test Commission (ITC). (2011). Constitution of the International Test Commission.

Ivankova, N. V., Creswell, J. W., & Stick, S. L. (2006). Using mixed-methods sequential explanatory design: From theory to practice. *Fields Methods, 18*(1), 3–20. Doi:10.1177/1525822X05282260

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14–26.

Jones, K. (2002). Issues in the teaching and learning of geometry. In, Haggarty, L. (Ed.), *Aspects of teaching secondary mathematics: Perspectives on practice* (pp. 121–139). London: Routledge Falmer.

Kane, M. (2010). *Errors of measurement, theory and public policy.* Presentation at the 12th annual William H. Angoff memorial lecture at Education Testing Service, on 19 November, at the Education Testing Service, Princeton, New Jersey.

Kanjee, A. (2006). Assessment research. In M. Terre Blanche, K. Durrheim, & D. Painter (Eds.), *Research in practice: Applied methods for the social sciences* (2nd ed., pp. 476–498). Cape Town: University of Cape Town Press.

Kgosana, M. C. (2012). Affirmative action and psychometric tests use in the South African National Defense Force: Are they complementary or conflicting Forces? *J Def Manag, 2*(4). Doi:10.4172/2167-0374.1000112

Kline, P. (2013). *Handbook of psychological testing* (2nd ed.). New York: Routledge

Kline, R. B. (2013). Assessing statistical aspects of test fairness with structural equation modelling. *Educational Research and Evaluation: An International Journal on Theory and Practice, 19*(2-3), 202–222. http://dx.doi.org/10.1080/13803611.2013.767624

Klingner, J. K., & Boardman, A. G. (2011). Addressing the “research gap” in special education through mixed methods. *Learning Disability Quarterly, 34*(3), 208–218.

Krauss, S. E. (2005). Research paradigms and meaning making: a primer. *The Qualitative Report, 10*(4), 758–770.

Krebs, D. E. (1987, December). Measurement theory. *Physical Therapy, 67*(12), 1834–1839.

Kriek, H. (2001). *Guidelines for best practice in occupational assessment in South Africa.* Pretoria: SHL Group plc.

Krishnan, V. (2011). A comparison of principal components analysis and factor analysis for uncovering the Early Development Instrument (EDI) domains. *Community-University Partnership (CUP), Faculty of Extension, University of Alberta, Edmonton, Alberta, Canada*.

Kuhn, T. (1970). *The structure of scientific revolutions*. Chicago, ILL: University of Chicago Press.

Kunda, M., McGreggor, K., & Goel, A. (2009, October). *Addressing the Raven’s Progressive Matrices test of general intelligence*. AAAI Fall Symposium on Multimodal Representational Architectures for Human-Level Intelligence.

Kunda, M., McGreggor, K., & Goel, A. (2010, August). *Taking a look (literally) at the Raven’s intelligence test: Two visual solution strategies.* In Proceedings of the 32nd Annual Meeting of the Cognitive Science Society, Portland, Washington.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic, & C. Weir (Eds), *European language testing in a global context* (pp. 27–48). Cambridge, UK: CUP.

Kurnaz, F. B., & Kelecioğlu, H. (2008). Investigation of Peabody Picture Vocabulary Test from the Point of Item Bias Peabody Picture Vocabulary Test. *World Applied Sciences Journal, 3*(2), 231–239.

Kvist, A. V., & Gustafsson, J.-E. (2008).The relation between fluid intelligence and the general factor as a function of cultural background: A test of Cattell's Investment theory. *Intelligence, 36*, 422–436.

Lantano, A. S. (2010, October). *A test anlaysis report of the DT English year 1 using the Rasch model.* Paper presented at the 11th National Convention on Statistics (NCS), EDSA Shangri-La Hotel.

Leech, N. L., & Onwuegbuzie, A. J. (2007). An array of qualitative data analysis tools: a call for data analysis triangulation. *School Psychology Quarterly, 22*(4), 557–584, Doi: 10.1037/1045-3830.22.4.557

Leedy, P. D., & Ormrod, J. E. (2010). *Practical research: Planning and design* (9th ed.). Upper Saddle River, NJ: Pearson Educational International.

Levin, K. A. (2006). Study design 111: Cross-sectional studies. *Evidence-based Dentistry, 7*, 24–25. Doi: 10.1038/sj.ebd.6400375

Li, R. (1996). *A theory of conceptual intelligence: Thinking, learning, creativity and giftedness*. London: Praeger.

Li, Z., & Zumbo, B. D. (2009). Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. *Psicologica, 30*, 343–370.

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement, 2*, 266–283.

Linacre, J. M. (2005). Rasch dichotomous model vs. one parameter logistic model. *Rasch Measurement Transactions, 19*(3), 1032.

Linacre, J. M. (2009). Local independence and residual covariance: A study of Olympic figure skating ratings. *Journal of Applied Measurement, 10,* 2–13.

Linacre, J. M. (2011a). Constructing valid performance assessments: The view from the shoulders of giants. *Samuel J. Messick Memorial Lecture. LTRC, Michigan, June 2011*. Retrieved from www.rasch.org/memos.htm.

Linacre, J. M. (2011b). *Winsteps* (version 3.71.0) [Computer Software]. Chicago: Winsteps.com.

Linacre, J. M. (2012). *Winsteps* (version 3.75. 0) [Computer Software]. Beaverton, Oregon: Winsteps.com.

Lipka, J. (1994, May). Culturally negotiated schooling: Toward a Yup'ik mathematics. *Journal of American Indian Education, 33*(3). Retrieved from http://jaie.asu.edu/

Lipka, J., & Adams, B. L. (2004). Some evidence for ethnomathematics: Quantitative and qualitative data from Alaska. *DG 15: Ethnomathematics at ICME-10 in Denmark*.

Lohman, D. F. (2005). The role of nonverbal ability tests in identifying academically gifted students: An aptitude perspective. *Gifted Child Quarterly, 49*(2), 111–138.

Louw, J., & Van Hoorn, W. (1997). Psychology, conflict, and peace in South Africa: Historical notes. *Peace and Conflict: Journal of Peace Psychology, 3*(3), 233–243.

Louw, W. (2009). Africanisation: The dilemma to Africanise or to globalise a curriculum. *Conference of the International Journal of Arts and Sciences, 1*(6), 62–70.

Magis, D., & De Boeck, P. (2010). *Identification of differential item functioning in multiple-group settings: A multivariate outlier detection approach.* Paper presented at the 75th annual meeting of the Psychometric Society, July, GA, Athens.

Malda, M., Van de Vijver, F. J. R., & Temane, Q. M. (2010). Rugby versus soccer in South Africa: Content familiarity contributes to cross-cultural differences in cognitive test scores. *Intelligence, 38*(6), 582–595.

Maree, D. J., Maree, M., & Collins, C. (2008). Constructing a South African Hope measure. *Journal of Psychology in Africa, 18*(1), 167–178.

Maree, K. (2010). Editorial: Assessment in psychology in the 21st century: A multi-layered endeavour. *SA Journal of Psychology, 40*(3), 229–233.

Mason, P. (2005). Visual data in applied qualitative research: lessons from experience. *Qualitative Research, 5*(3), 325–346. Doi: 10.1177/1468794105054458

Matsumoto, D. (2001). Cross-cultural psychology in the 21st century. In J. Halonen, & S. Davis (Eds.). *The many faces of research in the 21st century (chap. 5)*. Retrieved from http://teachpsych.lemoyne.edu/teachpsych/ faces/ch05.htm

Mayes, R. L., Rittschof, K., Forrester, J. H., Schuttlefield Christus, J. D., Watson, L., & Peterson, F. (2015). Quantitative reasoning in environmental science: Rasch measurement to support QR assessment. *Numeracy: Advancing Education in Quantitative Literacy, 8*(2), 4.

McCamey, R. (2014). A primer on the one-parameter Rasch model. *American Journal of Economics and Business Administration, 6*(4), 159–163. Doi: 10.3844/ajebasp.2014.159.163

McCoy, K. M. (2000, Spring). Statistics versus measurement? *Popular Measurement,* 30.

McEvoy, P., & Richards, D. (2006). A critical realist rationale for using a combination of quantitative and qualitative methods. *Journal of Research in Nursing, 11*, 66–78. Doi: 10.1177/1744987106060192

McGlone, C. W. (2008, July). *The role of culturally-based mathematics in the general mathematics curriculum: A case for presenting culturally-based mathematics to all students.* 11th International Congress in Maths Education, Monterrey, Mexico.

McGrew, K. S. & Flanagan, D. P. (1998). *The intelligence test desk reference (ITDR): Gf-Gc cross-battery assessment*. Boston: Allyn & Bacon.

McMillan, J. H. (2000). *Basic assessment concepts for teachers and school administrators*. Baltimore, MD: ERIC Clearinghouse on Assessment and Evaluation College Park.

McMurphy-Pilkington, C., Pikiao, N., & Rongomai, N. (2008). Indigenous people: Emancipatory possibilities in curriculum development. *Canadian Journal of Education, 31*(3), 614–638.

McShane, L. S. & Von Glinow, A. M. (2005). *Organisational behaviour* (3rd ed.). Boston, MA: McGraw-Hill Irwin.

Meiring, D., Van de Vijver, A. J. R.; Rothmann, S. & Barrick, M. R. (2005). Construct, item, and method bias of cognitive and personality tests in South Africa. *SA Journal of Industrial Psychology, 31*(1), 1–8.

Miller, E. (1993). *From dependency to autonomy: Studies in organization and change*. London: Tavistock Institute of Human Relations.

Mingers, J., Mutch, A., & Willcocks, L. (2013). Critical realism in information systems research. *MIS Quarterly, 37*(3), 795–802.

Mingers, J. & Willcocks, L. (2004). *Social theory and philosophy for information systems.* John Wiley Series in information systems. Chichester, West Sussex, England: Wiley.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Education Measurement, 33*(4), 379–416.

Miyata, M. (2007). A Rasch analysis of the ELI listening placement test. Retrieved from http://hdl.handle.net/10125/20193

Moerdyk, A. (2015). *The principles and practice of psychological assessment* (2nd ed.)*.* Pretoria: Van Schaik.

Morales, R. A. (2009, April). Evaluation of mathematics achievement test: A comparison of CTT and IRT. *The International Journal of Educational and Psychological Assessment, 1*(1), 19–26.

Moran, J. (2010). *Interdisciplinarity* (2nd ed.). New York: Routledge.

Mouton, J. (2001). *Understanding social research.* Pretoria: Van Schaik.

Msila, V. (2009, June). Africanisation of education and the search for relevance and context. *Educational Research and Review, 4*(6), 310–315.

Muchinsky, P. M., Kriek, H. J. & Schreuder, A. M. G. (1998). *Personnel psychology*. Durban: International Thomson Publishing (Southern Africa).

Mueller, C.E., Bullock, E.E., & Leierer, S.J. (2010). Examining psychometric and measurement properties of the Career Thoughts Inventory: Demonstration and use of the Rasch measurement model in career assessment research. Technical Report No. 51. *Florida State University, Florida*. Retrieved from http://www.career.fsu.edu/techcenter/TR51.pdf

Mustafa, R. F. (2011). The P.O.E.Ms of educational research: A beginners' concise guide. *International Education Studies, 4*(3), 23–30.

Nelson-Barber, S., & Trumbull, E. (2007). Making assessment practices valid for Native American students. *Journal of American Indian Education, 46*(3), 136–152.

Nettleton, A. (2010). Life in a Zulu village: Craft and the art of modernity in South Africa. *The Journal of Modern Craft, 1*(3), 55–78. Doi: 10.2752/174967810X 12657245205189

Ngulube, P., Mokwatlo, K., & Ndwandwe, S. (2009). Utilisation and prevalence of mixed methods research in library and information research in/South Africa 2002-2008. *South African Journal of Library and Information Sciences, 75*(2), 105–116. Retrieved from http://sajlis.journals.ac.za

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Nzama, L., De Beer, M., & Visser, D. (2008). Predicting work performance through selection interview ratings and psychological assessment. *SA Journal of Industrial Psychology, 34*(3), 39–47.

Nzimande, B. (1995). *Culture fair testing? To test or not to test*. Paper presented at the Congress on Psychometrics for Psychologists and Personnel Practitioners, Pretoria, 5 to 6 June, Pretoria.

Olckers, C. (2013). Psychological ownership: Development of an instrument. *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde, 39*(2), Art. #1105, 13 pages. http://dx.doi.org/10.4102/sajip.v39i2.1105

Onwuegbuzie, A. J., & Collins, K. M. T. (2007). A typology of mixed methods sampling designs in social science research. *The Qualitative Report, 12*(2), 281–316. http://www.nova.edu/ssss/QR/QR12-2/onwuegbuzie2.pdf

Onwuegbuzie, A. J., & Combs, J. P. (2011). Data analysis in mixed research: A primer. *International Journal of Education, 3*(1), 1–25.

Onwuegbuzie, A. J., & Johnson, R. B. (2006). The validity issue in mixed research. *Research in the Schools, 13*(1), 48–63.

Onwuegbuzie, A. J., & Leech, N. L. (2007). Sampling designs in qualitative research: Making the sampling process more public. *The Qualitative Report, 12*(2), 238–254. Retrieved from http://www.nova.edu/ssss/QR/QR12-2/onwuegbuzie1.pdf

Onwuegbuzie, A. J., Leech, N. L., & Collins, K. M. T. (2010). Innovative data collection strategies in qualitative research. *The Qualitative Report, 15*(3), 696–726. Retrieved from http://www.nova.edu/ssss/QR/QR15-3/onwuegbzie.pdf

Osterlind, S. J., & Everson, H. T. (2009). *DIF series: Quantitative applications in the Social Sciences* (2nd ed.). Thousand Oaks, CA: Sage.

Owen, K. (1992). *Test – item bias: Methods, findings and recommendations.* Pretoria: Human Sciences Research Council.

Owen, K. (1998). *The role of psychological tests in education in South Africa: Issues, controversies and benefits.* Pretoria: Human Sciences Research Council.

Oye, N. D., A.Iahad, N., & Ab.Rahim, N. (2012). Using mixed method approach to understand acceptance and usage of ICT in Nigerian public university. *International Journal of Computers and Technology, 2*(3), 47–63. Retrieved from www.ijctonline.com

Pan, Y. (2009). The impact of test design on teaching. *The International Journal of Educational and Psychological Assessment, 3*, 94–103.

Papalia, D. E., & Olds, S. W. (1988). *Psychology* (2nd ed). New York: McGraw-Hill.

Paterson, H., & Uys, K. (2005). Critical issues in psychological test use in the South African workplace. *SA Journal of Industrial Psychology, 31*(3), 12–22.

Pauwels, L. (2008). Taking and using: Ethical issues of photographs for research purposes. *Visual Communication Quarterly, 15*(4), 243–257.

Pedrajita, J. Q., & Talisayon, V. M. (2009). Identifying biased test items by differential item functioning analysis using contingency table approaches: A comparative study. *Education Quarterly, 67*(1), 21–43.

Penrose, L. S., & Raven, J. C. (1936). A new series of perceptual tests: Preliminary communication. *British Journal of Psychology (Medical Section), XVI*, 97–105.

Perrone, M. (2006). Differential item functioning and item bias: Critical considerations in test fairness. *Teachers' College, Columbia University working papers in TESOL and Applied Linguistics: The Forum, 6*(2), 1–3. Retrieved from http://www.tc.columbia.edu/tesolalwebjournal.

Peter, J., Leichner, N., Mayer, A.-K., & Krampen, G. (2015). A short test for the assessment of basic knowledge in psychology. *Psychology Learning and Teaching*, 1475725715605763.

Peters, M., & Mergen, B. (1977). "Doing the rest": The uses of photographs in American Studies. *American Quarterly*, 280–303.

Pheto-Moeti, M. B. (2005). *An assessment of seshoeshoe dress as a cultural identity for Basotho women of Lesotho.* Unpublished masters dissertation. Bloemfontein, University of the Free State.

Piaget, J. (1972). Intellectual evolution from adolescence to adulthood. *Human Development, 15*, 1–12.

Plake, B. S., & Jones, P. (2002). *Ensuring fair testing practices: The responsibilities of test sponsors, test developers, test administrators, and test takers in ensuring fair testing practices.* Paper presented at the meeting of the Association of Test Publishers, February, Carlsbad, California.

Polikoff, M. S., May, H., Porter, A. C., Elliott, S. N., Goldring, E., & Murphy, J. (2009). An examination of differential item functioning on the Vanderbilt assessment of leadership in education. *Journal of School Leadership, 19*(6), 661–679.

Poortinga, Y. (1995). Cultural bias in assessment: Historical and thematic issues. *European Journal of Psychological Assessment, 11*(3), 140–146.

Potter, W. J., Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research, 27*, 258–284.

Pour, I. M., & Ghafar, M. N. A. (2009). The analysis of Iran Universities' 2003-2004 entrance examination to detect biased items. *Jurnal Teknologi, 50*(E), 21–27.

Pratt, D. D., & Gutteridge, R. G. (2014). The role of the social mechanism in social transformation: a critical realist approach to blended learning. Paper presented at the 8th annual conference on World Wide Web applications held in Bloemfontein, 6–8 September. Retrieved from http://er.dut.ac.za/handle/12345 6789/113

Prokosch, M. D.; Yeo, R. A., & Miller, G. F. (2005). Intelligence tests with higher g-loadings show higher correlations with body symmetry: Evidence for a general fitness factor mediated by developmental stability. *Intelligence, 33*, 203–213. Doi:10.1016/j.intell.2004.07.007

Rapoport, S. I. (1999). How did the human brain evolve? A proposal based on new evidence from *in vivo* brain imaging during attention and ideation. *Brain Research Bulletin, 50*(3), 149–165.

Rasch, G. (1979). Letter from George Rasch to Ben Wright – personal communication, 19 October. Retrieved from http://www.raschmeasurement.com

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*, 1–48.

Ray, J. L., & Smith, A. D. (2011). Using photographs to research organizations: Evidence, considerations, and application in a field study. *Organizational Research Methods*, 1–28. Doi:1094428111431110.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied psychological measurement, 1*(21), 25 – 36.

Ridgen, C. (1999). The eye of the beholder: Designing for colour-blind users. *British Telecommunications Engineering, 17*, 2–6.

Roever, C. (2005). *"That's not fair!" Fairness, bias, and differential item functioning in language testing*. Retrieved from http://www2.hawaii.edu/~roever/brownbag.pdf

Roodt, G. (2013a). Basic measurement and scaling concepts. In C. Foxcroft, & G. Roodt (Eds), *Introduction to psychological assessment in the South African context* (4th ed., pp. 29–45). Cape Town: Oxford University Press Southern Africa.

Roodt, G. (2013b). Reliability: Basic concepts and measures. In C. Foxcroft, & G. Roodt (Eds), *Introduction to psychological assessment in the South African context* (4th ed., pp. 47–55). Cape Town: Oxford University Press Southern Africa.

Roodt, G. (2013c). Validity: Basic concepts and measures. In C. Foxcroft, & G. Roodt (Eds), *Introduction to psychological assessment in the South African context* (4th ed., pp. 57–67). Cape Town: Oxford University Press Southern Africa.

Rosa, M., & Orey, D. C. (2010, January-June). Culturally relevant pedagogy: Ethnomathematical approach. *Horizontes, 28*(1), 19–31.

Rosa, M., & Orey, D. C. (2011). Ethnomathematics: The cultural aspects of mathematics. *Revista Latinoamericana de Etnomatemática, 4*(2), 32–52.

Rothmann, S., & Cilliers, F. V. N. (2007). Present challenges and some critical issues for research in Industrial/Organisational Psychology in South Africa. *SA Journal of Industrial Psychology, 33*(1), 8–17.

Rushton, J. P., & Jensen, A. R. (2005a). Thirty years of research on race differences on cognitive ability. *Psychology, Public Policy and Law, 11*(2), 235–294.

Rushton, J. P., & Jensen, A. R. (2005b). Wanted: More race realism, less moralistic fallacy. *Psychology, Public Policy and Law, 11*(2), 328–336.

Ryans, D. G. (1938). The concept of intelligence. *Journal of Educational Psychology, 29*(6), 449–458.

Salkind, N. J. (2014). *Exploring research (8th ed.)*. Upper Saddle River, NJ: Pearson Education International.

Sattler, J. M. (2001). *Assessment of children: Cognitive applications (4th ed.)*. San Diego, CA: Sattler Publishers.

Schaap, P., & Vermeulen, T. (2008). The construct equivalence and item bias of the PIB/SPEEX conceptualisation-ability test for members of five language groups in South Africa. *SA Journal of Industrial Psychology, 34*(3), 29 – 38.

Schmidt, C. (2006). Validity as an action concept in IO psychology. *SA Journal of Industrial Psychology, 32*(4), 59–67.

Schnohr, C. W., Kreiner, S., Due, E. P., Currie, C., Boyce, W., & Diderichsen, F. (2007). Differential item functioning of a family affluence Scale: Validation study on data from HBSC 2001/2. *Social Indicators Research. Doi: 10.1007/s11205-007-9221-4*.

Sick, J. (2011). Rasch measurement and factor analysis. *Shiken: Jalt Testing and Evaluation SIG Newsletter, 15*(1), 15–17.

Shabani, K., Khatib, M., Ebadi, S. (2010). Vygotsky's zone of proximal development: Instructional implications and teachers' professional development. *English Language Teaching, 3*(4), 237–248.

Smit, G. J. (1996). *Psychometric aspects of measurement.* Pretoria: Kagiso.

Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology, 15*, 201–293.

Spearman, C. (1930). Disturbers of tetrad differences. *The Journal of Educational Psychology, 21*(8), 559–573.

Stepchenkova, S., & Zhan, F. (2013). Visual destination images of Peru: comparative content analysis of DMO and user generated photography. *Tourism Management.* http://dx.doi.org/10.1016/j.tourman.2012.08.006

Sternberg, R. J. (1984). A contextualist view of the nature of intelligence. *International Journal of Psychology, 19*, 307–334.

Sternberg, R. J. (2004, July-August). Culture and intelligence. *American Psychologist, 59*(5), 325–338.

Sternberg, R. J. (2006). The nature of creativity. *Creativity Research Journal, 18(*1), 87–98.

Sternz, J. E., Plano Clark, V., & Matkin, G.S. (2012). Applying mixed methods to leadership research: A review of current practices. *The Leadership Quarterly*, http://dx.doi.org/10.1016/j.leaqua.2012.10.001

Suppes, P., & Zinnes, J. L. (1962). Basic measurement theory. *Psychology Series: Technical Report, No. 45*. Stanford, California: Institute for Mathematical Studies in Social Sciences, Stanford University.

Syed, J., Mingers, J., & Murray, P. A. (2009). Beyond rigour and relevance: a critical realist approach to business education. *Management Learning*, 1–15. Doi: 10.1177/1350507609350839

Tan, X., Xiang, B., Dorans, N. J., & Qu, Y. (2010). *The value of studied item in the matching criterion in differential item functioning (DIF) analysis.* Princeton, NJ:ETS. Retrieved from http://www.ets.org/research/contact.html

Teddlie, C., & Yu, F. (2007). Mixed methods sampling: A typology with examples. Journal of Mixed Methods Research, 1(1), 77–100, Doi: 10.1177/2345678 906292430. Retrieved from http://mmr.sagepub.com

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis and Rheumatism (Arthritis Care and Research), 57*(8), 1358–1362. Doi: 10.1002/art.23108

Terry, C. L. (2011). Mathematical counter story and African American male students: Urban mathematics education from a critical race theory perspective. *Journal of Urban Mathematics Education, 4*(1), 23–49.

Theron, A. (2009). Theoretical perspectives in psychology. In Z.C. Bergh, & A.L. Theron (Eds), *Psychology in the work context* (4th ed., pp. 2–15). Cape Town: Oxford University Press South Africa.

Theron, C. (2007). Confessions, scapegoats and flying pigs: Psychometric testing and the law. *SA Journal of Industrial Psychology, 33*(1), 102–117.

Thompson, N. A., & Barnard, J. J. (2009). *Introduction to modern psychometrics: A workshop on the theory and applications of measurement*. Workshop hosted at the University of South Africa, 13–15 October, Pretoria.

Thompson, B., & Vacha-Haase, T. (2000, April). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*(2), 174–195. Doi: 10.1177/0013164400602002.

Thurstone, L. (1936). The factorial isolation of primary abilities. *Psychometrika, 1*(3), 175–182.

Thurstone, L. (1938). *Primary mental abilities*. Chicago, ILL: University of Chicago Press.

Trafford, V. N., & Leshem, S. (2008). *Stepping stones to achieving your doctorate: Focussing on your viva from the start*. Maidenhead, UK: Open University Press.

Urbina, S. (2004). *Essentials of psychological testing*. Hoboken, NJ: Wiley.

Van de Vijver, A. J. R., & Rothmann, S. (2004). Assessment in multicultural groups: The South African case. *SA Journal of Industrial Psychology, 30*(4), 1–7.

Van de Vijver, F. J. R., & Phalet, K. (2004). Assessment in multicultural groups: The role of acculturation. *Applied Psychology: An International Review, 53*(2), 215–236.

Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: an overview. *Revue Europeenne de Psychologie Appliqué, 54,* 119–135. Retrieved from www.sciencedirect.com

Van den Noortgate, W., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics, 30*(4), 443–464.

Van Dulm, O., & Southwood, F. (2013). Child language assessment and intervention in multilingual and multicultural South Africa: Findings of a national survey. *Stellenbosch Papers in Linguistics, 42*, 55–76. Doi: 10.5774/42-0-147

Van Eeden, E., & De Beer, M. (2013). Assessment of cognitive functioning. In C. Foxcroft, & G. Roodt (Eds), *Introduction to psychological assessment in the South African context* (4th ed., pp. 147–169). Cape Town: Oxford University Press Southern Africa.

Vasquez, M. J. (2012). Psychology and social justice: Why do we do what we do. *American Psychologist, 67*(5), 337.

Venkatesh, J. P., Brown, S. A., & Bala, H. (2013). Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems. *MIS Quarterly, 37*(1), 21–54.

Viljoen, C. T., & Van der Walt, J. L. (2003). Being and becoming: Negotiations on educational identity in (South) Africa. *South African Journal of Education, 23*(1), 13–17.

Visser, D., & Viviers, R. (2010). Construct equivalence of the OPQ32n for Black and White people in South Africa. *SA Journal of Industrial Psychology, 36*(1). Retrieved from www.scielo.org.za

Vygotsky, L.S. (1978). *Mind in society: The development of higher-order psychological processes*. Cambridge, MA: Harvard University Press.

Wacker, J. G. (1998). A definition of theory: Research guidelines for different theory-building research methods in operations management. *Journal of Operations Management, 16*(4), 361–385. Doi: 10.1016/S0272-6963(98)00019-9

Wang, W. (2000). The simultaneous factorial analysis of differential item functioning. *Methods of Psychological Research Online, 5*(1), 57–76. Retrieved from http://www.mpr-online.de

Wechsler, D. (1939). *Measurement of adult intelligence*. Baltimore, MD: Williams & Witkins.

Weiner, I. B. (2013). Psychological assessment is here to stay. *Archives of Assessment Psychology, 3*(1), 11–21.

Welman, C., Kruger, F. & Mitchell, B. (2009). *Research methodology* (3rd ed.). Cape Town: Oxford University Press Southern Africa.

Whiting, G. W. & Ford, D. Y. (2006). Under-representation of diverse students in gifted education: Recommendations for non-discriminatory assessment (part 2). *Gifted Education Press Quarterly, 20*(3), 6–10.

Wiberg, M. (2004). *Classical test theory vs. Item response theory: An evaluation of the theory test in the Swedish driving-license test.* Department of Statistics, Umea University, Sweden, EM No. 50.

Wicherts, J. M., Dolan, C. V., Carlson, J. S., & Van der Maas, H. L. J. (2010). Raven's test performance of sub-Saharan Africans: Average performance, psychometric properties, and the Flynn Effect. *Learning and Individual Differences, 20*, 135–151.

Wilson, V. & McCormack, B. (2006). Critical realism as emancipator action: the case for realistic evaluation in practice development. *Nursing Philosophy, 7*(1), pp. 45–57.

Wong, M., & Lipka, J. (2011). Adapting assessment instruments for an Alaskan context. *Mathematics: Traditions and [new] practices*, 821–829.

Wright, B. D., & Stone, M. H. (1999). *Measurement essentials* (2nd ed.). Wilmington, Del: Wide Range Inc.

Wu, M., & Adams, R. (2007). Applying the Rasch model to psycho-social measurement: A practical approach. *Educational Measurement Solutions*, Melbourne. Retrieved from http://www.edmeasurement.com.au/_docs/Rasch Measurement_complete.pdf

Yang, Z., Wang, X., & Su, C. (2006). A review of research methodologies in international business. *International Business Review, 15*, 601-617. Retrieved from www.elsevier.com/locate/ibusrev

Yau-Fai Ho, D. (1994). Introduction to cross-cultural psychology. In L.L. Loebner, & U.P. Gielen (Eds.), *Cross-cultural topics in Psychology* (pp. 1–32). Westport, CT: Praeger.

Yu, L., Lei, P., & Suen, H. K. (2006). *Using a differential item functioning (DIF) procedure to detect differences in Opportunity to Learn (OTL).* Paper presented at the annual meeting of the American Educational Research Association, 10 April, San Francisco, California.

Zhang, O., Shen, L., & Cannady, M. (2010). *Polytomous IRT or Testlet model: An evaluation of scoring models in small testlet size situations.* Paper presented at the annual meeting of the 15th International Objective Measurement Workshop, April, Boulder, Colorado.

Zumbo, B. D. (1999). *A Handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223–233.

Zumbo, B. D., Gelin, M. N., & Hubley, A. M. (2002). The construction and use of psychological tests and measures. In the Psychological theme of the *Encyclopedia of Life Support Systems (EOLSS).* Oxford, UK: Eolss Publishers. http://www.eolss.net

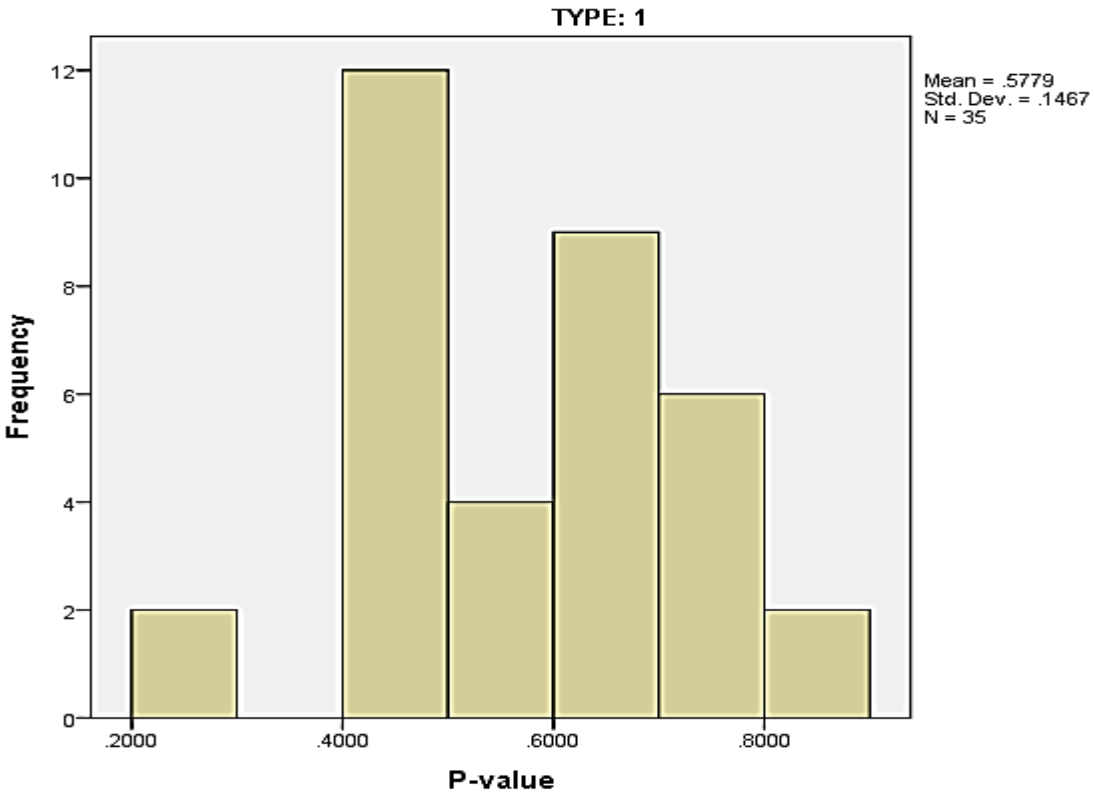
APPENDICES

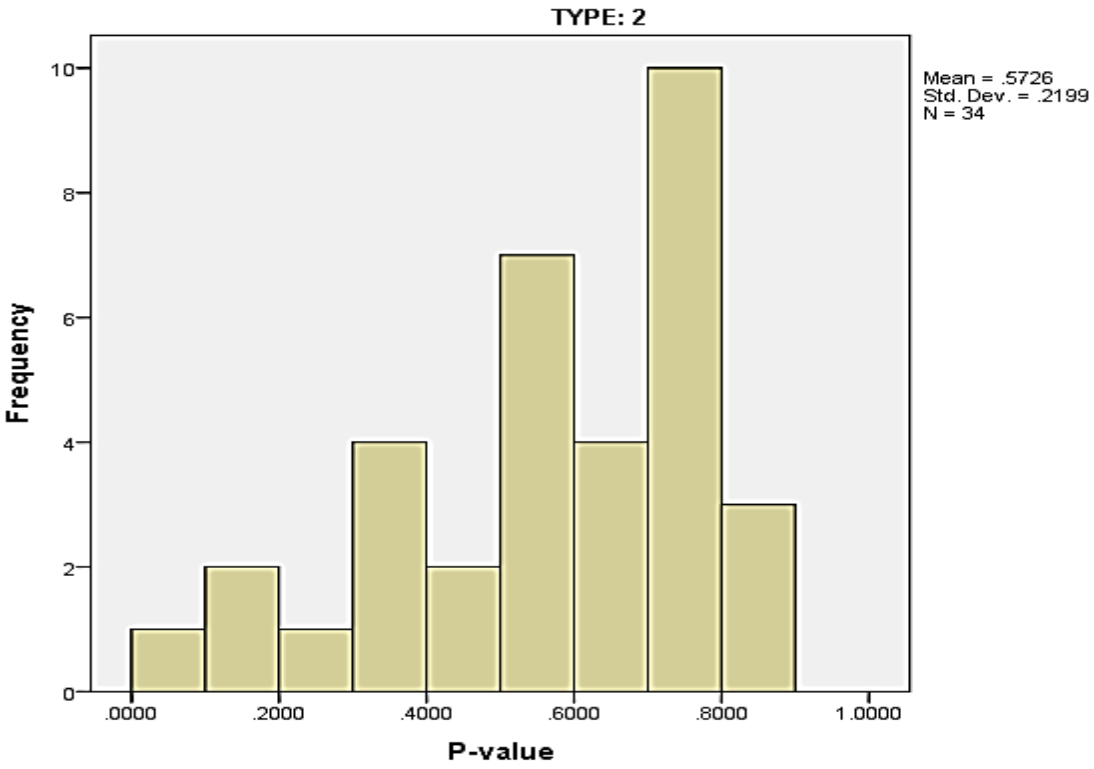
Appendix A: P-values for all the *new items*

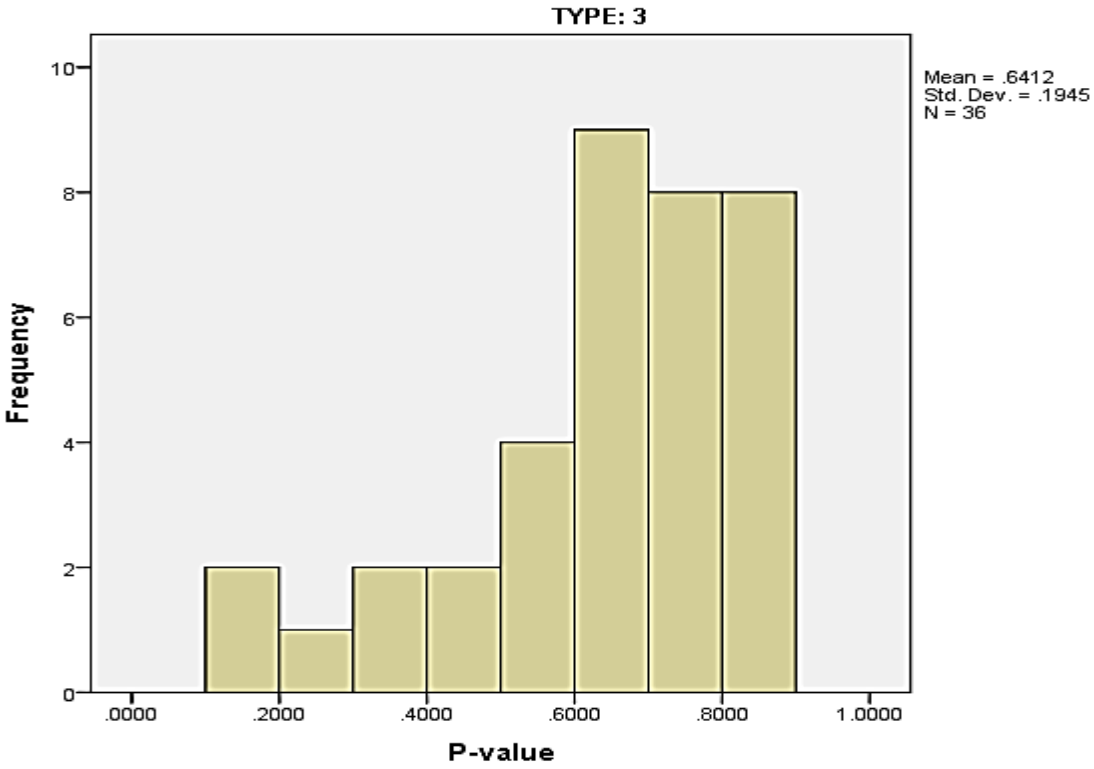
Item No.	*P*-value	Item No.	*P*-value	Item No	*P*-value	Item No.	*P*-value
1	0.566	26	0.906	51	0.559	76	0.808
2	0.77	27	0.617	52	0.316	77	0.482
3	0.872	28	0.716	53	0.812	78	0.614
4	0.694	29	0.909	54	0.903	79	0.821
5	0.65	30	0.752	55	0.793	80	0.794
6	0.654	31	0.853	56	0.756	81	0.837
7	0.853	32	0.784	57	0.139	82	0.694
8	0.842	33	0.592	58	0.537	83	0.647
9	0.534	34	0.55	59	0.234	84	0.343
10	0.618	35	0.571	60	0.817	85	0.482
11	0.125	36	0.771	61	0.064	86	0.286
12	0.872	37	0.788	62	0.833	87	0.651
13	0.717	38	0.316	63	0.622	88	0.796
14	0.646	39	0.798	64	0.497	89	0.474
15	0.284	40	0.618	65	0.337	90	0.729
16	0.729	41	0.769	66	0.85	91	0.434
17	0.781	42	0.524	67	0.645	92	0.711
18	0.647	43	0.663	68	0.864	93	0.731
19	0.261	44	0.294	69	0.844	94	0.697
20	0.771	45	0.653	70	0.671	95	0.427
21	0.686	46	0.235	71	0.588	96	0.781
22	0.458	47	0.14	72	0.161	97	0.605
23	0.529	48	0.808	73	0.207	98	0.135
24	0.795	49	0.831	74	0.731	99	0.518
25	0.669	50	0.707	75	0.702	100	0.462

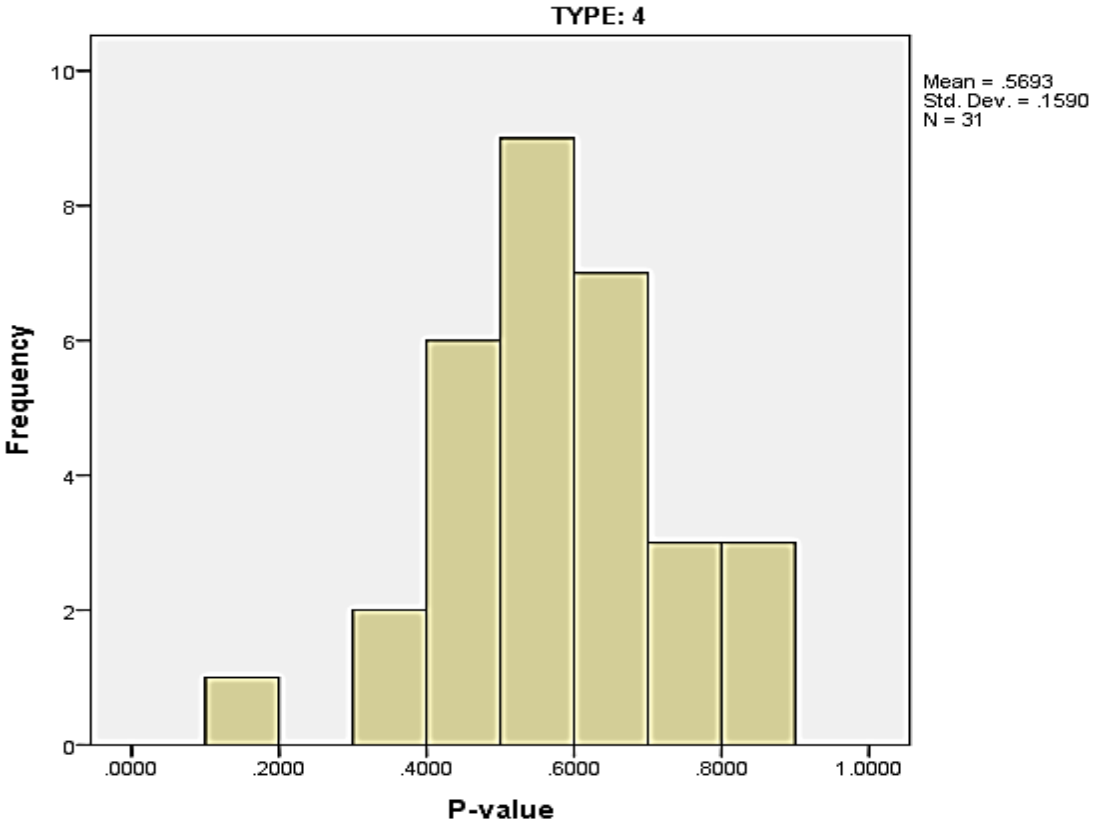
Item No.	*P*-value	Item No.	*P*-value	Item No	*P*-value	Item No.	*P*-value
101	0.804	126	0.542	151	0.753	176	0.591
102	0.441	127	0.515	152	0.31	177	0.623
103	0.631	128	0.829	153	0.712	178	0.515
104	0.649	129	0.636	154	0.805	179	0.756
105	0.686	130	0.71	155	0.536	180	0.417
106	0.457	131	0.505	156	0.707	181	0.429
107	0.477	132	0.599	157	0.602	182	0.573
108	0.561	133	0.582	158	0.784	183	0.456
109	0.668	134	0.548	159	0.314	184	0.356
110	0.328	135	0.756	160	0.771	185	0.512
111	0.54	136	0.556	161	0.845	186	0.641
112	0.354	137	0.475	162	0.492	187	0.362
113	0.662	138	0.412	163	0.461	188	0.604
114	0.679	139	0.399	164	0.687	189	0.686
115	0.521	140	0.577	165	0.061	190	0.743
116	0.683	141	0.551	166	0.858	191	0.436
117	0.52	142	0.804	167	0.746	192	0.62
118	0.229	143	0.809	168	0.494	193	0.55
119	0.583	144	0.717	169	0.545	194	0.709
120	0.689	145	0.522	170	0.264	195	0.417
121	0.62	146	0.653	171		196	0.551
122	0.66	147	0.571	172	0.477	197	0.471
123	0.497	148	0.472	173	0.63	198	0.787
124	0.568	149	0.458	174	0.625	199	0.486
125	0.568	150	0.295	175	0.535	200	0.794

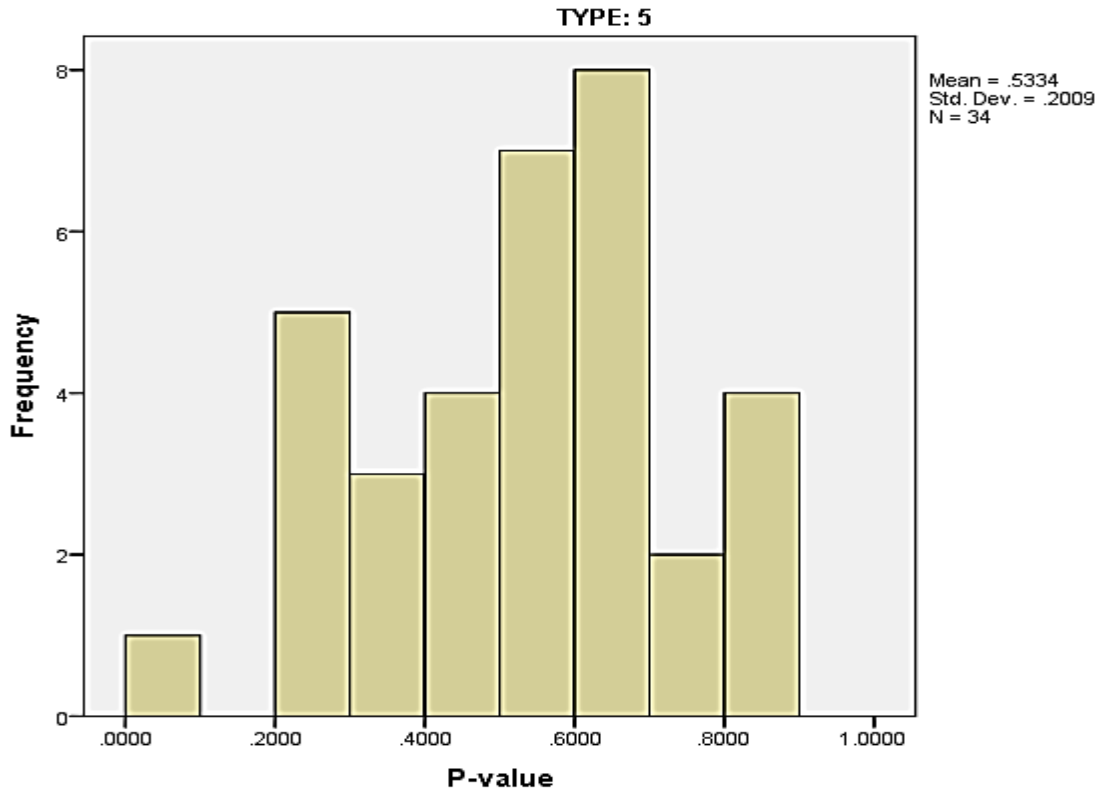
Appendix B: Graphical representations of the distribution of *p*-values

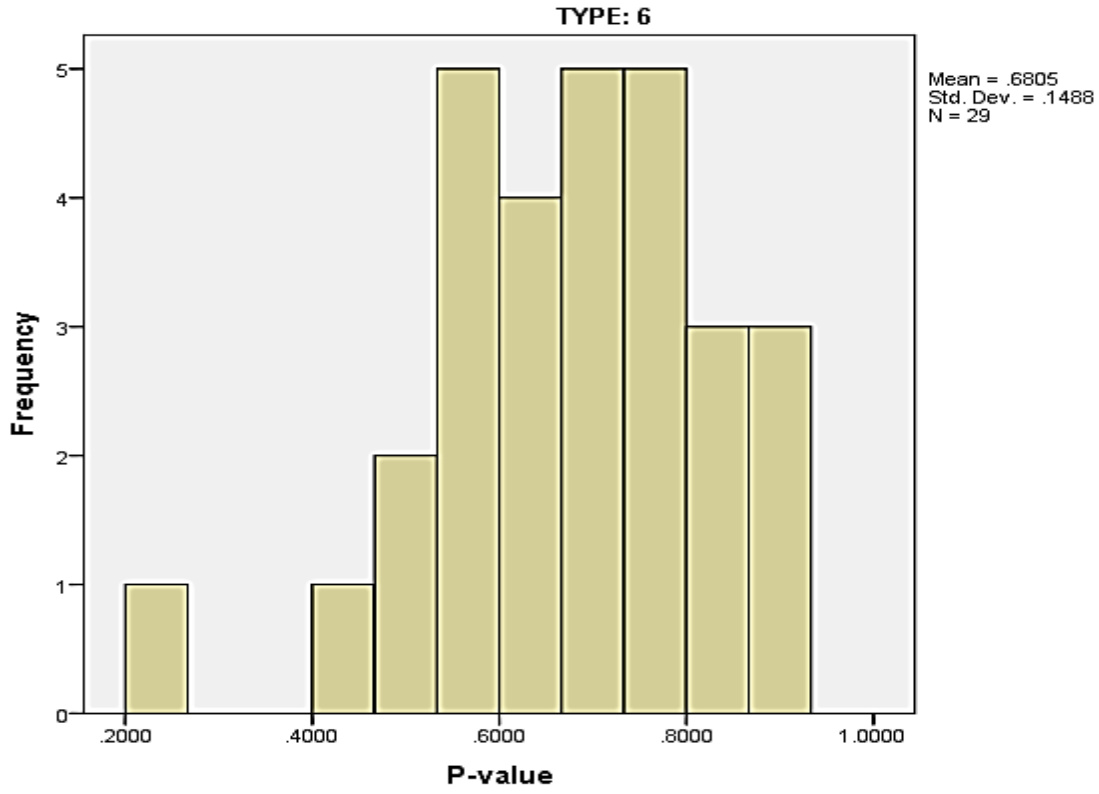


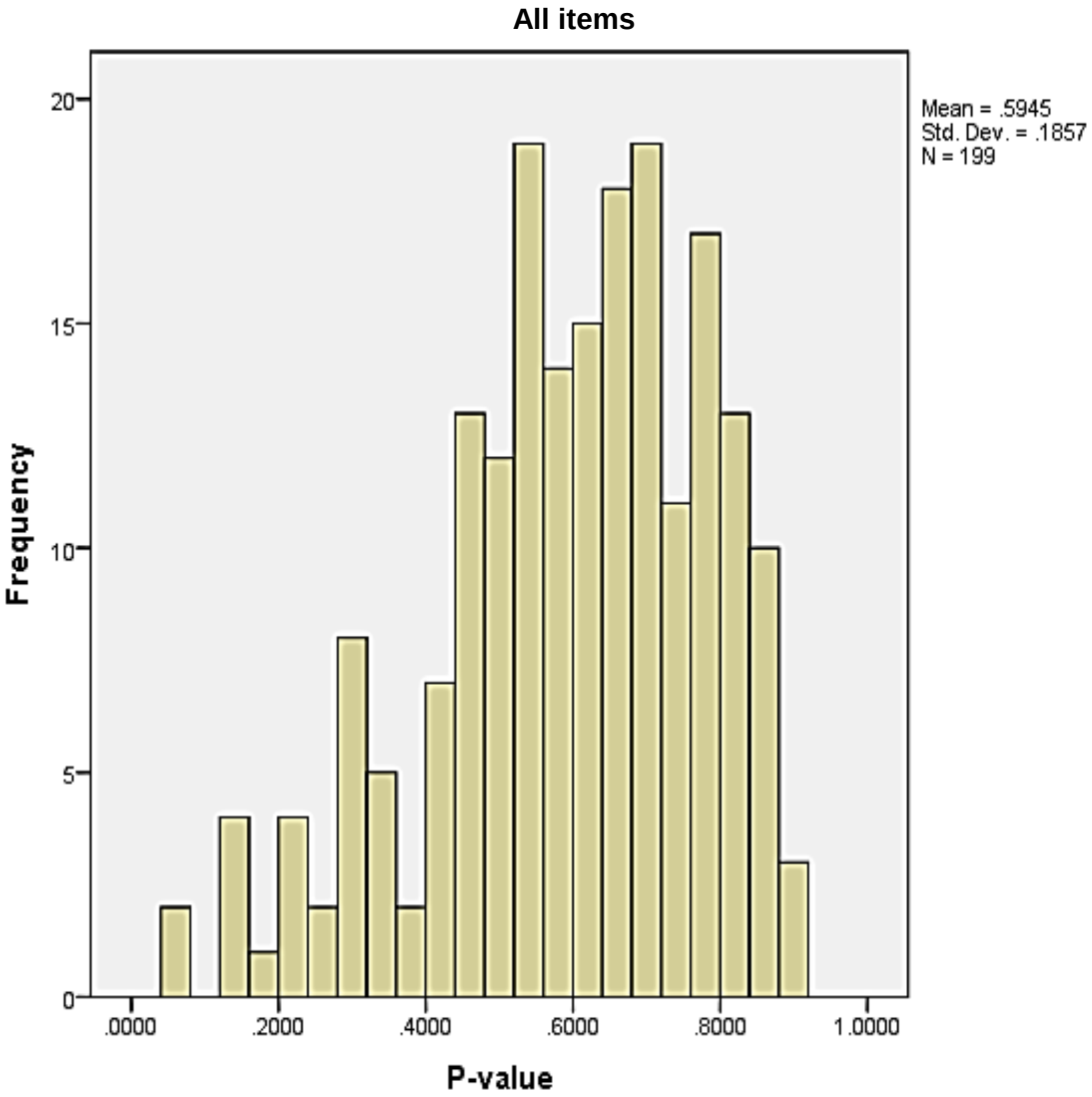



TYPE: 3
Mean = .6412
Std. Dev. = .1945
N = 36
Frequency
P-value
0
2
4
6
8
10
.0000
.2000
.4000
.6000
.8000
1.0000


TYPE: 4
Mean = .5693
Std. Dev. = .1590
N = 31
Frequency
P-value
0
2
4
6
8
10
.0000
.2000
.4000
.6000
.8000
1.0000

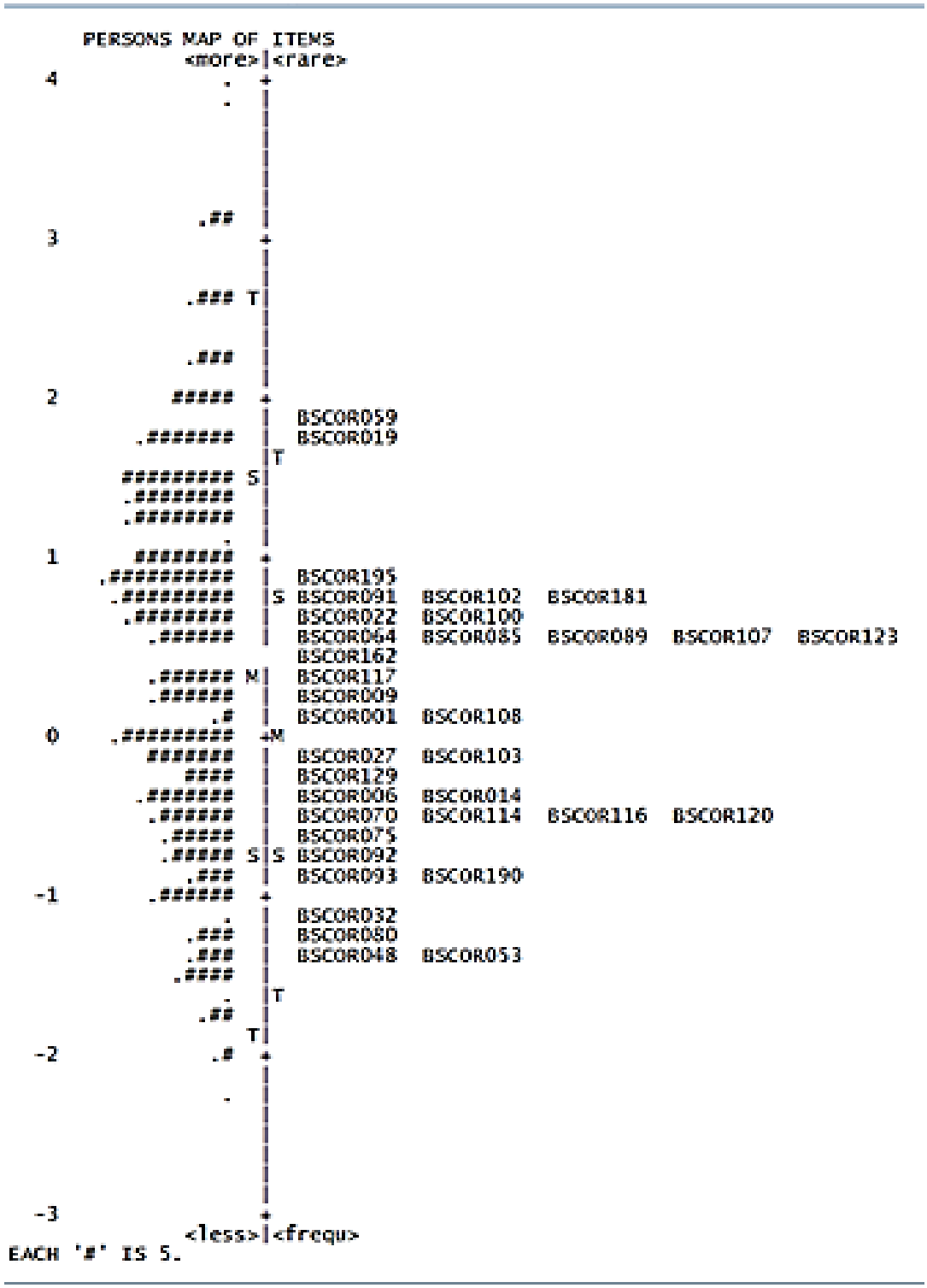

TYPE: 5
Mean = .5334
Std. Dev. = .2009
N = 34
Frequency
P-value
0
2
4
6
8
.0000
.2000
.4000
.6000
.8000
1.0000


TYPE: 6
Mean = .6805
Std. Dev. = .1488
N = 29
Frequency
P-value
0
1
2
3
4
5
.2000
.4000
.6000
.8000
1.0000


All items
Mean = .5945
Std. Dev. = .1857
N = 199
Frequency
P-value
0
5
10
15
20
.0000
.2000
.4000
.6000
.8000
1.0000

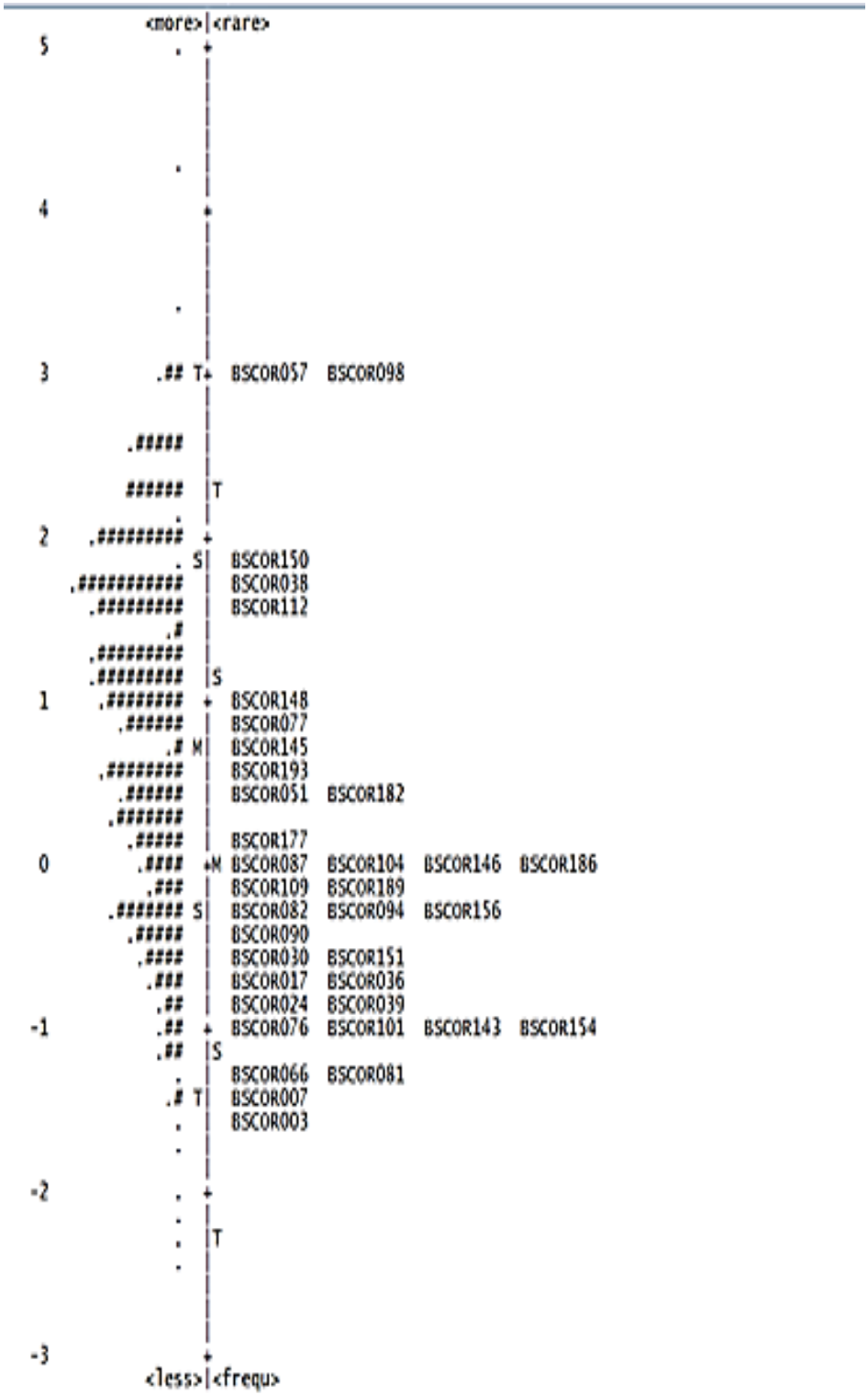
Appendix C: Person-Item maps for the *new items*

Person-Item Map for the *new item* type 1

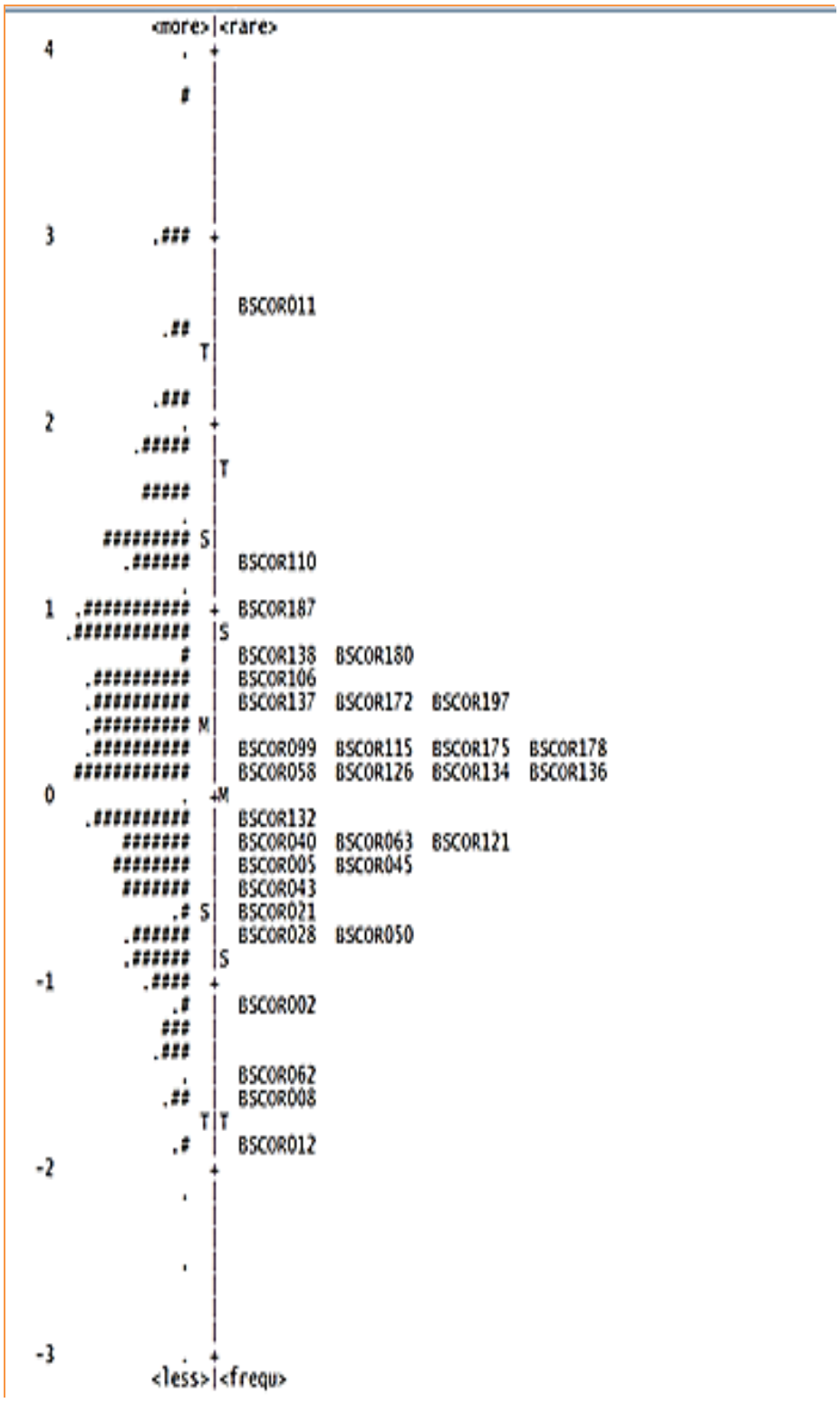


Person-Item Map for the *new item* type 2

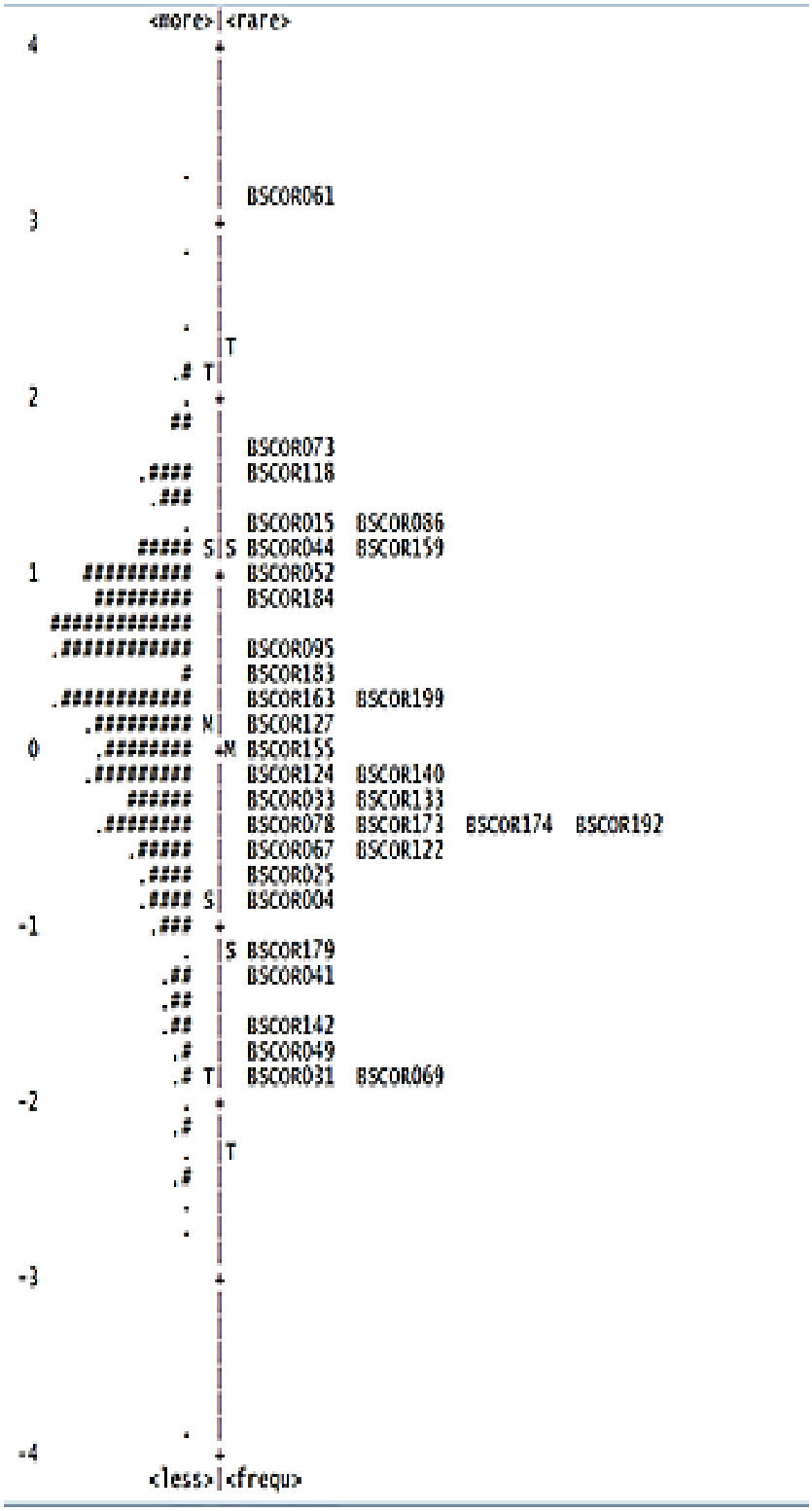
Person-Item Map for the *new item* type 3



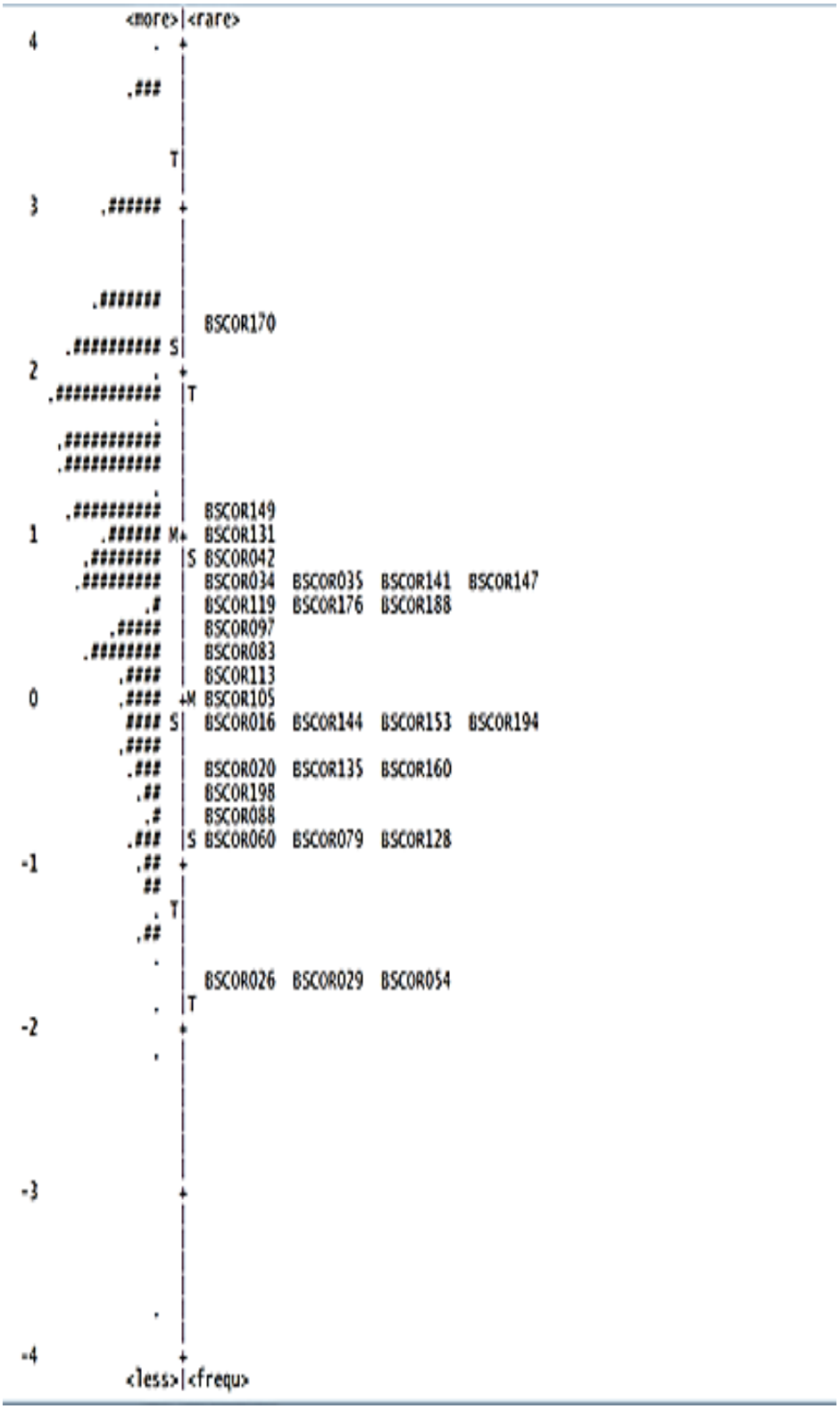
Person-Item Map for the *new item* type 4



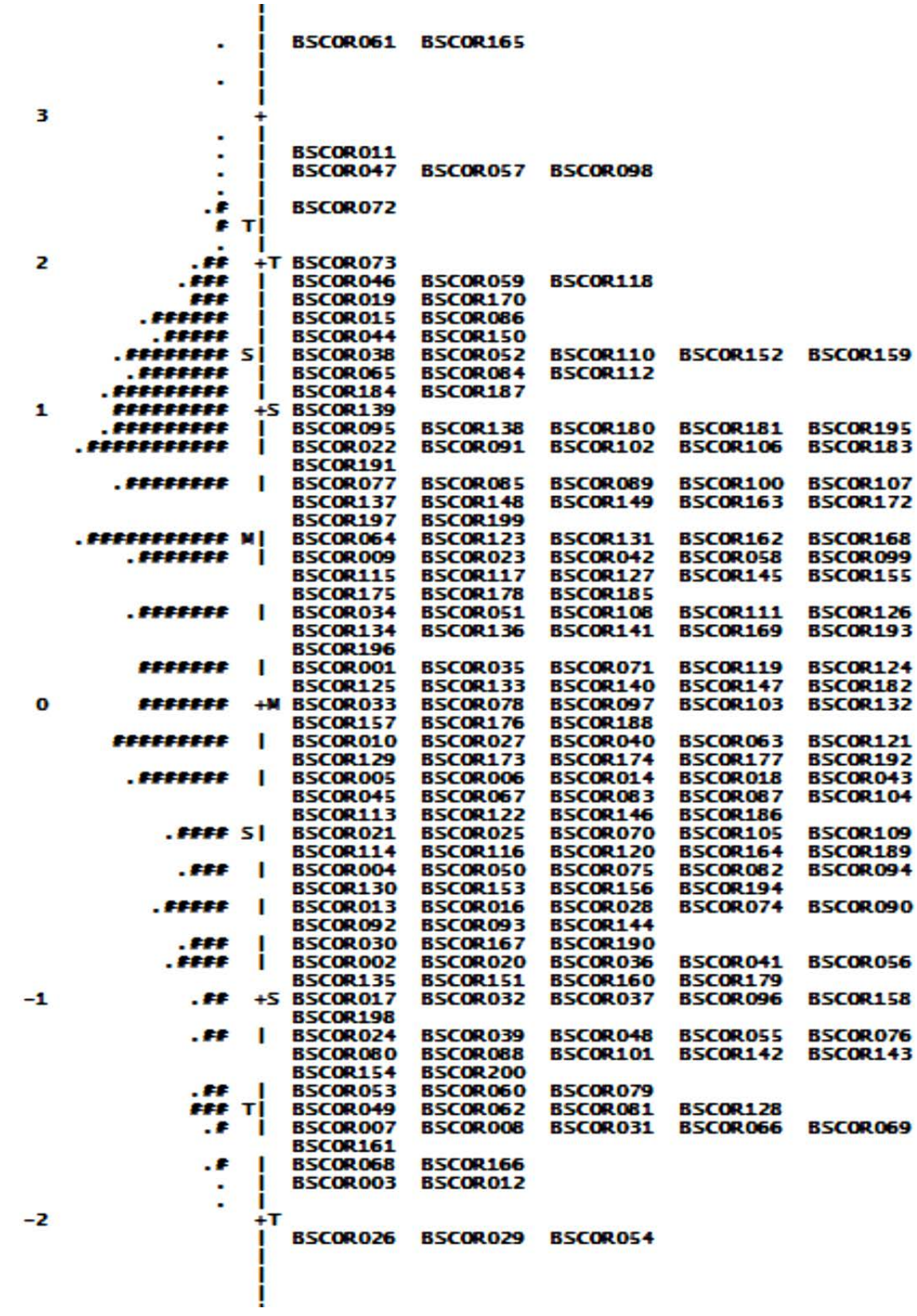
Person-Item Map for the *new item* type 5



Person-Item Map for the *new item* type 6



Person-Item Map for all *new items*



Appendix D: Summary of measured items for all the new item types

Summary of *new item* Type 1 measured items

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
11	211	902	1.89	.09	1.22	4.6	1.46	4.8	.21	75.1	79.0	BSCOR059
5	235	900	1.72	.08	1.09	2.2	1.26	3.3	.32	76.2	77.0	BSCOR019
35	376	902	.84	.08	1.25	7.7	1.30	5.8	.25	58.7	70.1	BSCOR195
33	386	900	.78	.08	.96	-1.2	.98	-.4	.48	72.0	69.9	BSCOR181
18	390	899	.76	.08	.89	-3.7	.88	-2.7	.53	75.3	69.9	BSCOR091
22	397	901	.72	.08	1.11	3.5	1.17	3.6	.36	65.0	69.8	BSCOR102
6	413	901	.63	.08	1.06	2.0	1.08	1.8	.40	66.6	69.6	BSCOR022
21	416	901	.61	.08	.98	-.8	.96	-.8	.47	71.3	69.5	BSCOR100
17	426	898	.55	.08	.91	-3.0	.89	-2.6	.52	74.4	69.5	BSCOR089
24	429	900	.54	.07	.91	-3.1	.88	-3.0	.52	73.0	69.5	BSCOR107
16	434	901	.52	.07	1.48	9.9	1.68	9.9	.05	51.2	69.5	BSCOR085
32	443	900	.46	.07	.94	-2.2	.99	-.3	.49	73.2	69.4	BSCOR162
12	445	896	.44	.08	.96	-1.5	.95	-1.3	.49	70.9	69.5	BSCOR064
30	446	897	.44	.08	1.02	.7	.99	-.1	.44	68.4	69.5	BSCOR123
28	466	896	.33	.08	.94	-2.2	.89	-2.5	.51	71.3	69.6	BSCOR117
3	481	900	.25	.08	1.03	.9	1.01	.3	.43	67.7	69.7	BSCOR009
25	504	899	.11	.08	.98	-.6	.95	-1.1	.47	70.4	70.1	BSCOR108
1	510	901	.08	.08	1.00	-.1	.96	-.8	.46	68.8	70.2	BSCOR001
23	547	892	-.15	.08	1.18	5.4	1.16	3.0	.30	63.3	71.3	BSCOR103
7	556	901	-.18	.08	.96	-1.2	.98	-.4	.47	74.4	71.5	BSCOR027
31	573	901	-.28	.08	.90	-3.0	.86	-2.6	.52	75.8	72.1	BSCOR129
4	581	900	-.34	.08	1.02	.6	1.03	.6	.42	71.3	72.4	BSCOR014
2	591	903	-.38	.08	1.00	.0	.91	-1.5	.45	72.0	72.8	BSCOR006
13	604	900	-.47	.08	.97	-.9	.92	-1.4	.46	75.1	73.5	BSCOR070
26	612	901	-.52	.08	.85	-4.3	.77	-4.0	.55	78.1	73.8	BSCOR114
27	613	898	-.53	.08	.81	-5.7	.79	-3.5	.57	81.0	74.0	BSCOR116
29	617	896	-.56	.08	.93	-2.1	.87	-2.1	.48	76.8	74.2	BSCOR120
14	634	903	-.65	.08	.93	-1.8	.90	-1.5	.47	77.3	74.9	BSCOR075
19	641	901	-.71	.08	.81	-5.2	.82	-2.7	.55	82.0	75.4	BSCOR092
20	659	901	-.83	.08	.79	-5.5	.68	-4.8	.57	82.3	76.5	BSCOR093
34	669	901	-.90	.08	.87	-3.3	.83	-2.2	.51	81.1	77.2	BSCOR190
8	709	904	-1.18	.09	1.12	2.5	1.21	2.1	.28	76.7	79.9	BSCOR032
15	713	898	-1.24	.09	1.01	.3	.97	-.3	.37	80.2	80.6	BSCOR080
9	724	896	-1.35	.09	1.04	.7	1.03	.3	.34	80.8	81.7	BSCOR048
10	732	901	-1.39	.09	1.04	.8	1.02	.3	.33	80.8	82.0	BSCOR053
MEAN	518.5	898.7	.00	.08	1.00	-.3	1.00	-.2		73.1	73.0	
S.D.	130.0	2.4	.79	.01	.14	3.5	.19	3.0		6.6	3.9	

Summary of *new item* Type 2 measured items

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
26	55	897	3.45	.14	1.11	.9	3.69	8.7	-.11	94.0	93.9	BSCOR165
7	125	894	2.46	.10	1.16	2.3	2.41	8.9	.00	86.2	86.2	BSCOR047
13	145	902	2.28	.10	1.13	2.2	1.85	6.6	.11	83.7	84.2	BSCOR072
6	211	896	1.75	.08	1.10	2.4	1.55	6.3	.18	76.5	77.6	BSCOR046
21	278	897	1.32	.08	.91	-2.6	.97	-.6	.44	76.7	72.2	BSCOR152
10	303	900	1.17	.08	1.30	8.7	1.58	9.6	.05	61.6	70.8	BSCOR065
15	309	900	1.14	.08	1.22	6.8	1.33	5.8	.15	62.2	70.4	BSCOR084
20	359	899	.85	.07	.99	-.4	1.11	2.4	.38	70.7	68.2	BSCOR139
32	392	899	.67	.07	.89	-4.5	.87	-3.3	.50	72.6	67.3	BSCOR191
29	446	902	.39	.07	.90	-3.8	.91	-2.5	.49	72.1	66.8	BSCOR168
31	460	899	.31	.07	1.07	2.6	1.11	2.8	.33	64.0	67.0	BSCOR185
4	476	900	.23	.07	1.03	1.2	1.02	.7	.37	64.9	67.1	BSCOR023
17	485	898	.18	.07	.93	-2.9	.90	-2.8	.48	69.9	67.3	BSCOR111
30	488	895	.15	.07	.97	-1.1	.96	-1.0	.43	68.2	67.4	BSCOR169
33	494	896	.12	.07	.99	-.5	1.00	-.1	.41	68.1	67.5	BSCOR196
18	508	895	.05	.07	1.01	.5	1.03	.8	.38	67.9	67.8	BSCOR125
12	529	899	-.06	.07	.96	-1.3	.92	-1.9	.44	68.6	68.5	BSCOR071
22	539	895	-.12	.07	1.00	-.1	.98	-.5	.40	69.2	69.0	BSCOR157
1	557	901	-.20	.07	1.19	6.0	1.23	5.1	.21	61.5	69.5	BSCOR010
3	582	899	-.36	.08	1.11	3.5	1.15	3.0	.28	65.7	71.0	BSCOR018
25	616	897	-.55	.08	.86	-4.3	.77	-4.7	.53	76.6	72.8	BSCOR164
19	640	901	-.70	.08	.89	-2.9	.84	-3.0	.49	78.6	74.4	BSCOR130
2	647	903	-.74	.08	.96	-1.1	.93	-1.1	.42	75.1	74.9	BSCOR013
14	659	901	-.82	.08	.92	-1.9	.86	-2.3	.45	76.9	75.8	BSCOR074
28	670	898	-.90	.08	.87	-3.3	.78	-3.6	.50	78.4	76.8	BSCOR167
9	682	902	-.98	.08	.94	-1.3	.86	-2.2	.43	78.9	77.7	BSCOR056
16	705	903	-1.14	.09	.90	-2.1	.86	-1.9	.45	81.7	79.6	BSCOR096
23	703	897	-1.16	.09	.86	-3.1	.74	-3.6	.50	81.9	79.8	BSCOR158
5	711	902	-1.19	.09	1.00	.1	1.02	.3	.35	80.2	80.2	BSCOR037
8	714	900	-1.22	.09	.82	-4.0	.68	-4.4	.54	83.7	80.6	BSCOR055
34	714	899	-1.23	.09	.91	-1.8	.86	-1.7	.44	82.5	80.7	BSCOR200
24	758	897	-1.62	.10	.84	-2.6	.66	-3.8	.49	86.3	85.1	BSCOR161
27	771	899	-1.73	.10	.84	-2.5	.61	-4.1	.49	86.7	86.2	BSCOR166
11	775	897	-1.79	.10	1.03	.5	1.09	.8	.27	86.8	86.7	BSCOR068
MEAN	514.9	898.8	.00	.08	.99	-.3	1.12	.4		75.3	75.0	
S.D.	194.9	2.4	1.24	.01	.12	3.1	.57	4.1		8.4	7.2	

Summary of *new item* Type 3 measured items

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
19	122	902	3.02	.10	1.20	2.9	2.55	8.5	.03	86.8	86.8	BSCOR098
10	125	901	2.99	.10	1.30	4.2	3.48	9.9	-.10	86.0	86.5	BSCOR057
28	267	904	1.86	.08	1.12	3.3	1.43	5.8	.28	71.5	74.2	BSCOR150
7	284	900	1.75	.08	1.17	4.6	1.42	6.0	.24	70.9	73.0	BSCOR038
23	318	899	1.54	.08	.99	-.2	1.03	.5	.42	73.5	71.3	BSCOR112
27	422	894	.93	.07	1.05	1.8	1.09	2.2	.39	66.6	69.1	BSCOR148
13	434	900	.88	.07	1.12	3.9	1.16	3.6	.35	65.9	69.3	BSCOR077
25	468	896	.69	.07	1.05	1.7	1.08	1.8	.40	69.1	69.4	BSCOR145
36	496	901	.53	.08	1.00	.0	1.00	.1	.45	71.2	70.0	BSCOR193
9	505	904	.49	.08	1.26	7.9	1.39	8.1	.22	59.6	70.0	BSCOR051
33	515	899	.43	.08	1.03	1.0	1.02	.6	.42	67.1	70.3	BSCOR182
32	559	897	.16	.08	.88	-3.8	.83	-3.6	.54	76.5	71.8	BSCOR177
34	578	902	.06	.08	.85	-4.6	.79	-4.2	.56	76.7	72.5	BSCOR186
21	582	897	.02	.08	.96	-1.2	.94	-1.1	.47	74.3	72.9	BSCOR104
16	586	900	.01	.08	1.07	1.9	1.05	1.0	.39	69.1	72.9	BSCOR087
26	586	897	.00	.08	.95	-1.5	.91	-1.7	.48	75.0	73.0	BSCOR146
22	602	901	-.09	.08	.96	-1.1	.93	-1.2	.47	74.9	73.6	BSCOR109
35	617	899	-.19	.08	.96	-1.2	.92	-1.3	.47	74.9	74.4	BSCOR189
15	622	896	-.23	.08	1.05	1.4	1.03	.4	.39	72.7	74.8	BSCOR082
18	626	898	-.25	.08	.84	-4.5	.75	-4.2	.56	80.6	74.8	BSCOR094
31	633	895	-.32	.08	.91	-2.3	.98	-.3	.49	78.1	75.3	BSCOR156
17	658	903	-.45	.08	.86	-3.5	.77	-3.5	.53	80.8	76.4	BSCOR090
5	678	901	-.59	.09	.95	-1.1	.93	-.9	.44	79.6	77.8	BSCOR030
29	678	900	-.60	.09	.83	-4.2	.72	-4.0	.55	81.8	77.9	BSCOR151
6	696	903	-.72	.09	.87	-2.9	.79	-2.6	.50	83.0	79.1	BSCOR036
3	703	900	-.78	.09	.90	-2.1	.88	-1.4	.46	83.0	79.8	BSCOR017
4	717	902	-.89	.09	1.05	1.0	1.00	.1	.35	79.6	80.8	BSCOR024
8	719	901	-.91	.09	1.11	2.1	1.08	.9	.30	79.2	81.0	BSCOR039
20	725	902	-.96	.09	.85	-3.1	.70	-3.5	.51	83.9	81.5	BSCOR101
30	722	897	-.96	.09	.87	-2.6	.75	-2.8	.49	83.7	81.6	BSCOR154
24	722	893	-.98	.09	.86	-2.8	.71	-3.2	.50	83.6	81.8	BSCOR143
12	725	897	-.98	.09	.95	-.9	.86	-1.4	.42	82.0	81.8	BSCOR076
14	753	900	-1.22	.10	.97	-.5	.85	-1.4	.39	84.5	84.2	BSCOR081
11	765	900	-1.33	.10	.89	-1.9	.79	-1.9	.44	86.5	85.3	BSCOR066
2	771	904	-1.37	.10	1.05	.8	1.25	2.0	.28	86.5	85.6	BSCOR007
1	784	899	-1.55	.11	.93	-1.0	.82	-1.4	.39	88.0	87.4	BSCOR003
MEAN	575.8	898.6	.00	.09	.99	-.2	1.08	.2		77.4	76.9	
S.D.	172.6	2.8	1.12	.01	.12	2.9	.52	3.6		7.0	5.6	

Summary of *new item* Type 4 measured items

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
4	113	902	2.68	.11	1.12	1.6	1.47	3.3	.20	87.5	88.1	BSCOR011
17	296	902	1.23	.08	1.01	.2	1.04	.7	.40	72.7	72.7	BSCOR110
30	325	898	1.05	.08	1.08	2.6	1.17	3.4	.33	69.2	70.8	BSCOR187
25	369	895	.80	.07	.98	-.8	.97	-.7	.44	69.4	68.8	BSCOR138
29	374	897	.77	.07	.92	-3.0	.87	-3.0	.49	71.0	68.7	BSCOR180
16	410	898	.58	.07	.96	-1.6	.95	-1.2	.45	70.2	67.8	BSCOR106
31	424	901	.51	.07	.96	-1.4	.96	-1.0	.45	69.0	67.6	BSCOR197
24	425	895	.49	.07	1.07	2.5	1.09	2.3	.35	65.5	67.5	BSCOR137
26	428	897	.47	.07	1.24	8.4	1.38	8.4	.19	58.0	67.6	BSCOR172
28	466	904	.29	.07	.92	-3.1	.90	-2.5	.48	73.2	67.4	BSCOR178
15	465	898	.28	.07	.88	-4.8	.86	-3.7	.52	72.9	67.5	BSCOR099
18	466	894	.27	.07	.99	-.3	1.00	-.1	.42	67.0	67.4	BSCOR115
27	481	899	.19	.07	1.05	1.7	1.04	.9	.37	64.9	67.6	BSCOR175
12	484	901	.18	.07	1.22	7.8	1.33	7.1	.20	58.8	67.6	BSCOR058
20	484	893	.17	.07	.97	-1.0	1.03	.7	.43	70.1	67.6	BSCOR126
22	491	896	.13	.07	1.09	3.4	1.11	2.6	.32	63.2	67.7	BSCOR134
23	498	896	.10	.07	.95	-2.0	.91	-2.2	.46	70.2	67.8	BSCOR136
21	536	895	-.11	.07	.96	-1.6	.90	-2.1	.45	69.4	68.9	BSCOR132
8	557	902	-.21	.08	1.07	2.2	1.02	.4	.35	66.0	69.6	BSCOR040
19	557	899	-.23	.08	.92	-2.7	.97	-.7	.46	72.7	69.7	BSCOR121
14	560	901	-.23	.08	.98	-.6	.97	-.5	.41	72.1	69.7	BSCOR063
2	584	899	-.38	.08	1.11	3.4	1.11	1.9	.30	67.6	70.9	BSCOR005
10	589	902	-.40	.08	.91	-3.1	.88	-2.3	.47	74.9	71.1	BSCOR045
9	599	903	-.46	.08	.94	-1.9	.93	-1.2	.44	74.2	71.6	BSCOR043
6	621	905	-.58	.08	1.03	.9	1.03	.5	.35	72.1	72.8	BSCOR021
11	640	905	-.70	.08	.97	-.8	1.00	.0	.40	74.7	74.0	BSCOR050
7	646	902	-.75	.08	.99	-.2	1.02	.4	.37	74.9	74.6	BSCOR028
1	696	904	-1.08	.09	.98	-.4	1.06	.8	.35	79.0	78.4	BSCOR002
13	750	900	-1.54	.10	.85	-2.8	.67	-3.4	.46	85.2	83.7	BSCOR062
3	761	904	-1.61	.10	.88	-2.1	.82	-1.7	.41	85.5	84.5	BSCOR008
5	787	903	-1.88	.10	.88	-1.7	.70	-2.5	.40	87.6	87.3	BSCOR012
MEAN	511.3	898.7	.00	.08	1.00	.0	1.00	.1		71.9	71.8	
S.D.	141.7	3.4	.88	.01	.09	2.9	.17	2.7		7.1	6.0	

Summary of *new item* Type 5 measured items

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
10	58	902	3.18	.14	1.10	.9	3.38	8.0	-.04	93.6	93.6	BSCOR061
13	187	902	1.72	.09	1.06	1.3	1.20	2.3	.24	79.3	80.0	BSCOR073
17	206	900	1.57	.08	.99	-.1	1.12	1.5	.31	79.2	78.1	BSCOR118
2	255	899	1.25	.08	1.07	2.0	1.43	5.9	.23	74.5	73.7	BSCOR015
15	257	899	1.24	.08	1.10	2.8	1.58	7.8	.18	73.9	73.6	BSCOR086
7	265	900	1.18	.08	1.09	2.5	1.42	6.0	.22	73.1	72.9	BSCOR044
25	281	896	1.08	.08	1.13	3.9	1.37	5.6	.21	68.1	71.6	BSCOR159
9	283	896	1.07	.08	1.10	3.0	1.36	5.6	.23	70.6	71.4	BSCOR052
32	327	895	.82	.07	.97	-1.2	1.05	1.0	.38	72.1	68.5	BSCOR184
16	383	897	.51	.07	1.09	3.7	1.15	3.6	.28	63.7	66.4	BSCOR095
31	410	899	.37	.07	.97	-1.1	1.01	.4	.40	69.3	66.1	BSCOR183
26	414	899	.35	.07	1.07	2.7	1.06	1.5	.32	61.8	66.0	BSCOR163
34	436	898	.23	.07	1.01	.3	1.01	.2	.38	65.6	66.1	BSCOR199
20	464	901	.09	.07	1.07	2.8	1.05	1.4	.33	61.7	66.5	BSCOR127
24	481	897	-.01	.07	.96	-1.7	.92	-2.1	.44	67.1	66.9	BSCOR155
19	511	900	-.16	.07	.92	-2.9	.92	-2.0	.47	72.2	67.9	BSCOR124
22	517	896	-.20	.07	.98	-.8	.97	-.8	.42	68.8	68.2	BSCOR140
21	523	899	-.23	.07	1.03	.9	1.05	1.2	.37	67.2	68.4	BSCOR133
5	533	900	-.28	.07	1.03	.9	1.02	.4	.37	68.1	68.7	BSCOR033
14	553	900	-.39	.07	1.23	7.1	1.35	7.2	.17	60.4	69.6	BSCOR078
33	556	897	-.42	.08	.89	-3.7	.84	-3.7	.51	72.7	70.0	BSCOR192
29	564	902	-.44	.08	.91	-2.9	.91	-2.0	.48	73.9	70.1	BSCOR174
28	566	898	-.47	.08	.85	-4.9	.80	-4.7	.54	74.7	70.5	BSCOR173
11	579	897	-.54	.08	1.03	.9	1.05	1.1	.37	69.5	71.2	BSCOR067
18	591	896	-.62	.08	.90	-3.1	.85	-3.2	.50	74.6	72.0	BSCOR122
3	602	900	-.67	.08	1.04	1.1	1.03	.6	.36	70.8	72.6	BSCOR025
1	623	898	-.80	.08	.96	-1.0	.94	-1.1	.43	74.6	74.0	BSCOR004
30	680	900	-1.18	.08	.88	-2.9	.77	-3.7	.51	80.7	78.3	BSCOR179
6	692	900	-1.27	.09	.79	-4.8	.68	-4.9	.58	83.0	79.4	BSCOR041
23	722	898	-1.51	.09	.78	-4.5	.65	-4.7	.58	86.0	82.0	BSCOR142
8	746	898	-1.72	.10	.82	-3.2	.69	-3.5	.53	86.9	84.2	BSCOR049
12	759	899	-1.83	.10	.90	-1.6	.77	-2.4	.45	85.8	85.2	BSCOR069
4	769	901	-1.92	.10	.85	-2.5	.84	-1.6	.48	87.9	86.1	BSCOR031
MEAN	478.6	898.8	.00	.08	.99	-.2	1.10	.6		73.7	73.3	
S.D.	180.4	1.9	1.12	.01	.11	2.9	.46	3.8		8.0	6.8	

Summary of *new item* Type 6 measured items

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
26	237	897	2.33	.08	1.26	5.9	2.00	9.8	.15	73.8	77.1	BSCOR170
23	437	901	1.12	.08	1.16	5.3	1.26	5.2	.32	62.5	69.5	BSCOR149
18	454	899	1.01	.08	1.13	4.3	1.25	5.0	.34	66.6	69.6	BSCOR131
7	474	904	.91	.08	1.01	.2	1.04	.8	.45	69.0	69.8	BSCOR042
20	493	894	.78	.08	.99	-.2	.95	-1.0	.47	70.7	70.0	BSCOR141
5	496	901	.78	.08	1.14	4.4	1.20	4.0	.35	63.1	70.1	BSCOR034
22	513	899	.68	.08	.99	-.2	.98	-.5	.46	70.0	70.3	BSCOR147
6	515	902	.67	.08	1.08	2.5	1.06	1.3	.40	66.7	70.4	BSCOR035
16	522	896	.62	.08	.93	-2.2	.88	-2.5	.51	72.1	70.7	BSCOR119
27	534	903	.57	.08	1.00	.1	1.00	.0	.45	72.0	70.9	BSCOR176
28	541	896	.51	.08	1.04	1.2	1.10	1.9	.42	70.2	71.2	BSCOR188
13	544	899	.50	.08	1.12	3.6	1.13	2.4	.36	65.8	71.4	BSCOR097
11	580	897	.28	.08	1.05	1.6	1.04	.8	.41	70.0	72.8	BSCOR083
15	591	893	.20	.08	.99	-.2	1.01	.2	.44	74.5	73.4	BSCOR113
14	613	894	.06	.08	.99	-.4	.95	-.7	.45	72.8	74.6	BSCOR105
29	639	901	-.09	.08	.91	-2.3	.88	-1.7	.50	78.7	75.9	BSCOR194
24	635	892	-.10	.08	1.00	.1	1.08	1.2	.42	77.7	76.0	BSCOR153
21	642	895	-.14	.08	.83	-4.5	.73	-4.0	.56	80.8	76.3	BSCOR144
1	655	898	-.21	.08	.92	-2.0	.86	-2.0	.49	78.8	76.9	BSCOR016
19	683	903	-.39	.09	.88	-2.8	.84	-2.0	.50	81.6	78.7	BSCOR135
2	693	899	-.49	.09	1.02	.4	1.13	1.5	.38	81.0	79.6	BSCOR020
25	695	901	-.50	.09	.92	-1.9	.82	-2.2	.48	81.2	79.7	BSCOR160
30	706	897	-.60	.09	.85	-3.1	.74	-3.0	.51	82.4	80.7	BSCOR198
12	719	903	-.67	.09	.91	-1.8	.82	-1.9	.47	83.0	81.4	BSCOR088
9	738	903	-.84	.09	1.05	.8	.96	-.3	.36	81.7	82.9	BSCOR060
10	737	898	-.86	.10	.78	-4.3	.65	-3.6	.54	86.5	83.2	BSCOR079
17	743	896	-.93	.10	.83	-3.1	.68	-3.1	.50	86.7	83.9	BSCOR128
8	810	897	-1.70	.12	.98	-.3	.87	-.7	.33	90.6	90.4	BSCOR054
3	819	904	-1.73	.12	1.00	.1	.87	-.7	.31	90.9	90.7	BSCOR026
4	822	904	-1.77	.12	1.00	.0	1.09	.6	.30	91.3	91.0	BSCOR029
MEAN	606.3	895.9	.00	.09	.99	.0	1.00	.1		76.4	76.6	
S.D.	129.7	3.5	.91	.01	.11	2.7	.24	2.9		8.1	6.5	