THE DEVELOPMENT OF A SET OF GUIDELINES FOR THE REVISION OF PSYCHOLOGICAL TESTS AND THE USE OF REVISED PSYCHOLOGICAL TESTS

J.H. CRONJE

2020

The development of a set of guidelines for the revision of psychological tests and the use of revised psychological tests

By

Johan Herman Cronje

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy (Psychology)

In the Department of Psychology

Faculty of Health Sciences

At the Nelson Mandela University

April 2020

Promoter: Prof M. B. Watson

Co-Promoter: Prof L. A. Stroud

Declaration of Authenticity

I, Johan Herman Cronje (194112520), hereby declare that the thesis for the degree of Doctor of Philosophy (Psychology) is my own work and that it has not previously been submitted for assessment or completion of any postgraduate qualification to another University or for another qualification.

Johan Herman Cronje

Official use:

In accordance with Rule G5.6.3,

5.6.3. A treatise/dissertation/thesis must be accompanied by a written declaration on the part of the candidate to the effect that it is his/her own work and that it has not previously been submitted for assessment to another University or for another qualification. However, material from publications by the candidate may be embodied in a treatise/dissertation/thesis

Acknowledgements

I express my heartfelt appreciation for all those who assisted in making this study possible.

Thank you to God, who has brought me thus far, and who continues to shower me with blessings.

I thank my Promoters, Prof Mark Watson and Prof Louise Stroud, for guiding me throughout this study. I admire your ability and dedication to this project, your continued encouragement and faith in my ability to bring this vision to fruition. I could not have done this without you.

To my parents who have been my longest supporters on my academic journey. Thank you for believing in me even when I did not believe in myself.

To my partner Brian and our family. Thank you for supporting this study by taking care of our home and me. Thank you for pretending you find test revision more fascinating than anything else in the world.

Thank you to my colleagues in the Department of Psychology who accompanied me on this journey as they worked on and completed their own theses.

Thank you to the Association of Creative Thought for providing me with spiritual support and reminding me that we are all individualised expressions of God.

Thank you for the Association for Research in Infant and Child Development, and Drs Elizabeth Green and Paula McAlinden in particular, for their interest in this study, my progress, and my wellbeing throughout.

Thank you to Research Capacity Development at the Nelson Mandela University for allowing me time to focus on this study through the Teaching Relief Grant.

Abstract

The psychological test industry has produced a wide variety of psychological tests that are used by professionals to facilitate measurement and decision-making. Tests are updated and revised periodically in order to remain current, valid and reliable in what is a competitive psychometric industry. Despite the prevalence of test revisions, especially in recent years, a number of authors have commented on the lack of comprehensive guidelines for test revision. Guidelines should cover aspects such as what the different types of revision are, when to embark on a revision, what process to follow and how test users should use revised tests. Test revision differs from test construction in a number of ways. There are external factors that affect the regularity with which a test should be revised. Test revision also involves more role players than test construction, including the opinions of those test users who may be resistant to any change in the previous test edition. Finally, revised tests sometimes have to contend with requirements from the test publisher who purchased the test or distribution rights from the developer. Test revision is expensive and time consuming, which leaves little scope for experimentation or trial-and-error. The availability of expertise, as well as the human and financial resources required to complete test revisions can make such projects unaffordable, especially for professionals in developing countries, such as South Africa. It may be more feasible for such professionals to collaborate with international revision projects. By doing so they can gain experience in test revision, contribute indigenous information that could shape the revision of an international test, increase opportunities to engage with international users, and potentially source international funding for research in their own country. The current study developed a comprehensive and practical set of 30 guidelines to assist those involved in test revision. These guidelines were peer-reviewed and refined. Finally, the guidelines were field-tested using a case study of a recently revised

developmental test, the *Griffiths III*. Professionals from South Africa, including the present researcher, formed part of the international team for the extensive revision of the *Griffiths III*, which makes this test an ideal case study from both the perspectives of the developed test revision guidelines as well as collaboration of professionals from a developing country in an international test revision. The knowledge gained from the development of guidelines and international collaboration in test revision is reflected on.

Keywords: Test revision, developing guidelines, psychological tests

Table of Contents

List of Tables

List of Figures

Abbreviations and Acronyms

AERA	American Educational Research Association
ANC	African National Congress
APA	American Psychological Association
AOPPC	Assessment Oversight and the Personnel Psychology Centre,
ARICD	Association for Research in Infant and Child Development
BPS	British Psychological Society
CISA	Certified Information Systems Auditor
CBT	Computer-based testing
COTAN	Dutch Committee on Tests and Testing
CTT	Classical test theory
EFPA	European Federation of Psychologists' Associations
ETS	Educational Testing Service
GMDS-ER	Griffiths Mental Development Scales – Extended Revised
HPCSA	Health Professions Council of South Africa
ITC	International Test Commission
IRT	Item-response theory
MMPI	Minnesota Multiphasic Personality Inventory
NAS	National Academy of Sciences
NCME	National Council on Measurement in Education
PSI	Psychological Society of Ireland
SSAIS-R	Senior South African Individual Scale – Revised
UK	United Kingdom

USA	United States of America
WAIS	Wechsler Adult Intelligence Scale
WPPSI	Wechsler Preschool and Primary Scales of Intelligence

Chapter 1: Orientation and Research Review

Psychological tests are assessment tools created to provide an objective measurement of a specific construct or domain of functioning within test participants. As such, psychological tests can form an important component of a psychologist's toolbox. As with any created product, the usefulness of a psychological test can degrade over time. To combat the effects of time, tests are therefore periodically revised according to identified needs. A revision process creates a unique opportunity for revision teams to interrogate test components, make changes, and produce an updated, fit-for-purpose test. The present study relates to psychological test revision, how the revision process is undertaken, and what the professional responsibilities of test users are in the use of revised tests.

In this chapter, the background to the present study, the operational definitions, and the problem statement are discussed. The aim and objectives of the research are also presented, together with an overview of the remaining chapters of this thesis.

Background to the Study

In literature, many terms are used when referring to psychological tests. Some of these include psychometric test, tool, measurement, scale, questionnaire, inventory, and instrument (Swanepoel & Krüger, 2011). Whilst each of these terms has conceptual differences, they are all products created to assist psychologists in forming a clinical opinion of a test taker. As with any created product, psychological tests can become outdated over time and consequently require updating in order to remain useful (Wiberg & von Davier, 2017). This process is known as test revision.

There are many components to a psychological test, including underpinning constructs, test questions, test instructions, equipment, test manuals, standardisation information, and test norms

(Coaley, 2009; Groth-Marnat, 2009). All these components may become outdated at some stage in a test's lifespan, thereby signalling a juncture at which the test should be revised. The number of changes required determines the extent of a revision, with the revision process broadly falling within a 'light', 'medium', or 'extensive' revision classification (Butcher, 2000).

A benefit in revising a psychological test is that it has undergone much public scrutiny by both test users and researchers after years of use in different settings and a range of independent research studies. Any test can be improved on and a revision is the ideal opportunity for such changes to be effected (Brannigan & Decker, 2006). A drawback of revising a psychological test is that the original test creates a direct benchmark that the revised edition will be measured against. Further, some test users become reliant on a specific test and are resistant to changing to the revised edition.

Test revision is complex and requires considerable expertise and resources. The expense of 'medium' or 'extensive' test revisions can exceed the resources available to many projects, and in particular those in developing countries such as South Africa. Professionals in the latter countries need to consider participation in international test revision projects that would allow them to inform the revision of tests that may be used in their own countries.

Given the complexity of updating an existing test, guidelines would be helpful for revision teams and for users of revised psychological tests. An analysis of guidelines from notable international organisations, such as the International Test Commission (ITC), Educational Testing Service (ETS), American Educational Research Association (AERA) and the American Psychological Association (APA) however revealed a lack of comprehensive guidelines dedicated to the revision or use of revised psychological tests. A review of 280 guidelines for psychological tests from guideline documents that also mention test revision, found that only 17

(6.1%) of the guidelines specifically referred to test revision or the use of revised tests. What makes this even more concerning is the open call for such guidelines from authors (Adams, 2000; Butcher, 2000; Strauss, Spreen, & Hunter, 2000) in a special edition of the APA journal, *Psychological Assessment.* This lack of attention from international organisations that set the benchmarks for professional practice in the discipline of psychology, as well as the subdiscipline of psychological testing, implies that test revisions have had to continue within this vacuum of industry-produced guidelines.

One reason for this is the assumption that the processes of test development and test revision are essentially similar, meaning that test development guidelines should also apply to test revision (Bush, 2010). This assumption fails to understand however that test development occurs in an expectation vacuum whereas test revision does not. A revised test faces an immediate benchmark in its predecessor, and the expectations of a community of test users that have been informed by the use of the previous test. A test revision team therefore has to tread carefully as it endeavours to modernise, innovate and update an existing test. The opinions of existing test users will be important to the economic survival of the revised test, and test users will have different suggestions and expectations about changes in a test they have become familiar with and proficient in using. As test revisions increase in frequency, it is imperative that a set of guidelines are created to assist professionals involved in test revision.

In general, guidelines exist to provide direction to professionals on the discipline-specific standards they need to adhere to (Jaeschke, Jankowski, Brozek, & Antonelli, 2009). Such guidelines are indispensable to newer members of a profession as they acquire the skills to progress competently in their work duties. Guidelines also serve to protect the autonomy and reputation of a profession by providing methods of internal checks and balances within the

community (Weisz et al., 2007). Given the importance of guidelines in general, they need to be developed by means of a rigorous and transparent process. An analysis of the procedures used to develop guidelines in the discipline of psychology revealed however that many guideline documents do not stipulate the process through which their proposed guidelines were developed. This means that readers are not able to gauge the validity of the guidelines from the rigour of the process or from a supporting evidence-base. The implication of such guideline documents is that readers are expected to adhere to guidelines based on the reputation of the authors or the publishing organisation, rather than the evidence reviewed or the academic rigour of the development process followed.

Problem Statement

There appears to be limited guidelines available on psychological test revision, with those documents that contain guidelines for test revision only mentioning this topic in 6.1% of their guidelines about psychological tests (AERA, 2014; ETS, 2015; ITC, 2015). The guidelines that exist primarily address the relationship between test publishers and users, operational aspects related to the rollout of revised tests, and the roles and obligations of test users. The present researcher has experience in both test revision and test use. From my perspective and experience, the existing guidelines are too fragmented and inadequate to meet the needs of practitioners concerning the process of conducting a test revision, the key markers of such a project, and the potential pitfalls that revision teams need to guard against. The reviewed extant guidelines also separate revision teams from existing test users. However, test users are important sources of experience and information, and may have valid opinions on the future direction of a test. A successful revision process should not only produce an improved test, but a strengthened relationship between test users, revision teams, and test publishers. This relationship will

increase the likelihood of a revised test being accepted by existing users of the previous edition (Adams, 2000). I therefore felt that there was a need for a more comprehensive set of guidelines for test revision that spans the lifecycle of a revision project, to provide clear and comprehensive guidance for test users and revision teams. Important aspects such as ethical considerations for revision, procedural arrangements, resource allocation, and management of different role players would be necessary for test revision guidelines to be of practical use to practitioners. This was the intention of the present study, as reflected in the research aim and objectives below.

Research Aim and Objectives

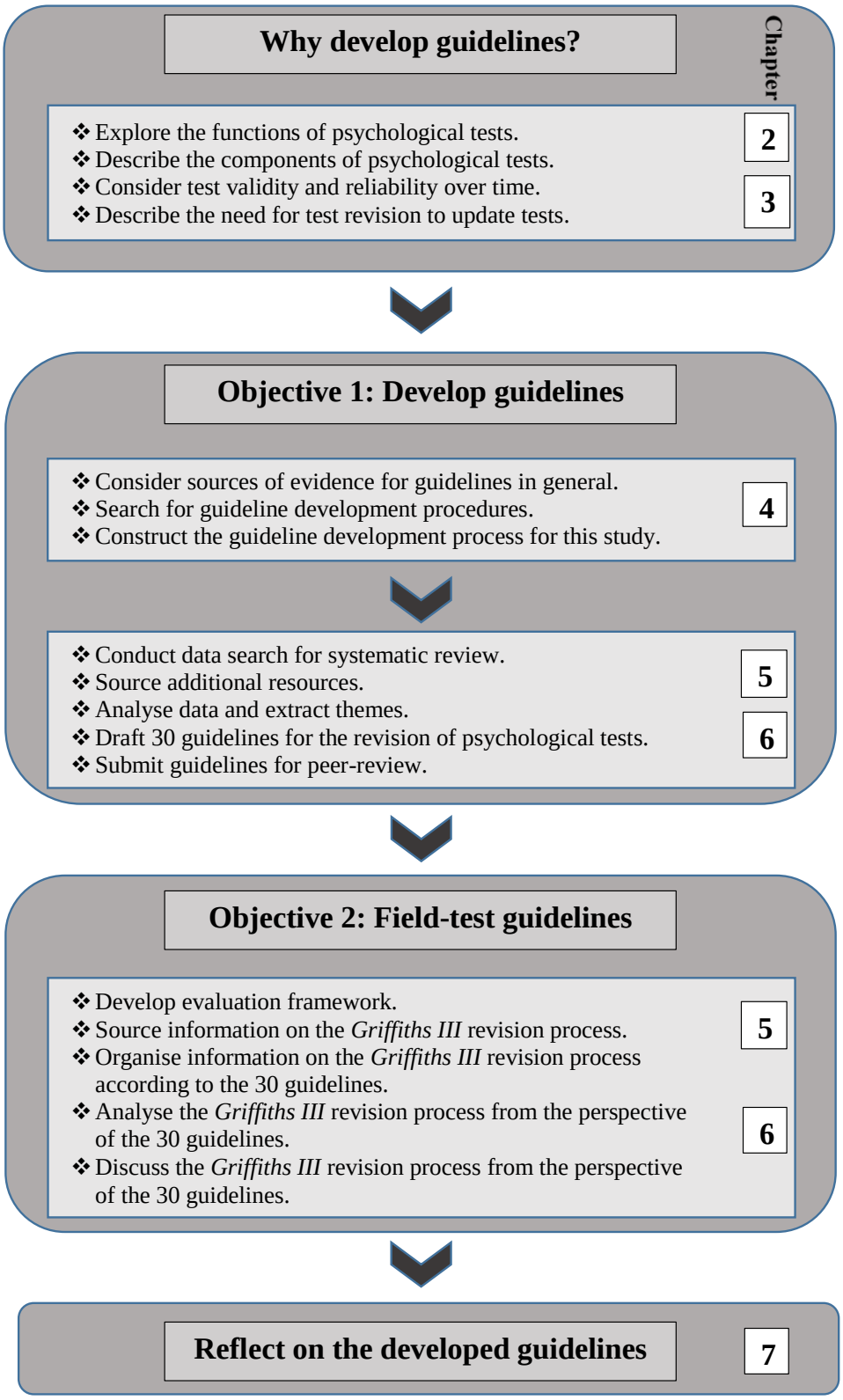
The aim of the present study was to develop a comprehensive set of guidelines for test revision that covers the full process of test revision, including the use of revised tests, and to field-test the proposed guidelines using the case example of a revised psychological test, the *Griffiths III*. Thus the purpose of this study was not only to construct guidelines but, by examining an extant revision, to critique the revision process of the *Griffiths III* in order to develop a clearer understanding of how the proposed guidelines could operate in practice. The *Griffiths III* was launched in 2016, making it a recent example of a revised test. In addition, the present researcher formed part of a team of South African psychologists who worked on the revision of the *Griffiths III*, thus providing an opportunity for the researcher to reflect on the collaboration of a professional from a developing country in an international test revision.

The objectives of the study were:

1. To develop guidelines for the revision and use of all types of revised psychological tests.
2. To explore the test revision guidelines with a specific psychological test revision.

A map of this study's research process is displayed in Figure 1. In particular, the process detailed in chapters four to seven must be emphasised.

Figure 1. Research process for the present study



Definition of Key Terms

According to the International Test Commission (ITC, 2013a), it is a challenge to find consensus on the exact meaning and definition of many terms used within psychological testing. One reason for this is that the subdiscipline of psychological testing is an applied science that is ever changing and expanding in response to research, technological innovation, the discipline of psychology itself, and society in general. In the present study, the following definitions have been accepted for key terms:

Psychological test: Foxcroft and Roodt (2013) define a psychological test as an "objective, standardised measure that is used to gather data for a specific purpose (e.g. to determine what a person's intellectual capacity is)" (p. 5).

Psychological testing: Within the discipline of psychology, the subdiscipline of psychological testing has traditionally been called psychometrics. Psychometrics refers to the development of psychological tests by test developers, as well as the administration, scoring and interpretation of test takers' results by a trained professional. More recently, the term psychometrics has been used less frequently, and is increasingly being replaced in the literature by the term *psychological testing* (International Test Commission, 2013a).

Test revision: This is the process of making changes to an existing psychological test. It is an overarching term for all processes related to effecting changes to any component of a test, such as test questions, equipment, instructions, and test norms (Adams, 2000; Bush, 2010; Liu & Dorans, 2013).

Practice guidelines: Proctor and Staudt (2003) define guidelines as "systematically compiled and organised knowledge statements to help practitioners select and use the most effective and appropriate interventions for attaining desired outcomes" (p. 209).

Contributions of the Present Study

The present study represents unique contributions, particularly in six areas. The first is the development of a framework for the revision of psychological tests consisting of ten phases in a revision project. The second is the tailoring of a process to develop guidelines that could be applied to other fields in the discipline of psychology in general and, more specifically, its subdiscipline of psychological testing. The third contribution is the development of 30 guidelines that span the ten phases of a revision project, for the revision and use of revised psychological tests. These guidelines go beyond the guideline statements scattered throughout documents of national and international psychological testing organisations or the contributions of authors in articles and textbooks. The 30 guidelines developed by the present researcher are based on a synthesis of literature, as well as the researcher's decades of learning and experience in psychological testing and test revision. The guidelines developed in this study are organised according to the lifespan of a test revision, as envisaged by the present researcher. It is also the first guideline document internationally that is dedicated solely to the revision of psychological tests including the use of revised psychological tests. The fourth contribution is that this is the first known analysis of a recently revised developmental test, the *Griffiths III*, to demonstrate the usefulness of the guidelines developed in this study in analysing a test revision process retrospectively. The analysis of the *Griffiths III* further supports the value of the guidelines for practitioners involved in current and future test revision projects. The fifth contribution relates to the research design of this study. As the combination of two qualitative research methods in a mixed method study is still a developing field, the present study serves as an applied example of this research design, particularly in how a study can combine two qualitative methods to develop guidelines within one method and then to field-test it with another method. The sixth

contribution is the present researcher's reflection of his involvement in the revision of the *Griffiths III*. Being from a developing country, South Africa, I reflect on my contribution in an international test revision project, in terms of how this contributed to my professional development as well as how I could have contributed more to the project to further cross-cultural research on the test during its development.

Overview of Chapters

Chapter One orientates the reader to the topic of the revision of psychological tests and presents an overview of the relevant literature, the formulation of the problem that led to the present research, as well as the aim and specific objectives of the study.

Chapter Two explores psychological testing from conceptualisation as a structured process of observation to operationalisation as a test. The benefits of psychological tests as well as criticisms against them are presented. Finally, the key components of tests, such as test equipment, manuals, and test questions are discussed.

Chapter Three focusses on the meaning of test revision, the reasons for such a revision, and specifically the effects of change over time on the accuracy of psychological tests. The relationship between different role players during test revision is also explored. The researcher then presents a suggested process for test revision. Finally, test revision in South Africa is explored within the historical context of the country, to highlight the added challenges faced by professionals interested in test revision, but who come from a developing country.

Chapter Four details the importance of guidelines within professions. The processes by which guidelines have been developed historically are presented. Guidelines within the subdiscipline of psychological testing and especially those concerning test revision are then

discussed. Finally, the problem that motivated the present study is formulated, and the aim and objectives of the study are delineated.

Chapter Five focuses on the research design and the procedures followed with respect to sampling, data collection, analysis, and guideline development. For the case study of a recently revised developmental measure, the *Griffiths III* Scales of Child Development is discussed as well as the revision history of the test since the creation of the first *Baby Scales* in 1954. Developmental measures are developed to test constructs from the subdiscipline of developmental psychology. I provide additional information about my professional career as well as my involvement in the revision of the *Griffiths III*, and the journey I have taken as the researcher of this study and as the developer of the present guidelines for test revision. The steps taken to enhance the trustworthiness of the findings of the study are explained, as well as the legal and ethical considerations that guided the research.

Chapter Six presents the findings of the thematic review, the guidelines for the revision and use of revised psychological tests, and the case study of the *Griffiths III*. The chapter discusses the research findings against the documents selected for data analysis, as well as other relevant literature.

Chapter Seven describes the conclusion, strengths, limitations and recommendations of the study. Conclusions are presented regarding each of the two research objectives. Readers are informed of the strengths and limitations of the research, as well as how some initial drawbacks proved eventual strengths of the study. Recommendations are offered for future research on test revision for other researchers interested in following a similar methodology, for those involved in test revision, and for the *Griffiths III*.

Concluding Remarks

This chapter has briefly orientated the reader to the topic of this study. The commonly found professional occurrence of practice guidelines were explored in terms of their importance for psychological tests as well as the process of revising and using revised tests. The lack of guidelines in psychological testing, and test revision specifically, was highlighted as the main component of the problem formulation for this study. Terms in psychological testing can be confusing, but the definitions applied to specific terms used in this study were clarified. The aim and objectives of the research and the outline of the chapters in this thesis were also presented. In the next chapter the nature of psychological tests and psychological testing is explored in detail.

Chapter 2: Psychological Testing

The present study focusses on the revision of existing psychological tests. In order to conceptualise test revision within the broader context of psychological assessment it is important to understand where tests fit into the discipline of psychology, what constitutes tests, and what the components are that tests consist of. These topics will be discussed in three sections in this chapter. The first section relates to the place of psychological measurement within psychology, and the interplay between different approaches to measurement. This is followed by a discussion about the role of psychological tests as an expression of the quantitative scientific approach in psychology. The benefits, as well as the criticisms, of the use of tests are discussed. Finally, the common components of a test, together with their complexities, are briefly described, to highlight how failings in components can become cues that a test revision is required. The purpose of this chapter is to highlight key aspects of psychological testing that will be further explored in Chapters Three and Four in terms of how such aspects would feature in guidelines in the subdiscipline of psychological testing, with a specific focus on test revision.

A number of terms including tool, measurement, scale, questionnaire, inventory, instrument, and developmental measure have been used interchangeably when referring to psychological tests (Swanepoel & Krüger, 2011). According to Coaley (2009), these diverse terms have created both linguistic and conceptual confusion in this subdiscipline. The misunderstanding can deepen when further terms are added. Developmental measures, for instance, are tests that reflect knowledge about constructs within the psychology subdiscipline of developmental psychology. All products representative of the above terms are designed by humans to serve the function of psychological testing, and it is argued that as such, all would require revision at some point (Foxcroft & Roodt, 2013; Swanepoel & Krüger, 2011; Wiberg &

von Davier, 2017). The abovementioned terms are therefore treated interchangeably within this study dependent on the literature reviewed, although the more encompassing and generic term of 'test' will be used most frequently. The next section describes measurement as an objective of psychology, and how psychological tests fit into the discipline of psychology.

Measurement within the Discipline of Psychology

Psychology, through its subdiscipline of psychological testing, provides a variety of options to those who seek assistance from the psychological profession in making valid and objective decisions in diagnostic, counselling, or personal development (Strauss, Spreen, & Hunter, 2000). According to Carretero-Dios and Perez (2007), psychologists work with phenomena that cannot always be directly observed but that can be measured in order to facilitate the delivery of psychological services (Fried & Flake, 2018). This measurement process is referred to as psychological assessment, a prevalent component in the subdiscipline of applied psychology. As assessment theory evolves, assessment methods will need to be revised to remain true to its theoretical base (Butcher, 2009). Psychological assessment can take the form of two main approaches that are focussed on either subjective or objective observation of psychological phenomena (Australian Psychological Society, 2018; Coaley, 2009).

Subjective measurement in psychological assessment.

Subjective measurement relies on personal scrutiny of an individual from different levels (Moerdyk, 2015). The first level would be self-directed observation by the individual as documented through self-report. Associated with this level is observation from those who know the individual, such as family, friends, co-workers, and teachers. The primary concern with observations at this level would be the inter-personal history between the observer and the observed that may lead to distortions in observations and reporting (Fried & Flake, 2018;

Gregory, 2018). The second level of observation would be from someone who does not know the individual, such as a member of his or her peer group. The main concern at this level of observation would be that the lack of familiarity of the observer with the observed might lead to misinterpretation of what they observe, or that such observers may lack the professional knowledge to identify important behavioural cues (National Academy of Sciences, 2009; Schwarz, 1999). The third level would be observation by a trained psychological professional. Whilst professional observation is foundational to the diagnosis and treatment of psychological dysfunction, the accuracy of such observation is governed by the depth and currency of knowledge and the level of experience of the psychological professional in terms of the psychological construct being observed. Professional observation is also not available at all times and, as it relies on a sample of observations taken over a limited period of time, the reported level of occurrence of a behaviour or psychological trait being investigated often fails to represent the true level displayed by a client (Kaplan & Saccuzzo, 2013).

Whilst observation at all three levels can be important in creating a holistic picture of the individual in question, all these levels are subject to critique regarding personal distance between the observer and the observed (whether too near / personal or too far removed / impersonal), and the lack of knowledge of the nuances that comprise the observed's behavioural patterns, selective observation, and depth of relevant psychological knowledge of the observer (De Vos, Strydom, Fouche & Delport, 2014; Maul, 2017). According to Holman, Head, Lanfear, and Jennions (2015), scientific progress relies on verifiable and reliable data. Subjective observation is particularly susceptible to cognitive and sensory biases that can skew perception, even when observers are skilled (Fried & Flake, 2018). Observers may enter a situation with an expectation

of finding a specific result that will affect their interaction with the observed as well as the stimuli they observe.

To ameliorate this occurrence, it is sometimes advisable to enter an observation situation 'blind', meaning without any knowledge about the client being observed. Holman, Head, Lanfear, and Jennions (2015) reviewed 960 empirical studies and found those that did not use 'blind observation' reported a notably higher frequency of statistically significant results than studies that utilised 'blind observation'. This led the authors to support the practice of 'blind observation'. Whilst this may be feasible in some clinical settings, it would not be possible in continuing therapeutic situations during which a psychologist becomes familiar with a client. Thus, this would problematise subjective observation as the sole foundation for rendering a prolonged psychological service (Maul, 2017).

Another avenue suggested by authors (for example, Burghardt et al., 2012; Neuman, 2011) is that of triangulation of observation. By exploring layers of observation and the propensity for bias in subjective observation, social science research has attempted to strengthen the practice of observation by allowing for concurrent independent observation of a subject matter by multiple observers to create a holistic picture through learning by observation. Through a process of integration of evidence and verification of results, this methodology of triangulation has improved the accuracy of observation studies (Burghardt et al., 2012). Despite this advancement, utilising multiple observers raises concerns about inter-rater reliability. For instance, each observer enters the observation situation with a unique set of skills and points of view. Observer reports from a single observation setting have been known to differ widely thus creating a quandary as to which report to treat as accurate, and how to merge conflicting opinions or mediate between them (Kaplan & Saccuzzo, 2013; Maul, 2017). Kassin, Dror and Kukucka

(2013) further raise the issue of contextual bias where a panel of independent observers may still reach erroneous conclusions due to the context within which the observations occur. By inviting observers to conduct observations in clinical settings, for instance, the setting itself may lead to a clinical diagnosis being reached. The National Academy of Sciences (NAS, 2009) has highlighted the "large body of research on the evaluation of observer performance in diagnostic medicine and from the findings of cognitive psychology on the potential for bias and error in human observers" (p. 8). Bias in observations are not necessarily intentional, but this being said, it remains a consistent criticism and an undesirable aspect in some forms of measurement.

Additional concerns about subjective measurement are that observations are usually not quantified or quantifiable, thereby hampering the effectiveness of longitudinal monitoring of clients. As this type of measurement is largely dependent on factors internal to the clinician on the day of observation, as well as a change over time from one clinician to another, an absence of quantifiable results can affect the adequate supervision and care of patients (Laher & Cockcroft, 2013). Added to this is the preference of health agencies for the quantifiable data that assisted a psychologist in reaching a clinical diagnosis for a patient, a diagnosis, which in turn can influence the specialised support, disability grants, or health service benefits available to such patients (American Psychological Association, n.d; International Test Commission, 2015).

The National Academy of Sciences (2009) has encouraged disciplines to develop protocols to improve the reliability of measurement, especially subjective observations. According to Kassin, Dror and Kukucka (2013), independent verification of observations is a critical step in promoting the quality of service rendered. Objective measurement is one method of achieving such an impartial perspective and this is explored in the next section.

Objective measurement in psychological assessment.

Objective measurement provides a counterbalance to subjective approaches in several ways. Moerdyk (2015) has identified five properties of reliable objective measuring techniques. These are:

- to attach an observable phenomenon to that which cannot be directly observed
- a correspondence between observable and unobservable phenomena
- that the observable phenomenon should be measurable on some scale
- a consistent and reliable relationship between the observable and unobservable phenomena, and
- measurement systems that are transparent and consistently applied, thus allowing different observers to agree on the value assigned to the phenomenon.

The above properties are important foundations of objective measurement, which seeks to apply consistent measuring practices that are intended to be free of observer bias, which is of such concern in subjective measurement. Objective measurement adopts a narrower definition of measurement by attaching a value to the observed phenomenon to quantify it.

The history of quantified reporting within psychology, as well as its connection to psychological testing, can be traced to the latter half of the 19th century. Although the notion of psychology as a science had already been put forward by Christian von Wulf in 1732, exploring psychological constructs through quantified methods only gained traction with the work of Sir Francis Galton on heredity and genius in 1869 (Cohen & Swerdlik, 2018). Galton's work generated interest in intelligence as a psychological construct in the late 1800s. At that time researchers such as Wilhelm Max Wundt were exploring the objective and quantified measurement of behavioural traits within the newly formed subdiscipline of experimental

psychology (Wango, 2017). Wundt was instrumental in training James Cattell and Emil Kraeplin, who were early pioneers in developing intelligence tests. The convergence of quantification, experimental psychology, and human intelligence laid the foundation for psychological testing as an important subdiscipline of psychology (Wango, 2017).

In their seminal work on quantification in psychological testing Nunnally and Bernstein (1993) identified six advantages of quantification, namely, objectivity, precision, comparison, generalisability, communication, and economy. Objectivity relates to assuming a neutral position and removing observer bias. By promoting objectivity, measurement strengthens the connection with the scientific approach by employing known and agreed on rules that produce knowledge instead of speculation. According to Nunnally and Bernstein, objectivity in measurement has been a driving force in breakthrough knowledge creation within the discipline of psychology.

Precision refers to finer distinctions that can be made in observation that enhance accountability with regard to the nuances of observed behaviour. This allows for the flagging of deviation from what is considered average or normal, as well as pinpointing the level of deviation from normal in quantifiable terms (Cohen & Swerdlik, 2018).

Quantification allows for comparative analysis. With repetitive measurement this allows for comparison over time to determine behavioural trends (Nunnally & Bernstein, 1993). This is important in the management of disease. The data display options that are made available by quantification, such as graphs, can convey a wealth of information regarding the progression of a disease over a patient's lifetime.

Measurement creates the possibility of generalisability. Psychology as a discipline is interested in understanding and explaining individual human behaviour. Psychological theories are critical to this process of understanding, but theories also rely on interconnected data to

explain how aspects of human functioning fit together (Moerdyk, 2015). Quantification facilitates the process of generalising from the specific to the general, thereby allowing for classification of individual points of observation within the larger structure of behaviour. This classification then allows for a greater level of distinction between how an individual's behaviour reflects or differs from others that in turn leads to the more accurate diagnosis and tailored treatment of patients (Nunnally & Bernstein, 1993; Proyer & Häusler, 2007).

Quantified information is easier to communicate and interpret. The magnitude of stating that a client is more depressed in this week's appointment than last week can only be clearly conveyed when the observer's definition of 'more' is qualified. Stating however that a client's score on a known test of depression increased from 10 to 15 in the last week will foster understanding of the extent that the depression has increased (Gregory, 2018).

The final benefit described by Nunnally and Bernstein (1993) is that of 'economy'. In brief this means that numbers can be attached to an expanded verbal definition. In the above example, psychologists trained on a depression test would understand the meaning of test scores and they would be able to conceptualise the level of depression indicated by a score of 15 compared to that of 10. This benefit economises communication between professionals and replaces lengthy descriptions with compact information.

Objective measurement employs therefore a scientific view of assessment, with pre-determined and agreed on methodological approaches. At best it seeks to eliminate observer bias by removing individual interpretation of observed phenomena from the equation (Kaplan & Saccuzzo, 2013). It relates and communicates observations in terms of quantified terms. At heart, objective measurement is reliant on psychological theory as well as current research for

developing measurement instruments, scoring criteria, and verbal interpretation of quantified information.

The relationship between subjective and objective measurement.

From the previous two sections it may appear that objective measurement is preferable to subjective measurement. This would be a restricted view however. As stated earlier, psychology deals with the observation of what often cannot be directly observed (Fried & Flake, 2018). The process of relating unobservable psychological constructs to observable criteria can be challenging. Any mistakes in this process would introduce error in measurement. This in turn creates a cascading effect, including negative impact on the assumptions that can be made about unobservable constructs, thereby increasing the chances for misdiagnosis, with resultant negative outcomes for the client (Fröhner et al., 2017; Rhodes & Madaus, 2003). All this has implications for the need for test revision. As assessment methods are revised to reflect changes in assessment theory, practitioners are able to modify their use of subjective observation techniques faster than externally developed objective standardised tests. This is because subjective assessments can be revised more quickly than objective assessments.

In addition, psychological training is aimed at developing the observational skills of psychologists to identify the different nuances in human behaviour, and to integrate what is observed into a coherent overall picture. From this broader picture a psychologist can then reach plausible findings for a particular client. A final diagnosis is therefore never reached by a psychological test or measurement scale but by a professional. The role of the psychologist would be to review all available information and to form a diagnosis if the balance of probability is weighted in favour of a particular diagnosis (Altmann & Roth, 2018). At best, a psychological test can strengthen the case for a particular diagnosis, but due to the inference it makes in linking

the observed behaviour with the underlying psychological construct, it cannot form the sole base of support for a definitive clinical diagnosis (Paterson & Uys, 2005). With this in mind, Foxcroft and Roodt (2013) refer to a synergistic relationship between objective and subjective measurement. They refer to a balance that should be maintained when collecting information that employs multiple measures, domains, sources, settings, and occasions.

It is in harnessing the strengths of different types of measurement that a rounded picture can be developed, which should form the cornerstone of any psychological service (Altmann & Roth, 2018). The concept of triangulation, or viewing a subject from different perspectives, is well known in social research and professional practice. The benefit of triangulation is that it can be used to verify or validate the information that is obtained from the client. According to Bhattacherjee (2012), failure to validate findings through a triangulated approach is one of the main failings of some research studies, thereby leading to biased conclusions. By viewing each testing encounter within a research paradigm, the value of the abovementioned caution is evident within the process of psychological measurement.

Neuman (2011) refers to the four main types of triangulation as measure, observer, theory, and method. Psychological tests fall within the scope of triangulation of measure, with the need for multiple measurements being of primary concern. Multiple measures increase the chances of creating a rounded picture of a client, with differences in observation raising questions or issues that a psychologist would need to explore further with the client to promote the accuracy of the overall picture of the client (Neuman, 2011). This being said, triangulation of measures is still dependent on the accuracy of the individual instruments, which places the burden of reliability and validity onto the tests themselves. This can be particularly problematic when using tests on population groups that the tests have not been standardised for or normed on. In such instances, a

test can only be used if evidence exists of its validity and reliability for the other population group. Again, there are implications for the need for test revision here. This may require standardisation and/or adaptation of the test or even revision of the test to minimise cross-cultural bias, and guidelines are needed to address how this should be done. The role of psychological tests within measurement will be discussed next.

What are Psychological Tests?

There appears consensus in the literature on the definition of psychological tests. According to Kaplan and Saccuzzo (2013), a test is "a measurement device or technique used to quantify behaviour or aid in the understanding and prediction of behaviour" (p. 6). Urbina (2014) describes a psychological test as "a systematic procedure for obtaining samples of behavior, relevant to cognitive, affective, or interpersonal functioning, and for scoring and evaluating those samples according to standards" (p. 1). Gregory (2015) defines psychological tests as "standardized procedure for sampling behavior and describing it with categories or scores. In addition, most tests have norms or standards by which the results can be used to predict other, more important, behaviors" (p. 512).

Psychological testing is a branch of psychological assessment aimed at the development and administration of psychological tests. In the previous section subjective and objective measurement were discussed. Psychological tests can combine subjective and objective measurement, but there is a greater reliance on objectivity in reporting through the use of quantified data. Tests are observational instruments that are focused on measuring a limited number of psychological traits according to established theories about such traits, including operational definitions of behaviours associated with the traits in question (Foxcroft & Roodt,

2013). In terms of empirical evidence, psychological tests undergo intensive research to determine the accuracy of their measurements.

According to Moerdyk (2015), tests have to meet three criteria of proof to establish their status as valid instruments. A deficiency in any of these criteria is a cue that a test should be revised. The first criterion is that test results must remain constant. This implies that should a test be taken on different days, the test scores should remain the same, assuming that there has been no change in the trait measured for the test taker. The second criterion is that a test should measure what it claims to measure. A test should be a direct reflection of psychological theory, and it needs to measure the scope of a trait sufficiently to allow test results to be reflective of the level of performance on the trait in general (Dunbar-Krige et al., 2015).

The third criterion is that of fairness. This implies that a test would measure a trait without the interference of aspects not related to the trait (Dunbar-Krige et al., 2015; Paterson & Uys, 2005). Aspects of particular concern would be gender and ethnic background. In the event that a black female has the exact level of reading proficiency as an individual white male, their scores on a reliable and valid reading comprehension test should be identical. The test should not be a measure of their background, but it should be singularly reflective of the trait in question (Urbina, 2014). In reality, this can be the most difficult aspect of fairness for a test to meet. The reason for this is that psychological constructs are connected, making it difficult to isolate an individual trait. In the above example, reading comprehension may be affected by differences in language use in different cultures and this would not only apply to the background of the test taker but also that of the test developer. Fairness in psychological testing has long been a subject of debate, specifically in multicultural countries, such as South Africa, where intelligence testing was used by the Apartheid government to support the discriminatory social and political agenda

of that era. Some constructs like intelligence prove illusive to delineate as they are informed by culture and context. As test items would reflect gaps in theory, the result would be an element of unfairness in the test results. This is where the experience of test users become important to interpret test scores. As tests strive for standardised and objective assessment, test bias would be of concern and this serves as an important motivation for a test to be revised (Foxcroft & Roodt, 2013; Heuchert et al., 2000; Laher & Cockcroft, 2013).

Clients sometimes confuse psychological tests with pseudo-scientific measures that appear in popular media. Foxcroft, Paterson, le Roux, and Herbst (2004) also suggest similar confusion even amongst psychological test practitioners about the distinct identity of psychological tests, and how they differ from those published in magazines. The three criteria mentioned above should be the cornerstones of any debate about the identity of a test. What sets psychological tests apart from other popularised tests is the care that is taken in their development process, and the body of research evidence and expert opinion that forms part of the defence regarding the integrity of the measure.

As psychological tests are products created by individuals or groups within the discipline of psychology, questions are naturally raised regarding their precision, especially given the issues discussed earlier concerning the accuracy of different approaches to measurement. This is considered in the following section.

Psychological tests as measurement instruments.

As stated previously subjective measurements are subject to errors in observation and reporting. Despite the scientific nature of objective measurements errors can also occur, thereby affecting the accuracy of observations (Laher & Cockcroft, 2013). The subdiscipline of psychological testing accommodates this possibility of measurement error in the formula:

Observed Score = True Score + Error Score

In this formula, observed score refers to the test score obtained by a client on a psychological test. This score is comprised of the client's true score, or actual ability on the measured trait, and any errors. The errors in question are usually ascribed to aspects related to the process of testing, or problems inherent in the test, its marking, and interpretation, which may produce a result that differs from the client's true ability (Moerdyk, 2015). Test developers conduct research to determine the general level of error in measurement either in order to account for it or to reduce it in the final interpretation of an individual's test scores (British Psychological Society, 2017).

A common method of reducing error score is to convert a test-taker's observed score to a standardised score. This process is founded on the principle within psychology that traits and behaviours are normally distributed within a population, with most individuals possessing an average level of a trait, and an ever-decreasing number of people being further away from the population mean. Using this theory, psychological tests are standardised as part of the development process. Through this process the test is administered to a representative sample of the population and, should the test performance of this sample be above or below the expected mean of a population, a correction is built into the scores to realign test scores with the trait-mean expected for the population (Foxcroft & Roodt, 2013).

Accounting for error is standard best practice in psychological tests and is aimed at transparency towards measurement error whilst producing test scores that reflect true scores as accurately as possible. The amount of error is also presented in a quantifiable way that allows test users to account for it when they work with test scores and normative information (British Psychological Society, 2017. A large error score would be inconsistent with the theoretical foundation of objective assessment, and would thus be a strong cue that a test must be revised to

reduce the extent of measurement error (Butcher, 2009). The relationship between psychological testing and science is discussed next.

Psychological testing as science.

Mention has been made of the scientific rigour employed in psychological testing, but it should be acknowledged that psychology as a soft science has traditionally been viewed by the hard sciences as a non-science (Neuman, 2011). This is partially attributed to the fact that humans as subject matter are more variable and difficult to explain, predict, and control than the focus areas of the hard sciences (Altmann & Roth, 2018). Whilst psychology concedes that its subject matter is complex, the discipline has increasingly relied on the principles of scientific research to promote clarity about constructs related to the human psyche. Science relies on objective, valid, and reliable measurement, and psychology has applied these principles particularly in the subdiscipline of psychological testing (Foxcroft & Roodt, 2013).

In the 20th century psychological testing was a consistently developing subdiscipline as it combined the subject matter of psychology with a scientific research approach to promote the objectivity of clinical observations (Foxcroft & Roodt, 2013). The psychological tests that were created as a result have furthered, in turn, the cause of scientific enquiry as many research projects rely on such tests to facilitate the measurement and exploration of psychological constructs (Carretero-Dios & Perez, 2007). In the 21st century advancements in measurement theory, statistics, and information technology continue to advance psychological tests to new heights.

The Benefits of Psychological Tests

As can be gleaned from the above, the task of psychological testing is to assume a critical perspective on its own practice whilst endeavouring to produce tests that accurately reflect

psychological theory and deliver scores and interpretations that are consistent with test-takers' trait-ability, without interference from non-essential external factors (International Test Commission, 2013). International test publishing firms offer practitioners a selection of tests that cover a wide scope of human functioning to assist in creating a more holistic and unbiased picture of clients. There are a number of benefits associated with psychological tests that have supported the popularity of the psychological test industry.

The first benefit relates to standardisation of measurement. With the increased emphasis on fairness in the rendering of psychological services and the rights of the client, standardisation has emerged as an important standard for rendering a consistent, customer-oriented service (Eyde, Robertson & Krug, 2010). The process of psychological testing, from administration, item content, scoring, interpretation and reporting of results is standardised for all test takers of a specific test. According to Murphy and Davidshofer (2004), a test is a "sample of behaviour collected under standardised conditions" (p. 4). Although this is not disputed, it can be argued that in order to facilitate a more accurate measurement, a representative sample of behaviour of the measured trait should be obtained in order to yield conclusive evidence. This would be the aim of a scientifically developed psychological test (Kaplan & Saccuzzo, 2013). The conditions under which an observation is conducted further affects the depth, quality, and accuracy of data that is collected and by controlling each step of the measurement process, psychological testing ensures that the assessment experience is the same for each client, regardless of time, context or environment. This, in turn, lends credibility to the process and enhances the confidence that can be placed in the observations, and the ensuing outcomes of the test experience. It should be noted that tests are static and reflect the understanding of the underpinning construct, acceptable test

content, and expected test behaviour of test takers at the time the test is developed. A change in any of these aspects would necessitate test revision.

The second benefit is that of speed and ease of some testing procedures, such as computer-based testing. The process of forming a comprehensive picture of a client can be time-consuming and this can be at variance with the fast pace of modern living. Consumers have become accustomed to immediate feedback and results, and psychological testing can speed up this process by measuring key behaviours and constructs associated with human development and functioning. Tests are designed to be user-friendly and to put test takers at ease whilst assisting in providing clarity on psychological constructs. The digital age has allowed for computer-based tests. Computer-based testing has a number of advantages, but foremost for the client is the ability to tailor the test to the visual requirements of the test taker, including increased font size, ability to display each question on its own, immediate feedback from the programme, as well as the sense of comfort and connectedness that many people today feel with computers (ITC, 2005). An additional advantage is that computer-based tests can score and report test results immediately. Test results also lend themselves to reporting in visual formats, such as graphs and pictures that appeals to the visual senses of test users, and brings greater short-term clarity than lengthy verbal explanations. Computer hardware and software can date however, which would require revision of the test software associated with this mode of testing.

The third benefit of tests refers to the expertise of those involved in the development of a test. The level of service rendered to a client depends on the expertise of the psychologist. Psychological tests are generally developed by a team of experts on the construct in question, which results in a product that has been vetted by respected leaders in the construct and within the subdiscipline of psychological testing. The use of such tests can promote the clinical

objectivity of the psychologist, thereby resulting in a diagnosis that is supported by triangulated observed evidence (Fröhner et al., 2017). An important consideration for psychological tests would be the cultural representativeness of test developers, especially for tests that are used for different cultural groups. A development team that lacks cultural diversity may unintentionally introduce cross-cultural bias to test content, which will need to be corrected through subsequent test revision. The issue of cross-cultural diversity in revision teams cannot be understated, and this needs to be stated clearly in guidelines for test revision.

The fourth benefit of testing lies in the standardisation information and research evidence that underlies a test. The norms that are supplied with a test allow test users to compare a client's score to a group of their peers. This enables test users to contextualise scores based on convergence or divergence from a population norm that highlights those areas in which a client performs similarly or dissimilarly to the expected norm. This information can be a reference point for self-understanding in the client and a potential catalyst for personal development (Creswell, Hanson, Plano Clark & Morales, 2007). Despite an abundance of tests Foxcroft, Paterson, le Roux, and Herbst (2004) found that most practitioners have similar needs, which are to accurately and consistently measure psychological traits according to the most recent academic understanding of the constructs being investigated. The latter authors' research also underscored that tests are used by a sophisticated and discerning market that demands high quality products, with a specific emphasis on fairness in multicultural environments. The psychological testing subdiscipline of psychology strives to meet such high standards. Its singular focus is on developing and promoting the accurate, fair, objective and ethical measurement of psychological constructs within the individual. This ethos has endeared

psychological tests with many in the psychological profession, and it has resulted in the creation of a multibillion-dollar global test industry (Eyde, Robertson & Krug, 2010).

However, it would be erroneous to assume that tests are universally praised and accepted by those within as well as outside the profession of psychology. The more pertinent criticisms of psychological tests are discussed in the following section.

Criticisms of Psychological Tests

In some respects tests have fallen victim to their own success. The growth of the testing industry has not gone unchallenged, with detractors highlighting the shortcomings of tests. The main criticisms against tests relate to reductionism, oversimplification of how complex human beings are, their questionable value for diverse populations, and the costs associated with tests (Ferreira, 2016; Murphy & Davidshofer, 2004). Some of the specific arguments around these critiques are explained below.

Reductionism: Human behaviour is complex, with an endless combination of traits and behaviours adding to the uniqueness of each individual. No psychological test can measure all the possible nuances of human behaviour, which means that tests either measure those aspects that are most commonly found, or those behaviours that are most distinguishable. Such observed traits are further limited in scope to what can be expressed as a number. This narrowing of scope to fit a potentially limitless and unobservable psychological construct into a short psychological test can result in gaps in the depth and breadth of an assessment. Tests can only record a sample of behaviour or responses, and this leaves scope for behaviours or responses that are not observed during an assessment situation (Murphy & Davidshofer, 2004).

The Hawthorne effect is a concept in social science research that states that during experiments participants behave differently to how they normally would as a reaction to the

research environment (De Vos, Strydom, Fouche & Delport, 2014). This reaction refers to the realisation that they are being observed, as well as the artificiality of research experiments. By attempting to control the assessment setting psychological testing creates a similar environmental change that, by virtue of its artificiality and known observational purpose, can elicit changes in behaviour on the part of the client (Paterson & Uys, 2005). For more attuned clients this could in turn lead to 'demand factor', another threat to the external validity of experimental research (De Vos, Strydom, Fouche & Delport, 2014). During demand factor a client may become aware of the type of behaviour that will be a focal point for the psychologist, and modulate their behaviour accordingly. In fairness, responses similar to the Hawthorne effect or demand factor can be provoked by many other forms of observational procedures, but the performance element associated with tests could amplify this phenomenon (Murphy & Davidshofer, 2004). This places an additional burden on tests to overcome this concern, and to comment accurately on the samples of behaviour collected. Tests can therefore, at best, only form part of the observational tools employed by a psychologist.

Oversimplification: The efficiency with which tests can be conducted and reports generated electronically can sometimes lull practitioners into a false sense of security about the validity, reliability, and suitability of the tests they utilise (International Test Commission, 2013). The number of psychological tests that are commercially available has created a consumer-oriented industry. Many test-users may emphasise ease of administration, scoring, and reporting, with less focus on the actual quality of content, underpinning constructs, and research evidence that support a test. Such oversimplification can create opportunities for unscrupulous developers to produce tests that appear easier to use, but that are of lower quality. The advent of computer-based or internet-delivered tests has also raised issues regarding test security, and the possibility

for confidential and sensitive information of test takers to be accessed by hackers or as a result of computer glitches (ITC, 2005).

Diverse Populations: Many tests that were originally developed decades ago to perform a specific task for a specific population group continue to find an international market, sometimes with little research on the validity of the test in postmodern times or for different contexts (Foxcroft & Roodt, 2013). According to Laher and Cockcroft (2013), tests have faced considerable criticism for possessing little value for diverse populations. Tests have been described as having been developed by white middle-class males for white middle-class males. This is evident in many performance tests, such as the Scholastic Aptitude Tests (SATs), where historically white test takers have obtained better scores than their peers from other ethnic groups (McDonald, Newton, Whetton, & Benefield, 2001). Tests are usually taken for a specific purpose, and the likelihood for test takers from certain cultural groups to obtain lower scores on a test could mean that they would be less likely to have a positive recommendation based on test scores than a test taker from an ethnic group for whom this anomaly is not a general concern (Ferreira, 2016). This concern for reliability is however not unique to psychological tests, but is a concern in different fields of measurement, including the biological sciences (Fröhner et al., 2017). If a test is used in population groups that is what not initially standardised for, research should be conducted to provide the required standardisation information. If the use of the test continues to spread to other population groups, this broader population would need to be accounted for in the standardisation information of future test revisions. Guidelines on how to do this would be important for revision teams (Butcher, 2009; Foxcroft & Roodt, 2013).

Despite this caution, testing still produces more reliable and valid results than any other measurement alternatives available for psychological constructs (Laher & Cockcroft, 2013).

Many of the alternatives to psychological testing are often found wanting in scientific rigour. A well-developed test provides feedback on a sample of key behaviours, whereas other data collection methods such as interviews or a sample of a client's biographical writing can be interpreted in divergent ways by different experts due to the absence of the standardisation of scoring and interpretation that is a hallmark of a psychological test (Assessment Oversight and the Personnel Psychology Centre, 2009). The Assessment Oversight and the Personnel Psychology Centre (AOPPC) of the Canadian Public Service Commission identifies three risks of interview methods that can affect the quality of the decision-making process (which includes selection and diagnosis). The first is bias and inequity where information supplied by the client is misinterpreted by the interviewer as a result of conscious or unconscious bias, particularly in relation to race and gender (AOPPC, 2009). The second risk is associated with the inaccuracy and lower predictive ability of unstructured methods, including interviews. The AOPPC (2009) points to consistent research findings that interview performance does not accurately predict future behaviour, as interviewees are able to gauge and express the types of responses preferred by interviewers. The third risk is the legal vulnerability of unstructured assessments in court challenges (AOPPC, 2009). Psychological tests, on the other hand, strive to meet these challenges in the development, scoring, and interpretation of results, in order to provide assessment findings that are supported by research evidence (Eyde, Robertson & Krug, 2010).

Test Materials and Training: Camara (2007) expresses reservation about the growing number of test users who do not possess the levels of training or qualification required to use psychological tests in a professional manner, thus assisting clients with life-changing decisions without the necessary theoretical or practical foundation to do so. This could create overconfidence in psychological tests that, by their very nature, are devices that may contain

flawed design elements (Eyde, Robertson & Krug, 2010). A psychological test is a product that consists of a number of designed components that are developed sensitively to work together towards the goal of accurate measurement. The process of producing, marketing, and conducting ongoing research on a test requires resources in terms of time, expertise, and costs. These costs are recovered by the test publisher from consumers. Although the purchase price of a test can be expensive, test users may also find themselves having to pay for training costs, registration fees as users of the test, as well as automated scoring and reporting services offered by the publisher. Added to this, tests are not static, which has additional cost implications for test users. The critical failing that could result from flaws in a test is a level of measurement error in the scores of test takers, which would in turn undermine the goal of the test (BPS, 2017; Rhoades & Madaus, 2003).

In the previous sections the nature of psychological tests as a science was explored, together with the associated benefits and criticisms of this subdiscipline. A psychological test consists of different components that work in synergy to comprise a quality product. These components are created to serve a particular function, that is, to add value to a test. When a component fails to meet its design specifications, it introduces measurement error into a test that would be an important source of criticism of that test, which would need to be addressed through test revision. It is therefore important to explore these components, to understand their function, their benefit, and possible areas of criticism. These components of tests will be addressed in the following section.

Components of Psychological Tests

Psychological tests consist of various components that in total contribute to the overall quality of a test. The two main test components are the test itself and a manual that supports the

use of the test (Groth-Marnat, 2009). The test component is administered to the client, and it includes the actual questions or tasks the client is expected to answer or perform, as well as any physical equipment that the client needs to interact with in order to complete the test. The test manual contains information needed by the psychological practitioner, such as:

- information regarding the purpose of the test
- the constructs it is designed to measure
- the process by which the test was developed
- guidelines on how the test should be administered, including instructions that should be given to the client
- information on how a completed test should be scored and interpreted, and
- statistical information about the test, such as test norms, and research into the validity and reliability of the test for its intended target population (Coaley, 2009).

The focus of the present research relates to the revision of psychological tests. As tests are composed of the above-mentioned components, the following sections of this chapter will elaborate on each of these components, focussing on how each component contributes to the effectiveness of the overall product, and at what stage failings within components could impact the utility of the test.

The Administered Psychological Test

The test administered to the test taker consists of carefully developed questions or tasks that are designed to assess the psychological construct the test was designed to measure. There are different types of psychological tests, including tests of current ability, knowledge base or level of development, potential for future development, personality traits, attitudes, and behavioural patterns (Domino, 2006). Tests can be conducted in a variety of formats, such as

structured questions and engagement with sensory stimuli (such as sounds, pictures, and manual manipulation of a physical object).

Tests can assume a variety of formats. Some tests allow for open-ended answers that are scored afterwards, whilst others use multiple-choice formats that provide a set of possible answers to test takers, where they must choose either the correct answer or the answer that most resembles themselves (Domino, 2006). Most tests rely on a language-based format of delivery. Test instructions and questions can be administered to a test taker either verbally or in writing, to which they need to respond verbally or in writing. For any test that contains linguistic elements, proficiency in understanding and/or reading a language at a specific level is a fundamental skill required for completing the test (Foxcroft & Roodt, 2013). Language itself is fluid and subject to change and a reflection of society. It is used by a population and the subcultures within it to facilitate responsiveness and to introduce new concepts. New words are continuously entering the stream of societal awareness, and the popular meanings of words regularly adapt to suit the culture of the day. This has an effect on the use of language in language-reliant activities, including a large subset of psychological tests. A change in the meaning of a word may necessitate an update to a psychological test, as misinterpretation of the word may result in error in the accuracy of the test to measure a construct in test takers (Koch, 2005).

Knowledge-based tests are also susceptible to advances in our understanding of the world or change with time, which means that answers that may be considered correct at a certain time may cease to be so in the light of new evidence (such as whom the president of a country is, or what the smallest part of an atom is). This will require an update to answers of knowledge-based test questions, for such items to remain accurate, thus calling for a test revision process. Tests that make use of pictures or equipment are equally important within the subdiscipline of

psychological testing. A number of item types lend themselves to a picture-based stimulus. These can include cognitive tasks such as spotting the missing object, matching pictures by type, placing picture cards in a logical sequence, and so forth. Picture-based stimuli can date very quickly and may require updating regularly. This is particularly important when a test is used in new populations. A set of revision guidelines would be important to assist teams in sensitising them to cross-cultural fairness, to minimise cultural bias in revised tests. Non-cognitive or personality-based questions can also be focused on a picture, such as asking the test taker to explain what is happening in a picture, or what the picture reminds them of, as in the *Rorschach Test* (Meyer & Eblin, 2012; Rorschach, 1927).

Another aspect related to how a test is presented relates to the mode of assessment. The traditional method of completing a test has been through paper-based answer sheets. With the advent of computers a world of possibilities emerged, including stand-alone computer-based testing, internet-delivered testing, and measurement aided by the use of technology, such as virtual reality (Coaley, 2009). The shift from paper-based testing to computer-based formats has created opportunities and challenges for those involved in the test industry. Computer-based testing (CBT) outstrips traditional methods in terms of standardised administration, item delivery, scoring, and interpretation of test results, as it removes human error within the assessment process (Piaw, 2012). Piaw concludes that "the CBT mode is more stable and consistent in terms of internal and external validity" (p. 662) than traditional testing methods.

That being said, this mode of testing comes with its own set of challenges. Foremost is the ability and level of comfort of test takers with computers (Coaley, 2009). Even in its most unthreatening form, a psychological test is still an assessment that can raise the anxiety levels of test takers. In the event that a client is also not familiar with computers, this method of

administration can itself become a confounding variable to accurate measurement (Yu & Ohlund, 2010). Some countries, cultures, and age groups enjoy greater access and exposure to computers than others, which is a factor that needs to be considered when developing or using a test. Apart from the interface between humans and technology, the reality is that technology is fallible. Interruptions to the energy supply or internet connection can seriously impact on a test session. Computer hardware and software can also date and software can become infected with a virus that could simply render a computer useless, or in the event of malware or spyware, compromise test security or transmit personal information such as confidential test results to an outside party (Eyde, Robertson & Krug, 2010). Technologies associated with CBT also age quickly, and changes in technology will necessitate a revision to tests, in order to remain usable on new digital platforms.

Piaw (2012) argues for greater innovation in theory and modes of testing, as well as research to support the usefulness intended by such innovation. The caution remains however that the pace of innovation outstrips the current pace of understanding and research about the effect of such progressive assessment techniques on the accuracy of measurement of true scores as well as the value of test results. In light of this, the present researcher argues for more intensive research aimed at investigating new innovations in assessment and a concerted effort to develop measurement theories that can encompass the advance of technology.

Given the above-mentioned concerns, traditional paper-based methods may be more comfortable for some test takers, and more practical for situations where access to technology is a problem. This is a tricky tightrope for users of psychological tests, and it poses one of the most noteworthy challenges facing the test industry. The next section will examine in greater detail the test manual as a component of a psychological test.

The Test Manual

The test manual that accompanies a psychological test is developed to support the accurate use of that test. This means that an update to test manuals would be a standard component of test revision, as the manuals need to reflect changes made to any test components. The aspects mentioned earlier that comprise a test manual will be discussed individually below.

The Purpose of the test.

Each test is developed with a purpose in mind. This purpose serves as the motivation behind the test's development, what the test intended to measure, and what type of test taker it is intended for (Hogan, 2013). These considerations have a direct influence on the final test produced. The main reason why a clear and detailed test purpose is required is to convince the professional community of why this test should not only exist, but also why it should be used in professional practice. The purpose statement in the manual needs to specify what the test is designed to measure, whether it be aptitude, proficiency or personality traits, and if the measurement is designed to measure the construct for a clinical or non-clinical population or both. A test can also be designed to be an initial screening instrument for possible clinical issues, or to be an in-depth measure aimed at facilitating the diagnosis of clinical disorders (Hogan, 2013). Another aspect that flows from the purpose statement is the delineation of the specific population the test is intended for, including age group, nationality or cultural group, language proficiency, and the presence or absence of a clinical diagnosis (Ferreira, 2016).

The purpose of a psychological test is subject to change over time and should be updated to reflect such shifts. In some cases, the definition of the construct it is intended to measure may change, the grounds on which a clinical diagnosis is reached may alter, and subsequent research on the test may increase or decrease the population it can be accurately used with (Bush, 2010).

As this would impact on the purpose of the test, it could significantly affect the components of the test, and even invalidate the test to such an extent that it can no longer be used with confidence in its present form.

The Construct of the test.

In the previous paragraph the construct of the test was referred to. All psychological tests are by definition designed to measure some psychological construct. The psychological constructs being measured cannot be seen and measured in traditional terms. Thus the measurement conducted by the test is based on an operational understanding or definition of what the construct entails (Coaley, 2009; Maul, 2017). Each construct consists of subconstructs or components that contribute to its full understanding. Intellectual ability consists, for instance, of various aspects and behaviours that are considered markers of intelligence by the psychological profession. As an active discipline, research into psychology is constantly challenging and updating the understanding of its discipline that results in changes to how constructs are defined as well as what elements form part of a greater construct.

The construct definition serves as an anchor for the individual questions that are included in a test, as the questions must contribute to measuring performance on the construct (Maul, 2017). A change in how the construct is defined and conceptualised would in turn affect the suitability of some items and potentially necessitate test revision that includes changes to existing items, or alternatively removing items and including new questions, in order for the test to satisfy construct validity demands (Fried & Flake, 2018). Different tests become dated at different speeds, which means that revision teams should consider this aging process for their tests and revise tests within specific timeframes. According to Butcher (2000), tests of cognitive performance and achievement age faster than other tests and should be updated more regularly,

as these constructs are greatly influenced by the current social climate and therefore draw more from up-to-date social stimuli (such as pictures or tasks), whereas the stimuli of personality tests remain more constant over time. Given the different rates at which tests age, it can become difficult for revision teams to discern the optimal time to embark on a revision. Guidelines would assist revision teams by highlighting the important cues they need to be aware of.

The test development process.

The process by which the test was developed is addressed within its manual to clarify decisions that were made in the development process (Foxcroft & Roodt, 2013). This process allows for scrutiny of the rigour of the test construction process, which in turn allows for professional feedback and critique as an important step in validating the test. The development process often relies on common best practice at the time of the test's development. This permits however additional critique of the test in subsequent years if test development processes and practices have changed, thereby supporting the need for standardisation of test revision guidelines.

Test administration.

The guidelines for test administration and test instructions are important aspects of a manual as they create a framework for each test session to occur in a standardised and comparable manner. If instructions are given verbally to test takers, these instructions could become outdated over time and require updating to be clearer for test takers of subsequent eras. In the event that a test is administered electronically, changes to computer technology and operating systems may also require updated instructions on test administration.

Scoring Instructions and Interpretation of Test Results

As a psychological test is a construct-based measurement, instructions should be included on what constitutes a correct answer and/or how answers should be scored. Information on how an overall score is derived should also be provided (BPS, 2017). A test result in itself is meaningless, and for this reason a manual has to assist test users on how to interpret test scores. In criterion-referenced tests, this interpretation can be based on the extent that a test-taker's score reflects the overall construct, and for norm-based tests it would reflect a comparative interpretation of the test-taker's score with the reference group for the test.

Standardisation Information, Test Norms, and Evidence of Validity and Reliability

The final aspect of the test manual includes the statistical information from research on the test. This would include information on the primary sample used to research the test to determine its statistical reliability and validity (Coaley, 2009). Any psychological test is only as good as its validity and reliability, and a major form of evidence in this regard would be of a statistical nature. Even a criterion-referenced test that is more reliant on qualitative expert opinion needs to be supported by quantitative research evidence to endorse the quality of the test (BPS, 2017). A test manual usually concludes with information on test norms that allows the conversion of an individual test-taker's raw score into a score that compares performance with a larger reference group (Hogan, 2013). Increasingly, manuals also include research on additional research populations, including clinical samples.

The reason for testing further samples is to support the use of the test in populations other than those used to develop the standard norms. Some psychological tests are also used to assist in diagnosing psychological disorders and for this reason research needs to be included in the manual on how sensitive the test is in flagging potential psychological disorders (Proyer & Häusler, 2007). Most research on a test occurs however after it has been launched and marketed

to professionals (Ferreira, 2016). This research can identify the utility of the test for special groups not included in the development of the test. Alternatively, research may also identify flaws in the test and refute its usefulness either for the broader market or for certain subgroups (Paterson & Uys, 2005). This information contributes to the critical debate surrounding a test and can affect the lifecycle of a test.

Statistical information forms an important part of a test manual. This is the section that objectively interrogates the expert opinions relied on by the test developer in constructing the test. The statistical information seeks to provide the scientific evidence of the soundness of the test, known as reliability and validity (Coaley, 2009). Arguably the part of the manual that is most continuously used is the norms. This is the statistical information that transforms a test-taker's raw test score into an interpretable standardised score to facilitate comparison with a broader external reference group. Norms are therefore a key component of a test, as they place the test-taker's performance within the context of the broader population.

An important development in psychological testing in the 20th century was the observation that test norms were subject to degradation over time. This is known as the Flynn effect, named after Professor Jim Flynn who highlighted the concept in the 1980s (Flynn, 1987). The Flynn effect refers to the continuous rise in observed average scores (especially intelligence tests) from the 1930s onwards (Aylward, 2009; Aylward & Aylward, 2011). The effect of this gradual rise is that test norms lose validity over time and become significantly inaccurate after 10 to 15 years (Silverstein & Nelson, 2000), thereby necessitating research aimed at updating norms at regular intervals. Evidence from the early 21st century suggested that the Flynn effect had plateaued in highly developed countries, but was still evident in developing countries that would include South Africa (Daley et al., 2003; Teasdale & Owen, 2005).

Concluding Remarks

This chapter considered measurement as an objective of psychology, and the complexities associated with accurately observing and measuring the behaviour of clients for the purpose of predicting behaviour and rendering appropriate psychological services. Despite the criticisms described about psychological tests, its approach, based on a scientific stance towards measurement, offers benefits in terms of accurate measurement and reporting. Tests developed on the foundations of psychological testing have resulted in a flourishing industry with a shared commitment to objective, fair, and accurate tests. These tests consist however of components that are subject to the effects of time and change in society. As such, they need to undergo revision at regular intervals. The importance of guidelines on how to revise these test components were highlighted throughout this chapter. Given the need for test revision, this aspect of psychological testing is addressed in the next chapter.

Chapter 3: The Revision of Psychological Tests

There is room for improvement in any test, with continuous internal and external factors necessitating a re-evaluation process (Brannigan & Decker, 2006). An obligation rests on test developers, therefore, to develop and update psychological tests according to strict ethical guidelines to ensure their test's validity and accuracy (Foxcroft, 2011). Test revisions may have been sporadic in earlier decades but have become increasingly more frequent (Adams, 2000). Test publishers in most first world countries have met this obligation for revised tests by providing periodic revisions of their tests, despite a lack of formal guidelines on when and how to undertake this process. In contrast, the resources required to revise a test have meant that test developers in developing countries such as South Africa have struggled to meet the demand for revised tests. This chapter will discuss what test revision is, why tests are revised, what the revision process entails, the issues related to test revision, and how the history of psychological testing in South Africa has shaped test revision in this country.

Test Revision Versus Test Adaptation

As broadly stated in the previous chapter, tests are impacted by time and changes in society. Test revision is the process of making changes to an existing psychological test. It is an overarching term for all processes related to effecting changes to any component of a test, such as test questions, equipment, instructions, and test norms (Liu & Dorans, 2013). Each test revision is a unique process that develops according to the goals of the revision process. Nevertheless, Butcher (2000) identifies three types of test revision. The first is a 'light' revision that covers changes mostly to the test manual. Aspects that could fall within this type are minor updates to item wording or editorial changes. The second 'medium' revision type is more intensive and includes changes to or replacing non-performing items, and updating the norms of

a test. The third type is an ‘extensive’ revision that involves a complete reanalysis and reconstruction of the test. This could include re-examining the theoretical foundation of the test and major changes to items or subscales, together with a new set of test instructions. An extensive revision would also include new norm data, as well as validity and reliability studies (Butcher, 2000).

Some aspects of test revision may be aligned, at times, to test adaptation, but the two concepts differ through their intended use. According to the International Test Commission (ITC, 2016), test adaptation is a broader term that encompasses test translation and is the practice of “moving a test from one language and culture to another” (p. 6). The purpose of test adaptation therefore is to create other versions of a test, either in a different language or for a different cultural group, that are as identical as possible to the content and construct definition of the original version (Foxcroft, 2011). A growing need for test adaptation has emerged due to the popularity of some psychological tests in countries and contexts that these tests were not originally designed for. Another motivator for test adaptation is the practice of comparing test scores from test takers of different language backgrounds or cultures, either for comparative research or for employment or educational selection purposes (Bush, 2010). For these uses, it is important that the test accurately measures a construct by not introducing measurement error through cultural or language bias (ITC, 2016).

Test revision is a broader, and often different, practice in psychological testing. In test revision, an existing test is reverse re-engineered by first dissecting and interrogating each component of the test. According to the Educational Testing Service (ETS, 2014), once those aspects that are no longer completely valid are identified, they are revised, together with those aspects such as norms and standardisation information that may be affected by item changes.

This could include noteworthy changes to the underpinning constructs and items of a test, which would result in a revised test that looks different from the original source test. This is the defining difference between revision and adaptation. In adaptation, the primary goal is that all versions of a test should be the same in construct and difficulty level. Similarity of test content becomes of secondary importance, as content may differ across test versions, as long as the link between construct and content remains the same (Ferreira, 2016; ITC, 2016). In a revision process, the test has to be placed under a microscope and compared to the perfect ideal, and then changed to be as close to such perfection as possible, which may change the conceptualisation and operationalisation of the underpinning construct, the difficulty of the test, and the content (Aylward, 2009; Aylward & Aylward, 2011). Whether the revised test looks anything like the original may be of secondary concern, and is dependent on how the revision process unfolds.

If the test developer identifies a need to change any aspect of a test, the test developer should decide whether adaptation or revision would be the most appropriate procedure to meet the project goals. It could be argued that before a test is adapted the issue of whether the test should not be revised first should be answered by the test developer to avoid adapting an outdated or poorly functioning instrument.

The Reasons for Revising Psychological Tests

Psychological tests are to some extent creations of a specific moment and are therefore susceptible to becoming outdated through the effects of interlinked external factors. The *Standards for educational and psychological testing* (2014), a joint publication between the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), states that:

> Test specifications should be amended or revised when new research data, significant changes in the domain represented, or newly recommended conditions of test use may reduce the validity of test score interpretations. Although a test that remains useful need not be withdrawn or revised simply because of the passage of time, test developers and test publishers are responsible for monitoring changing conditions and for amending, revising, or withdrawing the test as indicated. (AERA, 2014, Standard 4.24, p. 93)

From the above quotation the passage of time, changes in knowledge and the context that the test is used for, including the professional sphere or society, can serve as indicators signalling the need for a revision of a test. They similarly determine when and to what extent a test should be revised. Each of these indicators will be explored in the subsections below.

The effects of time on a psychological test.

The concepts, constructs and content of a test, as well as the norm and standardisation information are reflective of the time of the test's creation (Aylward, 2009). This raises the question of how long a test is valid for? According to Silverstein and Nelson (2000), a test and each test revision is "intended to have an influence extending approximately one generation" (p. 298). This view places the effects of time as an overarching theme that affects all aspects of a test, including the decision about when to embark on a revision. With time come advancements or changes in the world and context in which a test is used that may eventually render the usefulness of a test less effective over time. Test questions or items usually draw from stimuli and contexts that are up-to-date (Bush et al., 2018; Foxcroft & Roodt, 2013) to promote test takers' familiarity and level of ease with the test. Questions related to picture cards, for instance, would employ pictures that are contemporary and familiar to test takers. If a picture card contained a computer, for instance, the picture would be current for the majority of test takers.

However, rapid changes, specifically of technologically based stimulus materials, mean that a computer picture card from 20 or 30 years ago would look quaint and unfamiliar to a child in the present times. This simple influence of modernisation or changes in time can influence a test taker's performance on such an item (Adams, 2000). Another impact of time is the Flynn effect alluded to in Chapter Two (Flynn, 1987). The decline in the accuracy of test norms over the span of a decade illustrates the impact of time and changes in society on a test, specifically concerning its ability to be used as a comparative instrument with a test-taker's peer group (Aylward, 2009; Aylward & Aylward, 2011).

Although opinions differ about when a test should be revised, Adams (2000) observes that new versions of psychological tests now appear at about 10-year intervals, a shorter time-span than the one generation postulated by Silverstein and Nelson (2000). This shorter turnaround time, and the period required to prepare the groundwork for a revision, means that any test should be treated as a work in progress, with research and development featuring as a standing item on the test development agenda (Brannigan & Decker, 2006).

The effects of new knowledge on a psychological test.

New knowledge within the field of psychology advances our understanding of psychological constructs. Psychology is a living social science that is enhanced through research globally. Through this research and theoretical conceptualisation, new constructs are introduced, while existing constructs are refined and expanded on. Research is a cornerstone of the scientific approach and plays an important role by providing the empirical evidence to verify knowledge. Research is performed however in basic and applied contexts, with basic research discovering new knowledge and applied research defining the scope or boundaries of knowledge (Neuman, 2011). Within psychological testing applied research may support the utility of an existing test or

highlight aspects of the test that are not aligned to a modern theoretical understanding of a construct. New research on populations for whom a test was not originally normed, or even intended for, may produce evidence for the responsible use of the test on such populations or raise questions about the test questions, test materials, or norms (Foxcroft, 2004).

It could be argued that advancements in the field of psychological testing have created new opportunities for testing, and an opportunity to reimagine the field and the practical application of psychological tests. One major advancement in the latter half of the 20th century was item-response theory. Classical test theory (CTT), which forms the foundation for conceptualising a true score as a fairly linear function of test score and measurement error, was advanced through computer-assisted modelling to facilitate the calculation of a true score as a function of several elements, that is tailored for each test taker (Murphy & Davidshofer, 2004). With the aid of item-response theory (IRT) the level of measurement error can even be calculated at item level, which was inconceivable through standard CTT calculations. As CTT was the benchmark for test development for the greater part of the 20th century, the majority of tests would have employed norm and standardisation information based on CTT (Anastasi & Urbina, 1997). Complex logistic functions based on IRT highlight how outdated the previous conceptualisation of true score was, that in itself can provide a plausible reason to revise a decades-old test (Geisinger, 2013).

A further advancement in psychological testing has been in the field of item development. For decades Blooms Taxonomy (Bloom, Engelhart, Furst, Hill & Krathwohl, 1956), consisting of such competencies as to know, comprehend, apply, analyse, synthesise, and evaluate, formed the basis for constructing items. Later taxonomies have questioned the simplicity of Bloom's model by highlighting the need for complex cognitive processing distinctions and

multidimensional models (Anderson & Krathwohl, 2001). These arguments draw attention to the complexities of human cognitive processing, and the diverse skills and avenues of processing that are required for different types of test items. They also introduce the concept of open scoring, where instead of scoring an item as correct or not, the respondent's answer yields valuable information on how they process information and view the world around them. These advances transform psychological testing from its traditional roots focussed on comparing test takers to a peer group into a unique individualised assessment focus. The emphasis on individuation is a reaction to the comparative focus of bygone eras. It underscores the unique complexity of individuals, thereby increasing the demand for testing focussed on identifying intraindividual strengths and weaknesses for the purpose of tailored development and capacitation, as opposed to inter-individual comparison. Such competency-based testing is more focused on comparing individual performance to the construct being measured, whilst also enabling normative comparisons with reference groups (Cohen & Swerdlik, 2018). This facilitates the balance between intra- and inter-personal elements that Betehenner (2009) argues is key to interpreting growth and progression for the individual.

Arguably, the greatest advancement in recent decades has been the advent of the information age and affordable access to computers. For psychological testing, the digital age has heralded a revolution in terms of development of the field and an ever-expanding plethora of opportunity. Computers have made the complex calculations of IRT possible, thereby advancing the accuracy of test scores. Computers have also introduced the possibility of adaptive testing aimed at accurately assessing the abilities of test takers in a shorter period. As humans increase their level of comfort with computers, electronic devices offer standardised alternatives in test administration that can replace this function of the test user, thereby promoting the validity of

test scores (Chau, 2014). Computers also enable immediate scoring and norming of test results, as well as standardised reports that appeal to the need for immediate satisfaction of modern consumers. Smartphones, tablets, and digital technology take these benefits even further, due to their portability, extended battery life, and sustained connectivity to the internet (Chau, 2014). These advances reverse the traditional model of a client coming to a psychologist's practice to undergo assessment, by taking the assessment to wherever is most comfortable for the client (Geisinger, 2013).

Electronic devices can not only shorten the assessment encounter through adaptive testing, but also conversely extend the amount of time over which an assessment can span. Traditionally time was a major limitation for psychological testing. To prevent over-exposure to test items, most tests may only be taken once or twice a year by a test-taker, to allow test takers to forget test questions, and to minimise the possibility of a test-taker falling into a certain response set on a test. The internet age has changed this paradigm. Through online connectivity, a new version of a test can be created quickly using stored item banks of usable test questions, thereby avoiding rote learning of items (Marimwe & Dowse, 2017; Wei & Lin, 2014). These tests may also be adaptive in nature, allowing test users to pinpoint the true score of test takers with fewer items (Hambleton & Xing, 2006; Wyse & Albano, 2015). In addition, shortened assessments can be conducted periodically, thereby extending the assessment period (Geisinger, 2013). This creates a more holistic picture whilst addressing the traditional criticism that a test only offers a limited snapshot of a test taker, based on an observation of behaviour over a short period, and within a highly artificial test environment. Electronic devices can further be worn by test takers as part of their daily lives and record relevant data for a subsequent time-span analysis that would be more holistic than a once-off test session (Bers, 2012). By introducing psychological testing into the

daily life of test takers, through the aid of electronics and the internet, assessment can become less of a disruptive event for the client, and can allow for continuous monitoring of clients.

An exploration of change in knowledge and the internet introduces a further motivation for test revision. The internet is a platform that makes knowledge readily available, often without regard for copyright or trademark infringement. A test is a created product that facilitates objective and standardised observation of a test taker (Laher & Cockcroft, 2013). Once a test appears on the internet much of its value is lost; this is especially the case for ability or performance tests. A code of conduct for registered psychological test users is that they need to maintain test security, to prevent a test being circulated within the broader society. The internet can become problematic in terms of test security, especially if a hacker breaches the security measures of an online test (Foster & Miller, 2010). When test security is breached to the extent that anyone can gain access, the test may lose its validity, which would then enforce the decision to revise the test.

The effects of changes in the profession on a psychological test.

As a living science, any discussion of the field of psychology would be remiss without mentioning globalisation. We live in a global village that both broadens our access to new knowledge and understanding about psychology, as well as narrows the connective reach between us (Kames & McNeely, 2010). Tests operate in an evolving context and, as they are designed to provide valid results for a specific situation, tests are susceptible to changes in external stimuli. Iconic tests, such as the Minnesota Multiphasic Personality Inventory (MMPI) (Butcher, Graham, Tellegen, & Kaemer, 1989; Hathaway & McKinley, 1942; Tellegen & Ben-Porath, 2008/2011) or the Wechsler Adult Intelligence Scale (WAIS) (Wechsler, 1955, 1981, 1997, 2008), for instance, have found markets beyond their intended original scope, which in

turn necessitates revision and norming that take such international markets into account (Silverstein & Nelson, 2000). Revising a test for international use has its own set of challenges. The most prominent challenge is the sociocultural relevance of underpinning constructs and test items. Test developers should interrogate tests from the vantage point of international relevance, as it is very likely that a test will be used beyond its intended market, with the possibility that the test will be used, researched, standardised or adapted for test populations outside the developer's original target market (Geisinger, 2013).

Globalisation highlights the issue of who may use certain tests. In those countries that have professional governing bodies, professional knowledge is related to the requirements of relevant professional boards. The regulations of such boards may include the registration and use of psychological tests. Professional boards of psychology often adhere to strict policies about how tests should be used, and what best practice guidelines or prescriptions would apply to test users or test publishers. Test developers face pressure from both test users and regulatory bodies to offer tests that are at the forefront of knowledge. In particular, test users have expressed a desire for more regular updates of tests to promote the confidence with which such tests can be used (Foxcroft, Paterson, le Roux, & Herbst, 2004). Test users have more options available when purchasing tests and, with the information-age and ready access to published research on tests through the internet, test consumers have become better informed and more demanding about the quality of the tests they use.

In order to promote the quality of tests, professional boards increasingly require that test users select from tests that have been submitted for quality assurance and approved by test standard bodies (AERA, 2014). Test classification systems are comprehensive, taking into account all test properties, with each subsection of the rating system (including norms, test

materials, and underlying constructs) contributing to the total classification. An example of this is the classification system used by the European Federation of Psychologists' Associations (EFPA, 2013). In the EFPA system a test is scored on numerous facets, including the accuracy and currency of the underpinning theoretical constructs, the standardisation information that is provided for the test, the published research evidence on the soundness of the test, and how dated the norms are. Each aspect of the test is scrutinised and awarded a mark that is summated to a total score on which a test is judged as satisfactory or not for professional use. Although the EFPA system sets no time limits on revisions, the norms of any test must be less than 10 years old to obtain an 'Excellent' classification (EFPA, 2013).

The EFPA standards were drawn largely from the Dutch Committee on Tests and Testing (COTAN) (EFPA, 2013). Evers, Sijtsma, Lucassen, and Meijer (2010) indicate that for COTAN classification norms are considered as outdated after 15 years. COTAN undertakes a re-evaluation of each test's norms on an annual basis and, if norms are over 20 years old, that subsection of the application is automatically downgraded to 'Insufficient' which could have serious implications for the total classification score of the test. Stringent requirements such as these place additional pressure on test developers to be involved in continuous research and revision to retain or improve the classification rating of a test.

Professional requirements, such as test classification, play an important role in the decision to revise a test. Professional bodies, such as the American Psychological Association, publish diagnostic criteria for psychological disorders. The manuals that contain these standard criteria are regularly updated which, according to the ITC (2015), leads to a likely revision of tests that are used for diagnostic purposes in response to changes in the criteria (Standard 5.4).

Although tests operate within the scope of the profession of psychology, they are used within the broader sphere of society. As such, some tests may be influenced by societal norms or the political climate of the day. Test stimuli may draw from what is viewed as normal, usual, or reflective of society at a given point in time. As society changes, so too will what is considered relevant for such tests (Foxcroft & Roodt, 2014). Societal change may accompany political shifts that would also affect a test. The words that filter into society and political dialogue, such as terms relating to psychological illness or distress, would affect the text that is used in test manuals and test reports. Over time, certain words may no longer be politically correct or understood, thus requiring language changes to test materials.

A test is designed to engage with the broader community at multiple levels, and different users will place different expectations on their interactions with a test. As such, test publishers should have a proactive stance towards test revision. Test user feedback and published research should be collected continuously from different sources to determine the best time to embark on a revision (AERA, 2014). Due to the length of time required for the revision of a test, publishers cannot wait for a major event to derail a test before seeking information for the purpose of revision. The competitive nature of the test publishing industry does not support a reactive approach. The intricacies of the revision process are explored in the next section.

The Process of Test Revision

Test revision is a complex process that increases the scope for misunderstanding to occur. The first misunderstanding is that revision is not the same as development. Test revisions face a set of challenges that are different from a test that is being developed and launched. Once a test is launched, it becomes an entity that engages with multiple role players at different levels (Geisinger, 2014). When a decision is reached to revise a test, some of these role players may

play a role in the revision process. The individuality of each test revision process may in part contribute to the lack of clear guidelines for test revision, as it is difficult to visualise a generic process for all test revisions, although the present researcher believes that there is such a generic process. The post-revision process of launching, promoting and marketing a test additionally needs to be carefully managed in order to create a sense of continuity for test users. Each of the above aspects are explored in the sections below.

Tests as entities.

Tests can be viewed as entities that are expected to fulfil various roles. The primary purpose during their creation is to serve as academic entities that operationalise psychological theory and constructs in tangible and objectively measurable behaviours. After a test becomes eligible for purchase it also becomes an economic entity that is expected to provide an economic return to its creator and publisher through sales, licensing, scoring and reporting fees, and test-user training and support. At the point of contact with clients, a test is also a point of service in that it measures constructs and provides an objective reference for feedback to clients. The various entities that a test comprises create scope for the relevant parties to feed into future revisions of the test.

Role players in test revision.

Tests are usually developed by specialists in the field with years of knowledge about the domain and a passion for developing a test in line with their understanding of a specific construct. Today, test publishers purchase the most widely used tests from the original developer for ease of distribution and to maximise the associated economic benefits for the testing industry (Adams, 2000). The test revision process therefore becomes complicated, as more role players are involved, including subject-matter experts who may have a different theoretical slant to the

original developer, consultants in psychological testing, and business representatives from the test publisher. A revised test carries the added burden of having to advance the economic interests of the publisher, whilst opening up new markets for the brand. This creates pressure and conflict during the revision process that is often not anticipated at the outset of the process (Brannigan & Decker, 2006).

An example of conflict faced during revision is between cost and time constraints versus diligence in respect of the validity and reliability of a test. It points to an ethical dilemma where test publishers and academic development teams may be at opposing ends. From the test publisher's perspective, a test should meet the market's needs, maximise economic profit, whilst minimising risk to test takers and associated legal problems. The academic team, who would have their names and institutional affiliations associated with the test, would aim for a gold standard in terms of validity, reliability, and fairness of measurement, as their academic reputation and possibly their future career may depend on the quality of the test. These opposing forces can pull a project in different directions, and the decisions they inform will affect the final product. There are numerous examples of cases where insufficient piloting and pretesting, for instance, created serious problems for a test after it was published (Rhoades & Madaus, 2003).

These latent errors, or errors in the management of a process, are created when hasty process decisions have unanticipated effects on the final product, such as when a test is rushed to market without the required foundation of validity and reliability. Bush (2010) places much of the blame for this at the door of profit-driven test publishers, and argues that in some cases test revision processes should not include a test publisher. Although this opinion may appear to be sound, it may not be the best way forward given present economic realities combined with the

expense and complexities associated with test revision, product marketing, sales, liability, and copyright infringement.

A more considered approach may be to outline roles and responsibilities at the outset of the process. This recommendation is not new as Foxcroft, Paterson, le Roux, and Herbst (2004) suggested a model for test development that delineates the roles and responsibilities of those involved in the process, with the aim of improving the quality of the product as well as the service rendered to test users. As such, the four main role players that may form part of a revision are the academic team, the test publisher, test users, and clients that form the intended population for the test (ETS, 2014). The present researcher has identified the involvement of these role players at the various stages of test revision after a review of the literature used in this study as well as personal experience of test revision. As such, Table 1 reflects the researcher's original interpretation of the participation by role players in test revision. These levels of involvement formed part of the guidelines for test revision developed by the researcher in objective one of this study, and were peer-reviewed in step 10 of the guideline development process followed in this study, as detailed in Chapter Five.

Table 1. Roles and responsibilities of role players in test revision

STAGE	ACADEMIC TEAM	TEST PUBLISHER	TEST USERS	TEST CLIENTS
Prior to revision	Engagement in research	Monitoring research generated and collating feedback from test users	Collating personal test observations; engagement in research	Undergo testing and providing feedback to test users
Project management (process, timeline, budget)	Involved in project team	Involved in project team	Not involved	Not involved
Defining scope of revision	Providing insight on academic goals of the project	Providing insight into economic viability of revision process, including budget	Providing feedback on personal experience of using the existing test	Not involved
Developing test constructs, items, and equipment	Deciding on final construct definitions; developing items	Providing input on test equipment; support with publishing; steering the look and feel of the final test product.	Limited involvement, if any	Not involved
Gathering test data	Collecting data for experimental, pilot and final version of test	Providing budget support for data collection	Limited involvement, unless approached by project team to assist	Undergoing testing, if selected through sampling
Establishing test properties	Involved in data capturing, management, and data analysis	Involved in quality control of data analysis and presentation of results within test manuals	Not involved	Not involved
Launch of revised test	Involved in planning of launch	Manages test launch	Present at test launch	Not involved
Marketing and sale of revised test	Limited involvement, apart from sharing relevant insights through academic forums	Manages marketing and sales	Purchases test; informs colleagues of revised test	Not involved
Coordinating post-launch research, training, and test use	Involved in research and training	Involved in training and technical support; protecting copyright of the test; continuing the relationship with test users	Undergo updated training; use the test; engage in research on the test	Undergo testing and providing feedback to test users

As can be seen from Table 1, different role players are involved at various points in the revision process. The table also addresses the suggestion of Bush (2010) regarding the involvement of the test publisher. From Table 1 it is clear that the publisher plays a vital role in test revision, due to their experience of the publishing field and their production of a product geared to the market, as well as their continued relationship with test users (ETS, 2014). From Table 1 it would appear that the revision project team consists mainly of the academic test developers and the test publisher. This is the core of the project team, with additional members or experts co-opted when required at different times. The project team should decide on the target population for the test, the constructs measured, and specific items. These functions would rest mainly with the academic team, but with sufficient input from the test publisher to ensure that the revision satisfies market needs, and economic viability indices. The publisher will play the major role in branding, marketing, sales, and communication with test users. Test users and their clients are also included in the project for their valuable input of their experience with the previous version of the test, as well as their exposure to the revised version during pilot and standardisation testing (ETS, 2014).

Test revision is a time-consuming process that can easily veer off-course, therefore clear timelines, as well as a contingency leeway, may be advisable. However, this can only be formalised if the test revision team has sufficient knowledge about what a revision process entails, and can draw on an adequate set of guidelines that would alert them to hidden expectations as well as potential pitfalls (Aylward, 2009).

Towards a generic process of test revision.

One reason why guidelines on test revision have omitted specific information about the process of test revision may be the assumption that there is little difference between the

processes for test development and test revision. As explained in preceding sections, tests are entities that engage with a number of interested parties. The process of revising a familiar and often-used test should therefore be handled with great care to avoid effecting changes to the test that result in unintended consequences, such as changing its intended purpose, how it is used in practice, together with confidence in it as well as buy-in from test users (Bush, 2010).

Whilst cognisant of potential reservations about a one-size-fits-all approach, the present researcher argues that a generic process would assist in understanding the key operational aspects of test revision. The researcher therefore developed a generic test revision process to create an overview of suggested phases and tasks in the test revision process, taking the contribution of different role players into account. This process was peer-reviewed in step 10 of the development of test revision guidelines (objective one of this study), which will be detailed in Chapter Five. As the process is intended to be comprehensive, it would be more applicable to 'medium' or 'extensive' test revisions. Revision teams conducting 'light' test revisions would need to apply those stages that fall within their project scope. The phases of this generic process are presented in Table 2.

Table 2. Generic process of test revision

PHASE	SPECIFIC TASKS
Phase One: Pre-Planning	• Establish executive project committee • Assign roles and responsibilities to group members • Determine timeline for Phase Two
Phase Two: Initial Investigation	• Collating and reviewing available research outputs • Consulting test users about their experiences of using the test, their criticisms and suggestions for future revisions • Reviewing test materials (including manuals, test questions, test instructions, equipment, norms) • Consult experts in the construct and in psychological testing • Consider financial viability of the test (sales, training, and interest from the test-user community) • Identify the strengths, weaknesses, opportunities and threats of the test (using above information)
Phase Three: Project Planning	• Decide on the extent of the revision • Establish academic sub-committee, and identify ad-hoc members • Write project statement regarding the intended revised test (including intended population, test use, and test user) • Determine cost of the project • Plan the subsequent project phases and establish timelines
Phase Four: Academic Enquiry	• Review and refine test construct definition, including conceptualisation and operationalisation • Develop a test matrix • Compare existing test items, instructions and equipment to updated construct definition • Analyse existing item difficulty, discrimination, and fairness • Populate the test matrix with viable existing items • Decide if existing items adequately cover new test matrix
Phase Five: Item Development	• Identify item gaps in test matrix and construct new items, including backup items • Develop an extended experimental version of the revised test for field-testing • Review and / or develop test and item administration and scoring instructions • Apply for ethical clearance for data collection (consider ethical implications and permission for upcoming pilot testing, and standardisation sampling) • Gather and analyse data on experimental version, including feedback from administrators and experts • Refine test items

Table 2. (continued).

Phase Six: Test Piloting	• Develop a pilot version of the revised test • Administer pilot test to a larger sample • Perform preliminary data analysis and item analysis • Finalise composition of final test, including items, item-order, test materials, scoring and instructions
Phase Seven: Test Standardisation	• Select representative sample for test standardisation • Collect, store, capture, clean, and analyse data • Develop norms and standardisation information for the revised test
Phase Eight: Conduct Supporting Research	• Conduct research on clinical or non-clinical samples as required • Conduct research to establish correlation or relationship between previous versus of the test and the revised version • Conduct other relevant research required to accompany the launch of the revised test
Phase Nine: Test Product Assembly and Launch	• Develop final test manuals, including revision history and process, test and item administration, scoring, standardisation information, norms, and supporting research • Develop training materials for new and existing users • Launch the revised test
Phase Ten: Post-Launch Activities	• Conduct and disseminate additional research (such as correlation with other psychological tests and predictive validity for behavioural constructs) • Register test with required test classification agencies or professional organisations • Continue marketing revised test and engaging with test users

Table 2 presents a comprehensive overview of the key tasks within each of the ten phases of test revision, as developed by the researcher, thereby reflecting an original peer-reviewed contribution of this study. The scope and extent of each revision process rests on the decisions taken along the way. These decisions direct the process in a specific direction and, as such, require careful consideration as well as sufficient backing and documentation in order to defend or support the entire process (ETS, 2014). Each task requires meticulous effort as it becomes increasingly difficult to amend a previous task in the process of subsequent phases. According to van der Linden (2005), despite a century of development in test theory, no technology exists to

develop tests to the rigorous specifications intended by test developers. Van der Linden mentions further that the goal in each project is less to develop good tests and more to prevent the development of bad tests. This observation again points to the importance of the human component in developing and revising tests. It also underscores the value of a rigorous process that is planned with foresight, documented with care, and interrogated at every turn. Clarification of certain tasks within the test revision phases is explored below to highlight some of the available options and potential consequences of specific decisions for the project.

Phase One: Pre-planning.

In Phase One the establishment of the executive project committee is the key function. The reason for delineating the process of committee selection as an entire phase of the project is that the success of the revision depends largely on the ability of this group to work together and make the correct decisions throughout the process to deliver the final revised product. This group will have to start the initial investigation of the test and therefore the group should comprise suitably qualified members that are knowledgeable about the test and have insight into the test-user population, the broader psychological test market, and project management experience (ETS, 2014). Group members must share a passion for the test and a desire to preserve its legacy. As the test publisher may provide substantial financial support for the revision, they would need to be adequately represented on this committee. In most cases, the test publisher owns the test and therefore their needs should be considered during the process.

Phase Two: Initial investigation.

The purpose of Phase Two is to source as much information about the test as possible, including research reports and expert feedback on the construct and test properties (AERA, 2014). As the main purpose of the test is to form part of test-users' toolkits, registered test users

need to be consulted regarding how they use the test, what its main strengths are, as well as what they experience as problematic and in need of change within the test. Feedback from test users will also provide some indication about the prevalence of change resistance that the revised test may encounter. The voices of test users cannot be overemphasised as they provide fertile ground for suggestions about how to improve the test. By seriously considering these inputs test users will feel valued and included in the revision process, thus increasing their sense of investment in the test as well as the likelihood that they will be more open to the revised test (Geisinger, 2013). As the test publisher is the first point of contact with the test's users, their customer support, marketing, training and sales representatives should also be consulted for feedback on the test. At the end of this phase, the executive project committee should have a comprehensive overview regarding the strengths and weaknesses of the test, as well as how it compares to competing tests in terms of quality, innovation, usability, and customer service (Geisinger, 2013). Despite the importance of test user feedback, Gregory (2015) states that this step is usually overlooked in test revision projects.

Phase Three: Project planning.

There are three key aspects to Phase Three. Firstly, the extent of revision needs to be decided. This may take some time and negotiation between what would be ideal versus what is feasible within budget and timeframe constraints, whilst allowing some contingency for extraneous variables that affect the scope of the revision and the resources allocated for the project. The extent of the revision provides insight into the extent that an academic committee would need to engage with the test's constructs, content, and materials, together with who should form part of the academic subcommittee (Aylward, 2009). If changes are required in terms of the target population of the test, the contexts where the test should be used, or who the test users

would be, the second key aspect of this phase would be for the academic subcommittee to draft an appropriate project definition that encapsulates these aspects. The third key aspect is to establish a clear timeline and budget for subsequent phases of the project. This will involve considerable planning to keep the project moving forward whilst not moving too quickly and thereby failing to maintain professional due diligence throughout the revision process.

Phase Four: Academic enquiry.

The implementation of Phase Four depends on the extent of the revision. If during Phase Two gaps in theory or test content were identified that required correction, it would denote at least a 'medium' or 'extensive' revision. In such cases, the academic subcommittee should research and draft a new conceptualisation and operationalisation for the test. This is a critical aspect of the project, as Geisinger (2014) refers to the link between test quality and specifications used to construct test items and components. Given how crucial the conceptual definition and operational framework is to the eventual test, the development team should exercise great care and diligence in this task. A test matrix should be established for the new test by analysing the previous test matrix and individual items. Viable items from the earlier test should then be compared to the new test matrix in order to populate it. This phase would culminate in either a populated or partially populated test matrix.

Phase Five: Item development.

In Phase Five items are developed to fill gaps in the test matrix. It may be prudent to develop surplus items for an experimental version of the test in the event that items within the test matrix do not perform as expected and require replacement. The experimental version should be administered to a small sample for the purpose of obtaining qualitative feedback from experts,

administrators and test takers. This information will be valuable in refining items, together with administration and scoring guidelines.

Phase Six: Test piloting.

The purpose of Phase Six is to determine what items would form part of the revised test. Item placement should be approximated, and a near-perfect pilot version of the test should be developed. This version should be administered to a larger sample to obtain final feedback from administrators. The quantitative item data should be analysed to verify that items are placed in the correct position. The conclusion of this phase is to finalise the test items, their placement, materials, and administration and scoring guidelines.

Phase Seven: Test standardisation.

Some might argue that Phase Seven is the most important aspect of test construction and revision. A test speaks to users and clients through the norms that convert raw test performance into a standardised score that is comparative to a reference group or linked to directive explanations or definitions of what a specific test score means in practical terms. In real terms, the norm-development phase is when the accuracy of the preceding development process is tested through administration to a large sample of test-takers. The financial outlay and logistical challenge of this stage, in terms of sourcing a representative sample, training test administrators and facilitating the process of testing, recording and accurately capturing test information, is considerable. Thus much time and planning should be spent on every minor detail of this phase. Part of the planning for this phase is to determine the data analysis route to be taken to establish standardisation information, item statistics, and norms.

The two main options available are traditional norms and continuous / inferential norms. Traditional norming links to classical test theory whereby norms are developed on larger

samples, usually around 200 test-takers or more in each target group for whom norms are to be developed (Bechger, Hemker & Maris, 2009). The main benefit of traditional norms is that it is a known process used in most tests. The main drawback is that, as target groups are analysed separately, there could be a disconnectedness between the norms and item statistics that are developed, mainly because of errors in sampling or measurement (Zhu & Chen, 2011). These anomalies would have to be explained in the manual, or corrected in the norm table through human judgment, in a process called *hand-smoothing* (Zhu & Chen, 2011, p. 9).

Continuous norming questions the artificiality of different target groups. What qualitative difference would there be for a child one day before her birthday and the day after that would result in her being placed in an older year group? Continuous norming views different target groups as overlapping, thereby taking adjoining target groups into account when developing norms. There are several benefits to continuous norming. Smaller samples of 30-80 per target group are required, as samples are overlapped to create a smoother norming curve (Zhu & Chen, 2011). The norms are created through modern statistical techniques, such as polynomial regressions and item-response theory. The resultant output increases the confidence of norms, with minimised anomalies, and less disconnected norm tables or, in short, more accurate norms (Evers, Sijtsma, Lucassen & Meijer, 2010). The drawback of continuous norming is that it is still relatively unknown, although increasingly used, which could increase initial suspicion amongst test users and add to change-resistance from what is known and trusted to calculations and explanations that may be too complex for users to readily comprehend. As continuous norming is still a developing field it may also not be appropriate for all tests, which would exclude this option from being considered.

Phase Eight: Conduct supporting research.

Phase Eight is intended to address scepticism from existing test users. The benefit of previous versions of a test is the body of evidence that has been created over years of research. A revised test will not have the same extent of research which could lead to scepticism about how well the test performs in applied contexts (Bush et al., 2018). The revision team should therefore identify key areas of common use for the test and conduct research on the revised test to establish its usefulness for these contexts. It would also be useful to conduct studies on the qualitative and quantitative relationship between the new and old version of the test. This could create a connection between the two versions of the tests and illuminate the sense of legacy of the test as a continuing brand.

Qualitative studies could include analyses of the underlying constructs that are measured by the tests, as well as how they are measured. According to Liu and Dorans (2013), research should confirm the natural expectation that, after a minor revision, scores could still be interchangeable through equating formulas. After a medium revision, changes may be too substantial to equate scores, but some concordance between them could be established. The product of an extensive revision could yield a test that is so dissimilar from its predecessors that no linear relationship or even concordance can be established in raw scores (Liu & Walker, 2007)

Phase Nine: Test product assembly and launch.

Phase Nine is the envisaged product of the revision process. The efforts of the preceding phases culminate in the test, test manuals, and test materials. It is during this phase that the expertise of a seasoned test publisher is paramount, to develop materials that will capture the interest of test buyers. All elements of the test product must project a single and cohesive vision

that sets the test apart from its competitors. Psychological testing is an expanding and lucrative business and it involves major players in the industry, thus every effort must be made to make an individual test stand out. As part of the rollout of a test, training courses for new and existing users may be offered, to both generate income to offset the expense of the revision as well as to market the revised test. The maximal potential of the relationship between the test publisher and test users can be realised by offering client services, ranging from online scoring to membership of a test-user association to foster collegiality and common interests in the use of the test. The appropriate rollout of a test is fraught with difficulty and this will be discussed later in this chapter.

Phase Ten: Post-launch activities.

A potential oversight of some development teams is to consider the revision as finalised once a test is launched. This is far from accurate, as is highlighted by the activities of Phase Ten. It could be argued that the work really starts once a test is released into the market. Tests are usually launched with only sufficient research into validity and reliability to allow for general use. A newly launched test affords new research opportunities in its usefulness for special populations, cross-cultural studies, validation studies with other tests of the same construct, and local norms for different countries (Gilmore et al., 2015). The revision team should be active in such research to promote interest in the test and to maintain it in the spotlight within the professional and research community (ETS, 2014). Ongoing research will add to the body of knowledge about the test, and it is this evidence that will accompany the test's classification application with registration bodies.

This section described a generic process of test revision to shed light on the many intricate tasks such a process comprises. Due to the academic nature of many activities in the process,

revision teams can easily lose sight of the context that surrounds a test, specifically the human element. Test users have complex reasons for preferring a specific test, and some may object to any changes in the test. Despite a revision team's noblest aims, the test is created for test users to develop their own relationship and user-practice with it. This requires special focus from the revision team and test publisher, a focus that explored in the following section.

The process of managing the launch of a revised test.

A conflict that can be anticipated during test revision is resistance to change. Despite the best efforts of project teams Adams (2000) cautions about negative feedback from test users. Some people are simply averse to change and may have become so familiar with the previous version of a test that anything new will be received negatively (Silverstein & Nelson, 2000). Sometimes, however, there may be more to such resistance than just an unwillingness to change. Strauss, Spreen and Hunter (2000) refer to a body of research that has shown how even minor changes in tests, such as the *WAIS* and the *MMPI*, have resulted in significant changes in research findings on the factor structure and underlying constructs of the tests. Widely used international tests such as these are also often used for longitudinal tracking of changes in society. By implementing even minor revisions, a research project spanning decades could be affected and possibly rendered invalid.

McCauley and Strand (2008) also report notable differences in the level of research evidence supplied with psychological tests. This may lead to confusion in test users about the relationship between test scores of previous and revised versions of a test, a situation that could be avoided to some extent had the test revision team conducted and disseminated such studies to inform test users.

A further issue relates to the transition period between revisions of a test. According to Bush (2010), the industry norm is for test users to migrate to the latest revision within the first year of its release. Ethical standards by various organisations (including the APA) also advise test users against using tests that are obsolete. There is no consensus however on when a test becomes obsolete. In general terms a test is viewed as obsolete when the underpinning theory, test questions, norms, or technical aspects are no longer fit for the purpose of accurate, fair, and professional assessment, and if a test's continued use may lead to inaccurate diagnosis and recommendations (ITC, 2015, Guideline, 7.1). With this said, the ITC (2015) leaves this judgment call to the individual practitioner, a standpoint that does little to alleviate general concern (Bush et al., 2018).

In making the transition to the latest version of a test one benefit of such a shift at an early juncture is that it provides fertile ground for new research, which could inform the greater user population (Adams, 2000). One drawback identified by Bush (2010), however, is that it often takes years for revised tests to be validated on clinical samples. This information would be available for the previous test version and means that if a specialist requires it, the old test would be preferable in the interim over the revised alternative. With the increasing rate of test revisions, local norms for an older test sometimes only appear after the publication of a revised test edition in the source country (Gilmore et al., 2015). This would mean that test users in some countries are continuously using older tests than practitioners from other countries, which may affect their ability to contribute to knowledge generation internationally.

Considering these opposing views, it would then be advisable for a test publisher to have a changeover strategy in place whilst sufficient data is gathered for special populations. This would see the previous version of the test used for clinical diagnosis and management, and the

revised version employed for test takers similar to the sample used to develop the initial test norms (Bush et al., 2018). Test users should be advised throughout a changeover strategy and updated on any new developments, together with a general caution issued by the developer to use good clinical judgment when selecting the test version that is appropriate for each client (Bush et al., 2018; ITC, 2015; Silverstein & Nelson, 2000).

Finally, Adams (2000) highlights the difficulty that test publishers face in convincing test users to part with money for more expensive revised tests. This is a particular concern for test users from developing countries. Such practitioners face a purchase price for international tests, in which the costs are comparatively inflated by transport fees, import duties, and a weaker exchange rate of the local currency (Gilmore et al., 2015; van Dulm, 2013). According to the ITC (2015), test developers have a reciprocal relationship with test users. Developers respond to market needs that have been communicated to them by consumers. Consumers constitute an international base of expertise concerning developments in the domain being assessed as well as new ways of thinking about and assessing psychological constructs (Kames & McNeely, 2010). One suggestion is for test publishers to develop a more sustained relationship with test users by increasing their level of service through online forums, telephonic assistance, online scoring and report writing using more frequently updated norms, a service level that has been in existence for years from information technology companies (ETS, 2014).

By facilitating an ongoing dialogue between test users and test developers, test revision can become a more continuous process, with extensive revisions viewed as revision moments rather than another opportunity to cash in on a beloved test (Foxcroft, Paterson, le Roux, & Herbst, 2004). It may also foster product loyalty in test users and raise their status from passive consumers to active participants in the creation of a test and ongoing research. It is this body of

research that is generated on a test that should form an important component of the information circulated by the developer to users in newsletters, journals, as well as test manuals of subsequent revisions of the test to develop a rationale and justification for the need for a revised test (ITC, 2015, Guideline 2.4). This practice of open line communication will allow consumers to understand how a revised edition of a test adds fresh perspective and value, thereby serving as motivation for the additional expense of purchasing a revised test, as well as updated product training (ITC, 2015, Guideline 2.5).

According to the American Educational Research Association (AERA, 1999), tests should only be advertised as revised if they have been changed in significant ways. If only minor changes are effected, the test should rather be marketed as "with minor modification" (AERA, 1999, Standard 3.26, p. 48). This would be applicable to a 'light' test revision where slight changes are made to item instructions in the test manual. It would not apply to a 'medium' or 'extensive' test revision that involves changes to test items and the statistical information of the revised test. By embracing the benefits of an online platform, test developers can use this guideline to greater effect, by seamlessly making smaller adjustments as required, without releasing a new version each time.

Despite the increasing frequency of test revisions, this function of psychological testing has faced major obstacles in South Africa. This is explored in the next section.

The Revision of Psychological Tests in South Africa

Test revision projects require expertise, and human and financial resources (Gregory, 2015). The cost implications can be prohibitive for developing countries such as South Africa. As a formerly colonised country, South Africa is coming to terms with the calls for decolonised education and acknowledgement of indigenous knowledge systems (Burke et al., in press). In the

discipline of psychology, this is reflected through the term *African Psychology*, which requires an interrogation of the suitability of theories and psychological products generated outside South Africa, and the development of theories and services that are relevant to the South African population.

Decolonisation has extensive ramifications on former colonies, and South Africa in particular. In the early 20th century, psychological theories and test products developed in the USA and Europe informed tests developed in South Africa (Van Eeden & De Beer, 2013). During Apartheid test development and revision were more focussed on the white population, with such tests as the Senior South African Individual Scale – Revised (SSAIS-R) (Van Eeden, 1997), largely ignoring the majority black population. This was particularly evident in the 1992 revision of the *SSAIS-R*, where the black population was not considered in the construct delineation, item development, instructions, standardisation and norming. Despite this, the *SSAIS-R* was used on all South Africans, with the understandable underperformance of many black test takers in comparison to white test takers. Such biased psychometric tests were misused for selection in higher education and employment settings, thereby reducing the access and advancement opportunities for black South Africans in particular (Sehlapelo & Terre Blanche, 1996).

The end of Apartheid halted large-scale test development and revision in South Africa, given the mistrust of psychological tests expressed by the African National Congress (ANC) led alliance who assumed majority leadership of the country's national government in 1994 (Laher & Cockcroft, 2013). The cautions regarding psychological tests are clearly expressed in the *Employment Equity Act (Act 55 of 1998),* which states that tests may only be used in South Africa if they can be demonstrated to provide assessment results that are fair and unbiased to test

takers from all intended groups within South Africa. This places a great burden on test development and revision projects to provide adequate research support of cross-cultural validity.

The major objections and obstacles to psychological testing were mostly overcome in the early 2000s but by that stage much expertise in the development and revision of psychological tests had been lost. Since then, no large-scale national test has been developed in psychology in South Africa (Foxcroft & Roodt, 2018). Tests have tended to be on smaller samples or on single constructs. No test revision of large-scale South African tests has occurred either. Psychological testing has continued in South Africa, however, but with most tests being imported from the USA and Europe. Despite concerns about tests developed during the Apartheid era around three decades ago some tests, such as the *SSAIS-R* (Van Eeden, 1997) continue to be used with its original norms and not with updated psychometric information (Gadd & Phipps, 2012).

The delay in the development and revision of psychological tests in South Africa does not appear to be due to a lack of interest or calls for such projects, but as a consequence of other concerns. The effects of context, time and new knowledge on the validity of psychological tests were explored earlier in this chapter. The validity of many well-known tests in South Africa, such as the *SSAIS-R*, has been affected by all the above effects. The *SSAIS-R*, for instance, was developed and normed decades ago. The theoretical underpinnings of the test's construct would therefore need to be re-evaluated, to not only include the changed social context and political ideology of post-Apartheid South Africa, but also the underpinnings of intelligence from a non-Western, African perspective. The norms are also outdated, due to the observed Flynn effect, particularly in developing countries such as South Africa (Daley et al., 2003). As the sampling was not reflective of the country as a whole, a future revision of the *SSAIS-R* would also require

a comprehensive and inclusive sampling framework to produce valid norms for all South African population groups. Such a revision would be extensive in scope.

Important for South Africa has been ongoing research and theory development of *African Psychology*. The development and revision of psychological tests in South Africa is also constrained by limited staff capacity and expertise. The present funding opportunities in South Africa are inadequate to cover the expense of an extensive revision particularly for a large-scale test such as the *SSAIS-R* (Foxcroft et al., 2004; Gregory, 2015; Laher & Cockcroft, 2013). The test-user market has also developed a preference for the benefits of test products and support services of international test companies.

The advantages of employing widely used international tests are that they allow local test users to engage with counterparts in other countries, which allows for collaboration in international studies, thereby stimulating professional debate in international platforms and increasing the likelihood for publishing in international research journals. Considering these potential gains, one approach for South African professionals interested in test revision would be to collaborate with international revision project teams. By doing so, South African professionals would have a platform to inform construct delineation and item development, and perhaps even contribute South African data on international tests to inform final item selection and thereby reflect South African test takers in official standardisation information and test norms. International revision teams may also be more open to funding South African research projects linked to an international test, especially if such studies are by South African professionals with an established record of collaborating with the international revision team.

Concluding Remarks

This chapter considered the practice of the revision of psychological tests. The purpose of a test is to form part of the toolkit of professionals in the field of psychology. Such professionals depend on the reliability and validity of the instruments they use and, as such, tests need to measure constructs accurately and fairly (Dunbar-Krige et al., 2015). As with any created product, tests are subject to aging and lose their effectiveness over time. Test publishers are a key link in the psychological testing industry, connecting test developers with test users.

Test publishing is a lucrative and expanding field with multiple competing products vying for a share of the market. It is in the best interests of publishers to be proactive and to take heed of the different influences in the industry, including test users, test classification boards, professional boards, advances in theory, practice, and technology, and disseminated research. The feedback from these sources of information should be collated, considered, assimilated and acted on at the optimal moment to avoid costly mistakes. Test revision should be conducted with attention to accuracy and detail, whilst considering the abovementioned sources of information. In the end, the professional test-user community will judge the quality of the revised test and decide whether to utilise it or not. Despite the increasing frequency of test revisions internationally, this practice has faltered in South Africa. The example of South Africa highlights the dangers of psychological testing and how test revisions that are not informed by clear guidelines can fail to adhere to the standards of the psychological profession.

Although tests may be economic entities that have to satisfy professional boards and funding agencies, their core purpose and ethical obligation are to assist psychological professionals to reach accurate decisions for the benefit of their clients. The test publishing industry should strive therefore to maintain the highest ethical standards to promote the ethical

use of high-quality test products. The purpose and content of standards and guidelines for professional practice, particularly in psychological testing and test revision, are explored in the next chapter.

Chapter 4: Guidelines for the Revision of Psychological Tests

We live in a changing world. According to Blech (2007), the amount of knowledge doubles every one to two years and it is predicted that by 2020 this pace will increase to every 72 days. Interestingly, this prediction occurred before the advent of smartphones and apps in 2008. Thus, professionals in any discipline are on a lifelong journey of training and, given the rapid expansion of knowledge, much training only happens after leaving formal education (Merriam, Caffarella, & Baumgartner, 2007). Professionals improve on their discipline-specific knowledge by interacting with colleagues, furthering their training, keeping current on publications, and participating in research.

The process of formal, informal, and non-formal education is critical for professionals to remain current in their field and to acquire some level of expertise (Merriam, Caffarella, & Baumgartner, 2007). The way that this educational process occurs has changed dramatically since the advent of the information age. Ready access to knowledge is one of the benefits of the internet. Another benefit is that anyone with access to the internet has a platform to voice his or her opinion. However, this creates a repository of information where accurate, fact-based research and expert opinion is sometimes indistinguishable from uninformed opinion or outright misinformation. On the internet the lack of a peer-review that is a cornerstone of scientific research, is a disadvantage to the internet as a reliable source of information (Neuman, 2011). This chapter first explores how a specific profession deals with information by way of best practice guidelines for their field. The chapter then examines what guidelines are in general, the criticisms against guidelines, what the historical background is for professional guidelines, and how guidelines are developed. The chapter then reviews the current scope of guidelines in

psychological testing in general, as well as test revision specifically. Finally, the problem that forms the foundation for the present study is explored in detail.

What are Guidelines?

In general, professionals need to make management decisions in their daily practice. They have to weigh individual circumstances, the preferences of whomever they are working with, and anticipate desirable and undesirable consequences. Having integrated this information, they then need to select an appropriate response from a list of possible actions in order to deliver their service (Jaeschke, Jankowski, Brozek, & Antonelli, 2009). In many instances, this complex process flows smoothly and is well within the capabilities of the individual. Sometimes, however, the solution is less apparent, with several seemingly viable options available. In test revision such options would be the influence of extraneous variables on a revision project that requires critical decisions by the revision team. In these instances that present multiple options, a professional seeks external advice from experts about what route would be optimal. Sampson (1999) mentions six steps to use information effectively within a professional context, which is applicable to test revision. The first step is recognising a problem exists that requires additional information. The second is to select information relevant for the identified need. Thirdly, a professional must decide how to use the sourced information to address the need, followed fourthly by implementing the course of action they decided on. In the fifth step, the professional must decide if the course of action was successful in addressing the need. In the sixth step, help should be sought from other professionals or resources until the problem has been solved.

According to Weisz, Cambrosio, Keating, Knaapen, Schlich, and Tournay (2007), there are two dominant reasons for the existence of guidelines. The first is to expedite the process of accessing expert advice, whilst minimising the costs. This advice is often in the form of

published documents developed by experts to alert professionals to the best decisions relevant to the scenario they are facing (Jaeschke, Jankowski, Brozek, & Antonelli, 2009). The second reason is to protect the autonomy and reputation of a profession (Weisz et al., 2007).

There is ambiguity in the literature about terms such as ‘guidelines’, ‘standards’, and ‘policies’. The difference between these terms lies in the mandatory emphasis of the particular document, together with the level of operationalisation of the content. A *Certified Information Systems Auditor* study (CISA, 2011) offers insight into these terms from an institutional perspective. According to CISA (2011), policies are considered high-level documents that exercise control over staff, and they are usually enforced at managerial level. Standards are developed to promote uniform application of policy statements. Standards tend to state broad principles, but with compulsory compliance as determined by the organisation. According to the American Psychological Association (APA, 2017), standards tend to focus on broader issues such as acting with competence, dealing with ethical dilemmas, exercising respect for others, maintaining confidentiality, the right to privacy, seeking informed consent, and maintaining adequate records. In contrast, Proctor and Staudt (2003) define guidelines as “systematically compiled and organised knowledge statements to help practitioners select and use the most effective and appropriate interventions for attaining desired outcomes” (p. 209). Guidelines aim to have practical application and are developed when officially accepted standards are absent.

Apart from scope, an important difference between guidelines and standards is that guidelines are not necessarily formally adopted by an organisation, whilst standards are. National bodies, such as the AERA, APA, NCME and ETS have tended to name their documents ‘standards’ whilst the ITC as an international organisation has called their documents ‘guidelines’.

Whilst standards are produced through endorsement by an organisation, guidelines can be created by a single author without peer-review. A part of the process whereby a guideline can become a standard is through an external consultation and feedback system, and eventual adoption by an organisation. As such, guidelines may serve as the precursor to standards, by forming a foundation of explorative work. To facilitate the operationalisation of guidelines and standards, documents may also contain a procedural section. Procedures offer a systematic format for delivering the gold standard of service that is the basic goal of a policy, standard, or guideline. Procedures may be mandatory, but in most cases they offer a suggested route that may be deviated from or tailored to suit different situations (CISA, 2011).

A Critique of Guidelines

Practice guidelines in any profession are both praised and criticised. An explanation of the nature of guidelines would therefore be incomplete without some mention of the general critiques levied against them, as well as some of the issues around the implementation of guidelines. Guidelines exist to provide expert guidance to practitioners, but this purpose has raised concerns about the potential infringement on the autonomy of the individual professional (Weisz et al., 2007). According to the Educational Testing Service (ETS, 2014), however, guidelines are intended "to provide a context for professional judgment, not to replace that judgment" (p. 2). Guidelines are intended to build on the framework of the clinical judgement of professionals to assist them in making the best decision.

The limitation of guidelines is that they cannot foresee all circumstances, and therefore they cannot be applied without some level of interpretation (ETS, 2014). Additionally, guidelines draw from knowledge at the time and, with the rapid increase in knowledge and technological advances, may become outdated if such guidelines are authored in too rigid terms. The length of

time it takes to develop guidelines challenges developers to write definitive guidelines in such a way as to anticipate and possibly accommodate advances in the foreseeable future (ETS, 2014).

Some authors have expressed concern over the quality of guidelines (Siering, Eikermann, Hausner, Hoffmann-Eßer, & Neugebauer, 2013; Woolf, Schünemann, Eccles, Grimshaw, & Shekelle, 2012). The term ‘guideline’ has become linked with ‘quality’ and ‘best practice’, thereby inferring the authority of guidelines. A lack of a standardised format of guideline development can impede, however, the quality of guidelines as anyone with an interest in a field can develop guidelines, with little quality control over the final product. If quality is the intended outcome of a guideline document, then quality must also form the foundation for the development of the guidelines. The development of guidelines has become a burgeoning field in many disciplines, initiating guideline development centres and clearing houses in many countries and professions (Williams, 2017). As for quality, a reader can only speculate and look for evidence of validity and quality control in the guideline document to promote their confidence in the product.

Despite a plethora of guidelines, or perhaps as a result thereof, there has been concern about the low level of implementation of guidelines (Weisz et al., 2007). As guidelines tend to lack regulatory control, professionals can choose whether to follow them or not. Even if they are followed, personal interpretation of guidelines will determine how well they are implemented. Some have argued the lack of user-friendliness of specific guidelines (Kish, 2001). Guidelines prepared by experts can give voice to their years of experience. With this experience comes in-depth knowledge and even technical language that is outside the reach of many practitioners at an earlier stage in their career who may have the greatest need for such guidelines (Daly, 2005). It is therefore important that guideline statements are checked for clarity and readability.

Guédon and Savard (2000) echo the abovementioned criticisms. According to these authors, guidelines can fail due to two errors in interpretation. The first is that practitioners may not be able to interpret the sentiments of guidelines into specific individual situations. The second is that guidelines may come across as impersonal, thereby limiting the degree to which information is assimilated by readers. A facilitated process of human interaction and guidance can accommodate the social nature of humans, using social learning as a mechanism to facilitate transference of knowledge and experience. Guidelines that sound too abstract or too definitive may fail therefore in their primary goal as an educational mechanism based on the experience of experts (Guédon & Savard, 2000).

Some organisations, such as the American Educational Research Association (AERA), the American Psychological Association (APA), the National Council on Measurement in Education (NCME) and the Educational Testing Service (ETS), have adopted the practice of writing short and bold guideline statements that are then explained in a more technically focussed paragraph below each statement. Such a format tends to read well as it unpacks the debate around each statement, as well as the variety of options that may be considered in different contexts (AERA, 2014; ETS, 2014).

Despite these abovementioned concerns, Woolf et al (2012) argue that guidelines have become an indispensable part of efforts to improve quality of service. One challenge for practitioners is to assess the quality and importance of guidelines. Siering et al (2013, p. 4) have developed 13 quality dimensions for the appraisal of guidelines. Whilst these are generic, they would be applicable to most disciplines, including psychological testing. The quality dimensions are:

1. Information retrieval: This relates to the clarity of focus questions and the relevant outcomes intended by the guidelines. Further aspects include the process by which literature was sourced as well as criteria used to include and exclude literature.

2. Evaluation of evidence: This refers to the procedure to grade literature. An additional aspect is the accuracy with which research results are summarised by the guideline document, and the level of cohesion between literature and the guidelines.

3. Consideration of different perspectives: The norms and values that underpin the guidelines should be explored. The document should also include an evaluation of expert opinion, professional experience, and client input. The process by which these were considered, integrated into, and reflected in the guidelines should be explained.

4. Formulation of recommendations: The methods used to formulate guidelines must be explained. The strength of the recommendations should be stipulated to indicate the relative degree of certainty associated with a specific guideline, or with the guideline document as a whole.

5. Transferability: Clarity should be provided whether resources used to develop the guidelines were comparable to the settings for which the guidelines were developed. Implications for costs associated in implementing the guidelines in a specific setting should also be explored. For transparency, barriers or facilitators to implementing the guidelines in local contexts should be identified.

6. Presentation of guideline content: The benefits and the potential harm of implementing a guideline must be stipulated. There should be a direct link between guidelines and evidence.

7. Alternatives: In the event of conflicting evidence, alternative options to the specified recommendation must be presented. There should also be a description of situations where the

guidelines may not be applicable. For service-oriented guidelines, the right of clients to be offered alternatives, as well as their right to exercise that choice without fear or recrimination needs to be upheld in the document.

8. Reliability: Guidelines should be pilot tested before being released. An extensive peer-review of the document should be performed before broader dissemination.

9. Scope: The rationale and objectives for the guideline should be described. The topic and underpinning constructs should also be adequately explored within the document. The specific professional setting that the guidelines are intended for must be identified. The intended professional that the document is targeted at must be stipulated.

10. Independence: The group members that formulated the guidelines must be mentioned, together with their subject discipline or profession. The role of each member in the project should be highlighted. If a guideline is developed or funded by a specific organisation, this should be stated. Any conflict of interest in the membership of the development group should be declared and considered.

11. Clarity and presentation: The guidelines should be worded clearly and unambiguously. If the document contains procedures or a specific process, this must be presented with clarity.

12. Updating: The document should contain the date on which it is issued, as well as how current the evidence it contains is. If the guidelines have a lifespan, the end date that it can be applied must be stipulated.

13. Dissemination, implementation, and evaluation: The document should contain a section on how it is to be disseminated. It must include practical suggestions for implementation. Strategies for evaluating the guidelines, once they have been implemented by the broader community, should be thought through.

The above 13 dimensions developed by Siering et al (2013) offer a framework for evaluating a guideline document, as well as a structure for those that develop such documents. Whilst these dimensions have not been critiqued in the literature, they resonate with the earlier work of Savard, Gingras, and Turcotte (2002) who condensed many of the dimensions of Siering et al (2013) into five generic skills in information processing. The first skill is to communicate with others to identify a need. The second is to analyse a problem further to appropriately link the various subcomponents of the problem. Thirdly, a process of synthesis should be applied to information to create potential alternatives. The fourth skill is to attribute value to alternatives in order to prioritise them. Fifth, alternatives should be operationalised to develop executive strategies for different ends and means (Savard, Gingras, & Turcotte, 2002).

The abovementioned dimensions and skills point towards scientific rigour in collating and evaluating evidence, and a direct link between evidence and guidelines. The dimensions and skills emphasise clarity in how guidelines are conceptualised and written, as well as transparency in how and by whom they were developed. These aspects are closely linked to the quality of the final product, and therefore they offer greater precision for those concerned about the relationship between evidence, guidelines and best practice. Thus, they serve as a useful benchmark to assess the quality of the guidelines developed in the present study.

The next section will provide a historical overview of guidelines in broader professional contexts to underscore their increasing importance.

The History of Guidelines

A primary focus of professional bodies is to regulate the quality of service rendered by their members. Traditionally this has been accomplished by standardising professional accreditation, focusing on exit-level outcomes expected of professionals, and approving the

qualifications offered by different training institutions (Weisz et al., 2007). Internationally, the sufficiency of this practice has come under scrutiny (Gough, Thomas, & Oliver, 2012). One contributing factor for such scrutiny has been the rapid increase in knowledge that has resulted in diversification rather than simplification of understanding and practice. This confluence of factors generated a review of practices from the 1960s, with an increased focus on evidence-based practice. The 1960s not only heralded a time of political change that promoted globalisation, but also sparked a social revolution. The calls for equity and equality transformed into calls for social justice that were expressed through equitable standards of service for all and greater public accountability (Weisz et al., 2007; Wiener, 2000).

The focus on quality standards started with initiatives of employers to regulate the services by professionals in their employ. This practice of standards gained popularity, and eventually turned into industry-wide standards for different professional services (Weisz et al., 2007). This shift in regulation was alternately praised and criticised. In response to concerns about the rigidity of standards that undermine the personal decision-making of professionals, there was a shift towards developing practice guidelines. Such guidelines offered best practice advice, whilst creating opportunities for practitioners to use professional judgment. Although the ensuing evidence-based practice guidelines had limited legal force, they still conveyed the moral authority of the experts that developed them. The moral subtext of documents was seen, by some, as infringing on the personal freedom of individual professionals and over-regulation of their daily routines (Daly, 2005).

In the early 1990s the focus of guidelines was clarified as “systematically developed statements” (Field & Lohr, 1990, p. 38) aimed at assisting practitioners about appropriate actions for specific situations. This reconceptualisation improved the role of guidelines within daily

practice, whilst also redistributing the decision-making power to professionals. In a different sense, this move further narrowed the legal enforceability of such documents, as a black and white approach turned to grey areas that were open to interpretation. In either event, this redefinition highlighted the gap between suggestions and law that was subsequently filled by standards and policies.

During the 1990s in particular, the interest in professional practice guidelines expanded to social health sciences such as social work and psychology. In the field of psychology, prominent national and international organisations, including the American Psychological Association (APA), the British Psychological Society (BPS), and the International Test Commission (ITC), increased efforts to develop guideline documents. The process of developing guidelines in the social health sciences focussed more on expert input. One reason for this was the less available evidence-based research traditionally used to develop guidelines. The next subsection will describe the process of developing guidelines.

The Process for Developing Guidelines

As mentioned earlier in this chapter, different methods exist for developing guidelines. Some guidelines are research-based, whilst others focus on the experience of experts. Sometimes guidelines are developed by organisations or groups, whilst other guidelines are created by individuals (Jaeschke, Jankowski, Brozek, & Antonelli, 2009). There is no universally applied process. The reasons for this are as plentiful as the number of different methods. In some disciplines, a scientific fact-based approach works well as research evidence tends to come from quantitative studies. The emphasis on pure experimental research also produces findings that are more consistent, thereby allowing for definitive and absolute answers to research questions.

In contrast, some disciplines, including the social sciences, produce both qualitative and quantitative research (Kish, 2001). Quantitative research in the social sciences rarely meets the stringent requirements for experimental research, due to the influence of the mind and personality of research participants. The nature of qualitative research is to exert less control over data that is collected, but to strive for consistency and congruence between data collection, data analysis, research findings, and conclusions (Shenton, 2004). As such, qualitative studies rarely venture into the confines of classic experimental research. The interpretive lens through which qualitative data can be viewed, such as social constructivism, systems theory, and phenomenology, can introduce obstacles to the external validity of findings. The result is that, although findings may be accurate and consistent, the generalisability of findings beyond the scope of a specific research study may be limited (Neuman, 2011). This creates a body of research that is, at times, less cohesive and that requires greater expertise in interpretation. To produce guidelines for the social sciences, expert input is therefore indispensable in order to merge research with interpretive experience.

According to Woolf et al (2012), two major circumstances can occur during guideline development. The first is if the evidence is unclear, meaning that the best option amongst the alternative procedures cannot be determined solely based on research. This could either be the result of too little published research, or a strong body of evidence that points to divergent answers. The second is when, even in the face of evidence, there exists too much uncertainty about the relative benefit and harm associated with the evidence-based answer. The middle road would be to develop guidelines by blending research evidence and expert opinion.

Kish (2001) proposes a 12-step system for developing guidelines as an example of this middle road approach. These steps are outlined in Table 3.

Table 3. Steps for developing guidelines

Step	Description
1	Selection of panel
2	Introductory meeting of panel
3	Determine the scope of the guideline
4	Determine the target audience and the target population
5	Determine how the evidence will be selected
6	Select and review the evidence to be used in writing the guideline
7	Grade the evidence and determine what will be used and what will be discarded
8	Write the guideline, including an executive summary
9	Submit the guideline for outside review
10	Modify the guideline on the basis of the outside review
11	Submit the guideline to the parent organisation for review and publication
12	Review and update the guideline as appropriate

Table 3 demonstrates the importance of evidence in developing guidelines. Steps five to seven are the midpoint focus of the process, with time devoted to how evidence is selected, sourced, reviewed, and graded. Only after a guideline has been drafted is it submitted for outside review. It is however unclear whether such a review would be based on evidence, experience, or opinion.

For the present study, a literature search was conducted to establish steps outlining the process by which guidelines have been developed for the discipline of psychology, particularly psychological testing. Despite the existence of many guideline documents, the processes followed to develop them proved elusive. This is of concern, as any attempt to justify the quality of a guideline document that inherently is a product of research, should include a transparent outline of the process that was followed (Dijkers, 2013). From Table 3, however, it appears that a standard process (Jaeschke, Jankowski, Brozek, & Antonelli, 2009; Kish, 2001), irrespective of the specific discipline involved, would include the following five elements:

1. A clear demarcation of the scope or topic for the guideline, together with the intended audience. These aspects would create direction for the practice or scenario for which the guideline is written, as well as to which professionals it would apply.
2. Details on the process that was followed to develop and refine the guidelines.
3. Details of the sources of evidence used to formulate guidelines, how they were sourced, and the relevance of evidence assessed. Originally, guidelines were developed based on expert knowledge and experience, but they are increasingly being developed with a stronger focus on systematic reviews of published research evidence (Gough, Thomas, & Oliver, 2012; Gronseth, Woodroffe, & Getchius, 2011). The benefit of this practice in some scientific disciplines is that by virtue of controlled and replicated experimental research, a synthesis or meta-analysis of findings from different studies can offer an increased level of certainty about the most appropriate method of treatment. A concern however is that this practice veers away from the original purpose of guidelines, which is to seek advice from experts. Research also evidences variable degrees of formal peer-review prior to publication that may result in different levels of quality. A systematic review of evidence may also be compromised by publication bias, the practice whereby some journals are more inclined to publish studies that have statistically significant results over studies where no significant results were found.

Zuiderent-Jerak, Forland and Macbeth (2012) suggest that guidelines should reflect all knowledge, not just experimental research, which would require that literature searches are expanded beyond published articles to include internal reports and conference papers. Kish (2001) concurs stating that: “In many circumstances, scientifically rigorous material may not be available. In such circumstances, it is

appropriate to use expert opinion as long as it is clearly indicated and attributed" (p. 853). According to Siering (2013), this implies the presence of some informed human element in guideline development that reflects the learning of those with practical experience. The expert human element is indispensable therefore in drafting clear guideline statements. The process by which multiple sources of evidence were considered and merged into a definitive guideline statement through expert mediation, should therefore be explained in detail, together with indications as to whether certain guidelines are grounded more in expert opinion than experimental studies or peer-reviewed research.

4. The peer-review process should be detailed. As guidelines are informal documents written from an expert perspective for a non-expert audience, such a document would need to demonstrate some form of external review by experts prior to publication.
5. The lifespan of the guidelines should be communicated. Guidelines are based on evidence, and over time new evidence may surface that could challenge the certainty of previous evidence. A guideline document should detail therefore when it was developed, and whether it supersedes a previous document. There should also be clarity on whether the guidelines will be due for revision by a specific date, or whether the guidelines are contained within a living document that will be updated when required.

The above components work synergistically in the process of developing guidelines to promote the validity of the suggestions for best practice offered by such documents. In the next section, the landscape for guidelines in psychological testing is explored. Some notable examples of guidelines developed by organisations in the industry are briefly described to highlight the current scope of guidelines and to identify potential gaps.

Guidelines for Psychological Testing

As part of this study, the researcher performed a focussed national and international online search for guidelines and standards for psychological testing from organisations and associations. A number of guidelines were found, with guidelines most commonly focussing on test fairness, responsible test use, and qualifications for test users. Some of these documents have remained as guidelines, whilst others have become standards after being adopted by a specific national or international body. The most prolific organisation in this regard has been the International Test Commission (ITC), with several national organisations, including the Health Professions Council of South Africa (HPCSA), the British Psychological Society (BPS), and the Psychological Society of Ireland (PSI) either referring readers to the ITC or providing direct online links to the ITC website. The American Educational Research Association (AERA), the American Psychological Association (APA), the National Council on Measurement in Education (NCME), and the Educational Testing Service (ETS) in the United States have also authored documents, with the internationally acclaimed *Standards for educational and psychological testing* (AERA, 2014) being a notable example. Internationally two standards documents (AERA, 2014; ETS, 2014) and one guideline document (ITC, 2015) were found that mentioned test revision. No South African standards or guidelines were found for test revision from the HPCSA, other test organisations or private companies. Laher and Cockcroft (2013) have also noted the scarcity of South African guidelines about all aspects of psychological testing. From the present researcher's perspective, this lack of guidelines is of concern, given the difficult past of psychological testing in South Africa as explored in Chapter Three. In my opinion professional bodies in South Africa, such as the HPCSA, should be more proactive in providing guidelines on psychological testing to its members. Table 4 below reflects a selection of guidelines and

standards found through the online search. The documents have been placed below the broad category headings to which their content relates.

Table 4. Guidelines and standards for psychological testing

ORGANISATION	DOCUMENT TITLE	YEAR	NUMBER OF STANDARDS/GUIDELINES
Guidelines for Test Use			
American Educational Research Association, American Psychological Association, National Council on Measurement in Education	Standards for Educational and Psychological Testing	2014	241
International Test Commission	Guidelines on Test Use	2013	118
Psychological Society of Ireland	Policy on the use of Psychometric Tests in Ireland	2006	118
Test User Qualifications			
American Psychological Association	Test user qualifications	2000	39
Good Testing Practice			
International Test Commission	Guidelines on Quality Control in Scoring, Test Analysis and Reporting of Test Scores	2013	136
International Test Commission	Guidelines on the Security of Tests, Examinations, and Other Assessments	2014	103
Fairness in Testing			
American Psychological Association	Code of Fair Testing Practices in Education	2004	55
Educational Testing Service	Standards for quality and fairness	2014	86
Educational Testing Service	International Principles for Fairness Review of Assessments	2009	7
Mode of Testing			
International Test Commission	Guidelines on Computer-Based and Internet-delivered Testing	2005	249
Test Development and Adaptation			
Educational Testing Service	A Validity Framework for the Use and Development of Exported Assessments	2015	12
International Test Commission	Guidelines for Translating and Adapting Tests	2016	18
Process of Adopting a Revised Test			
International Test Commission	Guidelines on Practitioner Use of Test Revisions, Obsolete Tests, and Test Disposal	2015	27

From Table 4 it appears that guidelines and standards have covered an array of topics in psychological testing, including test use, test user qualifications, good testing practice, fairness, modes of testing, development and adaptation of tests for different countries, and managing the process of switching from a previous version of a test to a newer revised edition. National bodies tended to remain focussed on general guidelines for test use and fair testing practices. The ITC, as an international organisation, ventured into technical aspects such as guidelines for internet and computer-based testing (2005) that was released before this mode of testing gained popularity. The ITC also addressed the question of managing the transition between previous and revised versions of tests (2015) at a time when test revisions are becoming more commonplace, and with revisions being released with increased regularity. The ITC guidelines do not refer however to how to go about revising a test. The present study is therefore the first to offer procedural guidelines for practitioners conducting test revisions. The guidelines by the ITC (2015) on the use of test revision was relevant however for the present study and were adapted and included by the researcher in the guidelines developed in the present study.

In general, the guidelines and standards found related more to offering insights on best practice for test users, than on the task and process of developing or revising tests. Two notable exceptions were found. One is from the ETS (2015) who considered the practice of exporting tests to contexts for which they were not originally developed. Even though this is not necessarily directly related to test revision, the question of whom the target market for a test should be forms part of any test revision process. Another is the ITC Guidelines for translating and adapting tests (2005). This again is not closely aligned to test revision, as the purpose of translation and adaptation is to create a test of identical difficulty as the source test. The aspect that would be important here, however, relates to considerations about migration of tests across

cultures and languages, and the need to consider fairness on an international scale when revising a test.

No international or national guideline documents with a sole focus on test revision were found in the online searches. Despite this, some standards specifically mention test revision. These will be explored in the next subsection.

Standards for test revision.

As mentioned earlier, guidelines become standards when they are adopted by a professional body. Two standards documents and one guideline document specifically mention test revision. These are the *Standards for educational and psychological testing* from the AERA, APA, and NCME (AERA, 2014) and the *Standards for quality and fairness* (ETS, 2014). The ITC *Guidelines on practitioner use of test revisions, obsolete tests, and test disposal* (2015) offers the most extensive guidelines on test revision from the perspective of test users. The latter document highlights the importance of considering test users during test revision, as they would be the target market for a revised test. These three documents are explored below.

The *Standards for educational and psychological testing* (AERA, 2014) document isolates two standards specifically for test revision. The first is Standard 4.24, quoted in Chapter Three of the present thesis that relates to the amendment of test specifications when new research or conditions for use may affect test validity, and the active role expected of publishers to monitor, revise or withdraw a test (AERA, 2014). This is a useful standard as it provides clear direction to revision teams on when a test should be revised. The second Standard (4.25) places an obligation on publishers to inform test users of changes to a test, the impact such changes may have on the score scale, and the comparability of scores between the original and revised test versions. Again, Standard 4.25 provides sound advice on the necessity for revision teams to alert test users

on major changes to score scales between previous and revised test versions that would affect the comparability of scores. This is the type of information required by test users to prevent mistakes in scoring and interpretation of test results on revised tests. Standard 4.25 additionally strives to protect test users from economic exploitation by qualifying that the term 'revised', may only be used if "the test specifications have been updated in significant ways" (AERA, 2014, p. 93). Whilst the intention of this statement is important, it is not clear enough to allow for consistent interpretation by revision teams. It would have been better to state that this phrase would apply particularly to 'medium' or 'extensive' revisions.

In addition to the two test revision-specific criteria, the *Standards for educational and psychological testing* mentions additional standards that apply to revised tests (AERA, 2014). According to Standard 5.20, the changes to test specifications should be identified, and it should be acknowledged that scores between different versions might not be equivalent, even if statistical linking methods have been performed to equate test scores. If the test specifications have changed significantly, publishers should further be mindful of how scores are reported. They should consider creating a new scale reporting method to avoid confusion between previous and revised test versions. Standard 7.14 also addresses the potential confusion between different test versions and offers the advice that affected materials be adequately updated, whilst documentation should note the date of their publication, as well as for which edition(s) of a test they are relevant (AERA, 2014). This constitutes a total of four standards, none of which focus on the process of revising a test, from the AERA, APA, and NCME.

The *Standards for quality and fairness* (ETS, 2014) document mentions test revision in six standards. Standard 3.2 states:

> Document and follow procedures designed to establish and maintain the technical quality, utility, and fairness of the product or service. For new products or services or for major revisions of existing ones, provide and follow a plan for establishing quality and fairness (ETS, 2014, p. 12).

This practice speaks of a considered, planned, and documented approach to both promoting and confirming the quality and fairness of a revised test. Accurately reporting trait-ability without interference from nuisance variables is a focus of psychological testing (Foxcroft & Roodt, 2013). This seemingly simple aim can become complicated when test items unwittingly tap the influence of unintended variables, such as gender and culture, thereby affecting the accuracy of reported scores. The measures taken by test developers to minimise measurement error need to be well documented. Despite much effort, such errors may surface, and their extent should be accurately reported in test manuals to avoid misleading test users. Consideration of measurement error is important, as McCrae (2018) reports that about 40% of variability in test results can be attributed to differences in methods used during scoring and interpretation.

Standard 4.5 similarly encourages developers by stating: “If the intended use of a test has unintended, negative consequences, review the validity evidence to determine whether or not the negative consequences arise from construct irrelevant sources of variance. If they do, revise the test to reduce, to the extent possible, the construct-irrelevant variance” (ETS, 2014, p. 17). This standard highlights two aspects. The first is the need for developers to engage in research on the test, with a specific emphasis on quality, accuracy and fairness. The second relates to correcting errors in measurement through subsequent revisions, and the expectation for revised tests to improve on previous versions. A concerted effort to promote fairness is again emphasised in Standard 5.1 that encourages test developers to provide details of the plan, in its historical,

current, and future format to promote fairness in the test (ETS, 2014). Aspects of this proactive planning and monitoring of a test is also evident in Standard 7.7, which encourages developers to:

> Periodically review test specifications, active items, tests, and ancillary materials to verify that they continue to be appropriate and in compliance with current applicable guidelines. Revise materials as indicated by the reviews. Notify test takers and test users of changes that affect them (ETS, 2014, p. 32).

This standard operationalises the key aspects of a revision, which are to monitor all aspects related to a test, revise less than optimally functioning components, and maintain communication with the test user market. This open and inclusive stance with test users cannot be overemphasised. The professionals that use a psychological test comprise, by definition, a population that has training and experience in the psychological testing subdiscipline, and who develop expertise in the tests they utilise. A revision can capitalise on this expertise by obtaining input from test users. Such an act of consideration can stand test publishers in good stead, as registered test users are an accessible and existing market for the test brand, who may embrace a revision, with minimal sales and marketing.

Standard 3.3 encourages publishers and revision teams to "obtain substantive advice and reviews from diverse internal and external sources, including clients and users, as appropriate. Evaluate the product or service at reasonable intervals. Make revisions and improvements as appropriate" (ETS, 2014, p.12). This suggested practice highlights the importance of regular engagement with multiple sources of information and role players. This is a call to action for publishers, to be proactive in monitoring feedback about a test, and to make revisions when necessary.

In Standard 7.5 the ETS (2014) explores the technical aspects of pretesting items as part of test revision. In particular, it encourages developers to pretest items sufficiently, including further pretesting of revised items because of findings from an initial pretest. This may sound like common knowledge of good test construction practice but, taking into account that guidelines and standards are written for an audience that is seeking expert advice, it is encouraging that the ETS steers practitioners back to sound basic principles. The building blocks of psychological testing are also its foundations and they should be adhered to, especially when the pressures of time and resources can become a compelling argument for not paying due diligence to the entire process.

The ITC mentions test revision several times in the *Guidelines on practitioner use of test revisions, obsolete tests, and test disposal* (ITC, 2015). The guidelines, in brief, address three areas, namely the relationship between test publisher and test users, the communications from test publishers to test users, and the responsibilities of test users in relation to revised tests. The ITC refers to a reciprocal relationship between test publishers and test users (2015, Guideline 2.1). According to Guideline 3.1, test publishers should consider the economic concerns of existing users when determining the price for a revised test (ITC, 2015). This is particularly important for users from developing countries who may have limited funds for new test purchases. Guideline 3.1 further emphasises the economic concerns of test users to professional bodies who may demand that professionals use the latest version of a test.

Concerning communication, Guideline 2.4 requires test publishers to provide an adequate motivation as to why a revised version of a test is being released (ITC, 2015). Such motivation may include changes in the profession or theory, new research, or market influences. A key principle in test revision is that the revised test should improve on a previous edition and create a

better understanding of the measured construct, for the ultimate benefit of the client (ITC, 2015, Guideline 2.2). Test developers should therefore communicate adequate evidence on various forms of test reliability and validity for a revised test, as well as the relationship between scores of new and previous test editions (ITC, 2015, Guideline, 4.6).

The ITC affords more responsibilities to test users regarding adopting revised tests, but it also provides guidance that is more explicit. Guideline 2.5 acknowledges the role of the test practitioner in deciding whether to use an old or new version of a test (ITC, 2015). This decision should be reached after considering the best interests of each client, and after a thorough review of the research evidence, domain descriptors, standardisation information, and normative information for both old and revised versions of a test. Studies of clinical populations on a revised test tend to emerge after it is launched and, in some cases, a previous version may be preferable for a specific test taker if relevant research for the test taker's peer group is unavailable on the revised test. That being said, the ITC prohibits personal attachment to a previous test version as an acceptable reason for not utilising a revised version (ITC, 2015, Guideline 2.9). Finally, test users are advised to actively engage with publishers and pursue training and accreditation on a revised test to ensure that their assessment skills remain current (ITC, 2015, Guideline 2.7).

The above standards and guidelines reflect the published works specific to test revision found from professional bodies. The standards and guidelines covered a range of aspects, but most notably:

- when to revise (AERA, 2014, Standard 4.24; ITC, 2015, Guidelines 2.2 and 2.4);
- to monitor research and feedback on a test (AERA, 2014, Standard 4.24; ETS, 2014, Standard 7.7; ITC, 2015, Guideline 2.4);

- the need to improve issues related to fairness and construct-irrelevant variance with each test revision (ETS, 2014, Standards 3.2, 4.5, and 5.1; ITC, 2015, Guideline 2.2);
- to establish the relationship between old and new versions of a test (AERA, 2014, Standard 5.20; ITC, 2015, Guideline 4.6);
- to implement a cost-effective strategy to roll out a revised test, and avoid change resistance (ITC, 2015, Guidelines 2.7, 2.9 and 3.1);
- the responsibility of test publishers and test users to build and enhance on their reciprocal relationship (ITC, 2015, Guideline 2.1);
- to inform test users of changes to a test, and that scores on different versions may not be equivalent (AERA, 2014, Standards 4.25 and 5.20);
- to obtain input from multiple sources of information, including test users during test revision (ETS, 2014, Standard 3.3);
- to maintain updated records of changes to test materials and to date the changes (AERA, 2014, Standard 7.14); and
- to conceptualise and document plans for test revision, including exercising professional due-diligence through such practices as adequate pretesting of items (ETS, 2014, Standards 3.2, 5.1 and 7.5)

In principle, the guidelines advocate for: open communication and cooperation; being proactive in producing, collating and disseminating information; embracing change for the benefit of clients; honesty about the functions and limitations of a test; and striving for excellence and high standards in psychological testing. Despite the value of the available guidelines, their potential impact may be limited through their inclusion as dispersed comments in larger documents. In some documents test revision does not feature as a prominent aspect and

sometimes appears to have been included as a secondary concern (e.g., AERA, 2014; ETS, 2014). Of the 280 standard and guideline statements reviewed in documents that mention test revision (AERA, 2014; ETS, 2015; ITC, 2015), only 17 (6.1%) standards and guidelines referred specifically to test revision. The abovementioned standards and guidelines also offer little practical insight into how a test revision process should be conducted and managed.

As stated earlier, the AERA, APA and NCME offer four standards for test revision. The absence of more comprehensive work on this topic from these organisations, in particular, is interesting as in 2000 the APA's journal, *Psychological Assessment,* devoted an entire issue to test revision. In this special issue of the journal some articles commented on the lack of formal guidelines and issued a call for more definitive guidance from the psychological testing profession (e.g., Adams, 2000; Silverstein & Nelson, 2000). The journal special edition was largely motivated by the publication of the 1999 edition of the *Standards for educational and psychological testing.* A criticism of the 1999 standards was that they departed from their former structure of labelling guidelines as *primary* (required for all tests before they are used), *secondary* (desired but not required), and *conditional* (applicable in some cases) (Camara, 2007).

This placed the onus on test publishers and users to comply with standards according to their personal judgment, a move that has been criticised as it reduced potentially critical standards to suggestions (Camara, 2007). In defence, however, this reflects a trend towards the reduced legality of guidelines since the reconceptualisation of the term in the 1990s (Field & Lohr, 1990). A reading of the much anticipated 2014 *Standards for educational and psychological testing* indicates that AERA, APA, and NCME has made little or no change in their conceptualisation of the role and importance of test revisions, despite a burgeoning psychological test industry.

Problem Formulation

There appears to be limited guidelines available on psychological test revision, with those guideline documents for psychological tests that mention test revision only referring to this topic in 6.1% of their guidelines (AERA, 2014; ETS, 2015; ITC, 2015). Further, the guidelines in these documents focus more on the use of revised tests than the revision process per se. The guidelines that exist primarily address the relationship between test publishers and users, operational aspects related to the rollout of revised tests, and the roles and obligations of test users. From a practical perspective that needs to consider the complete test revision process, these guidelines are insufficient and fail to address the needs of test users, revision teams, and test publishers. Although these guidelines address, to some extent, the concerns of test users, the foundational principle of guidelines as expert sources of information for those performing a specific act, in this case revising a specific test, remains unmet for those individuals and teams seeking to embark on the process of revising a test. Although comprehensive guidelines would be useful for all professionals in psychological testing, it would be even more valuable for those in developing countries who are seeking to gain access to information and guidance on the practicalities of test revision. It is concerning that as a developing country no such guidelines have been drafted by organisations in South Africa, especially given the destructive legacy of Apartheid on the subdiscipline of psychological testing in the country.

Some issues facing a novice in this field, for instance, would be determining the appropriate moment to embark on a revision, which steps to undertake in the process, and what type of evidence to produce to ease the migration of users from an old to a newly revised version of a test. From the perspective of the present researcher, a more comprehensive approach to test

revision guidelines, which is the unique contribution of the guidelines produced in the present study, should include:

- who should steer the project,
- who should form part of the process,
- what evidence to consider as indicators to determine when to revise a test,
- how fairness should be included in item development,
- the critical concerns of test validity and reliability, and how these can be improved on from the previous test to the revised version,
- the relationship between previous and revised editions of a test, both in terms of statistics and in how the tests are used,
- acknowledgement of the importance of test users as the market of a revised test,
- the role of test users during the test revision process, and subsequent adoption of the revised test, and
- the ongoing responsibility of test publishers to monitor the use of the test, to engage in, and to encourage research on a revised test.

In respect of the above elements, the call over the last two decades for guidelines on the complete process of test revision, from conceptualisation to publication and beyond, appear to have been largely ignored by international psychological testing organisations. This oversight has created a situation where test revisions can differ with regard to due-diligence in the development of psychological tests, thereby resulting in tests of varying quality. Whilst it may be argued that multinational test revision projects backed by industry resources would be able to afford the expertise of a seasoned professional in test revision, many tests have a smaller following, or may not have access to expert resources. This is especially relevant for a

developing country such as South Africa. The original spirit of guidelines would speak to such smaller projects that still seek to deliver a quality product. An industry committed to promoting the accurate and fair measurement of all tests under its domain requires a set of guidelines that can inform test revision processes, and against which completed revision projects may be benchmarked. It would also be beneficial if those who develop guidelines for test revision are familiar with the challenges faced by professionals with limited resources in developing countries. The present researcher experienced the impact of this lack in guidelines in his professional career, both as a test user and as a member of revision teams, and was motivated to develop a comprehensive, clear and detailed set of guidelines for the revision of psychological tests and the use of revised psychological tests. The current study was undertaken therefore to deliver such a set of practical guidelines for practitioners involved in the process of test revision.

Research Aim and Objectives

There is a need for comprehensive test revision guidelines that cover all aspects related to the process of test revision, including the indicators that would highlight the need for revision, the process of test revision, guidance on issues that may arise during the revision process, and how test users should engage with revised tests. This study aimed to fill this vacuum by developing such guidelines, using a structured approach to guide those embarking on test revisions through the process from conceptualisation to completion and post-launch, whilst also offering guidance to the users of revised tests.

The aim of the present study was to develop a comprehensive set of guidelines for test revision that covers the full process of test revision, including the use of revised tests, and to field-test the proposed guidelines using the case example of a revised psychological test, the *Griffiths III*. Thus the purpose of this study was not only to construct guidelines but, by

examining an extant revision, to critique the revision process of the *Griffiths III*, in order to develop a clearer understanding of how the proposed guidelines could operate in practice. The *Griffiths III* was launched in 2016, making it a recent example of a revised test. In addition, the present researcher formed part of a team of South African psychologists who collaborated with an international team on the revision of the *Griffiths III*. This experience afforded me an opportunity to reflect on the collaboration of a professional from a developing country in an international test revision.

The objectives of the study were as follows:

1. To develop guidelines for the revision and use of all types of revised psychological tests.
2. To explore the test revision guidelines with a specific psychological test revision as a case.

Concluding Remarks

In this chapter, the definition and nature of guidelines and the demand for these benchmarks were explored. The concept of guidelines was explored, together with their benefits and criticisms, highlighting shifts in thought about guidelines. The chapter then narrowed its focus on the profession of psychology, and guidelines in the field of psychological testing in particular. The documents from the most prominent international organisations in the field, namely the APA, AERA, NMCE, ETS, and ITS, were discussed. The practice of test revision was then viewed through the lens of guidelines and standards from these organisations, highlighting both the scope as well as the limitations of published documents. The absence of guidelines for test revision in South Africa was highlighted, together with the need for such guidelines to be developed by a practitioner from South Africa who has insight into the legacy of discrimination in the past within South Africa and who has gained experience in test revision through collaboration in an international test revision.

As a final comment, the calls for greater clarity on the complete process of test revision for those engaged in such projects, from developers to test users, was presented. The need for relevant guidelines was emphasised with a reminder of the core purpose of psychological testing, and the concern that every effort should be made for the benefit of the client (Dunbar-Krige et al., 2015). The chapter concluded with the aim and two objectives of the present study. In the next chapter, the research methodology employed in this study to develop sound and comprehensive guidelines for test revision are detailed.

Chapter 5: Research Method

This chapter details the research method utilised for the present study. Given the lack of clarity in psychological testing guideline documents, the researcher developed the design steps used for this study. The specific steps of this study are explored according to each of the two objectives. This is followed by a description of the process of selecting the sample as well as information pertinent to the case study that formed the second phase of the research. The purpose of data analysis, some important options considered during the analysis process, and the steps followed by the researcher are explained. The importance of trustworthiness in qualitative research is reflected on, together with the measures that were taken to enhance the credibility, transferability, dependability and conformability of the research findings. Finally, the ethical considerations that guided the researcher during the study are presented.

Research Method and Design

The focus of the present study was exploratory and descriptive in nature. According to Noor (2008), decisions regarding the research method should be guided by the nature of the research problem. The study employed a mixed method approach focussed on the qualitative methods of systematic reviews and case studies. Mixed methods is a research approach that combines different research methods to achieve the aims and objectives of a study (De Vos et al., 2014). Mixed methods have become popular in the social sciences, including psychology, as the complex nature of some research topics requires investigation through different techniques within a single study (Creswell, 2009).

Initially, mixed method studies mostly incorporated both quantitative and qualitative methods. The reason for this is that quantitative and qualitative approaches are known for different ontological and epistemological approaches to generate knowledge. Ontology refers to

whether the object studied is a stand-alone reality or dependent on social construction and perception. Ontology affects epistemology, which is how topics are studied to generate knowledge (Babbie, 2016; Punch, 2014). Quantitative studies are generally viewed as objective, positivist, and deductive. Qualitative approaches tend to be inductive, interpretive and constructionist (Bryman & Bell, 2014).

By combining the strengths of both approaches, mixed method studies can simultaneously develop and test theory (Laher, Fynn, & Kramer, 2019). More recently, mixed methods have expanded to include combinations of either quantitative or qualitative methods (Laher, Fynn, & Kramer, 2019; Simpson, 2011). The reason for this is that some topics are more suited to quantitative or qualitative studies. A combination of qualitative methods can therefore be combined to crosscheck and refine findings, using their interpretivist stances to enhance the trustworthiness of research (Barnes, 2012). Developing guidelines for the revision of psychological tests, for instance, would involve more qualitative data, as the purpose of the study would be to develop or construct guidelines, and the sources of data would be expert opinion, and therefore more interpretive in nature.

Frost et al (2010) state that, with the preponderance of qualitative studies within the discipline of psychology, increasing numbers of purely qualitative mixed method studies are emerging. Bryman and Bell (2014) identify four types of mixed method designs:

1. Convergent parallel design (both methods are used at the same time and findings are compared or merged);
2. Exploratory sequential design (one method is employed and acts as a preparation for another method);

3. Explanatory sequential design (one method is used after another, with the second method used to explain or elaborate on the findings of the first;
4. Embedded design (both methods are used simultaneously in an integrative way to create a more complete or rounded picture).

In the current study, an explanatory sequential design was used. In the sequence of this explanatory study, the guidelines were developed first using data obtained through a systematic review and the experiences as both a member of test revision teams and a user of revised psychological tests. The guidelines were then refined through peer-review. The second phase in the explanatory sequential design was to field-test the guidelines through an instrumental case study.

According to Bryman and Bell (2014), qualitative approaches are research orientations that emphasise words instead of quantification when collecting and analysing data. Qualitative research also views "social reality as both constantly shifting and emergent, as interpreted by individuals" (Bryman & Bell, 2014, p. 31). Qualitative studies seek rich and deep data that create a valid framework for understanding the topic being researched (Silverman, 2011). A strength of qualitative methods is that they provide greater flexibility to the researcher in exploring the research topic by affording the opportunity to move forwards, backwards, and laterally during the process as the study unfolds (Bryman, 2016). Through the process of uncovering meaning, qualitative approaches can be useful methods in the development of guidelines due to their interpretivist epistemological and constructionist ontological orientations (Creswell, 2009). These orientations fit well with the viewpoint that guidelines are not found but constructed, and similarly developed after an interpretive and reflective process that not only synthesises multiple sources of information, but also in themselves reflect the effort, interpretation and emphasis of

their creators. A guideline development process should go beyond condensing what has already been stated by analysing the similarities and differences in available sources, mediating between different points of view, and creating comprehensive guideline statements to guide practitioners. A guideline development process should also look at extant gaps in the literature and address these with practical guideline statements that describes the procedural link for practitioners. Specific methodologies were used for each of the two objectives and these are described below.

Objective One: Developing guidelines.

In Objective One, guidelines were developed for test revision. The complexities of developing sound and usable guidelines were explored in Chapter Four. The greatest concern that face the development of guidelines relate to their applicability, soundness, the process by which they were developed, and the level of field-testing and peer-review they were exposed to prior to publication (Chilemba, van Wyk, & Leech, 2014; Steyn, 2011). For Objective One, a structured process adapted from the guideline development processes employed in clinical and medical settings was followed, as there appears to be greater procedural clarity and agreement in guideline documents from these fields than from documents within the discipline of psychology in general, and specifically within the subdiscipline of psychological testing. The various guidelines and standards documents discussed from notable national and international organisations in psychology, such as the APA, BPS, and ITC, did not document the *procedure* through which guidelines were developed. From a reader's perspective, this is disconcerting as it requires confidence in the guidelines to stem less from the soundness of the supporting evidence or rigorous process, and more on the reputation of the publishing organisation.

On the topic of evidence, the two main sources of evidence used to develop guidelines were identified in Chapter Four as research outputs and expert advice (Jaeschke, Jankowski,

Brozek, & Antonelli, 2009; Kish, 2001). Objective research evidence is an attractive option for some disciplines, especially when studies are available that meet the stringent requirements of experimental research. For other disciplines, including psychology and its subdiscipline of psychological testing, there is a mixture of qualitative and quantitative research, but few studies that meet the requirements of classical experimental research that includes random sampling, randomised experimental and control groups, double-blind administration and the collection of pre- and post-test data, and strict experimental control of confounding variables (Babbie, 2016). Evidence in psychology also relies on interpretive lenses and integration of different viewpoints that leaves the definitiveness of findings open to disagreement (Creswell, 2009; De Vos et al., 2014).

The review of guidelines in psychological testing, as detailed in Chapter Four, further highlighted the lack of clarity on the steps followed to develop extant guidelines in the subdiscipline. For the current study, the researcher took the viewpoint that a systematic procedure would support the quality of the developed guidelines. Transparency regarding the process followed would also promote the credibility of the guidelines. With this in mind, the researcher tailored the approach proposed by Kish (2001), as discussed in Chapter Four, to include quality checks in the guideline development process, such as systematic review, peer-review and field-testing. The steps constructed and followed by the present researcher to develop the guidelines for test revision and the use of revised psychological tests in this study are outlined in Table 5.

Table 5. Guideline development process

STEP	DESCRIPTION
Consideration of Author and Audience	
1	Information regarding the guideline author
2	Determine the target audience
Systematic Review	
3	Determine the review question or scope of the guideline
4	Perform literature search
5	Critically appraise literature
6	Extract relevant information
7	Synthesise information
Drafting of Guidelines	
8	Construct the framework for the guidelines
9	Write the guidelines
Assessing the Internal and External Validity of Guidelines	
10	Submit the guideline document for peer-review
11	Refine the guideline document based on the peer-review
12	Field-test the guideline document

According to Table 5, the process included consideration of author and audience, a systematic review, the drafting of guidelines, and assessing the internal and external validity of the guidelines as four phases of guideline development. The methods utilised in developing the guidelines is explored below in terms of these four phases.

Consideration of author and audience: In the first developmental step, the author and audience are considered. As guidelines are developed by an author (or authors) and could therefore be shaped by their personal opinions, the issue of who will draft the document and why they felt compelled to do so should be clarified. Similarly, the intended audience should be considered, as this will inform the language and tone of the guidelines, as well as the level of their content (American Academy of Neurology, 2011; Kish, 2001).

Systematic review: The second step of the developmental process consists of a systematic review, which is a structured methodology of sourcing data, extracting information, and

integrating findings (Bryman, 2016; ten Ham-Baloyi & Jordan, 2016). According to Petticrew and Roberts (2006), there is increasing pressure on decision makers to provide evidence that their practice-based decisions are based on the best information available. A key challenge when developing guidelines is how to find and evaluate potential sources of information, how to select sources to inform the guidelines, and how to integrate information from selected sources into a cohesive guideline document. To facilitate these processes, guideline developers have increasingly employed the technique of systematic reviews or meta-analyses (American Academy of Neurology, 2011; Dijkers, 2013; ten Ham-Baloyi & Jordan, 2016; Venter, 2016).

In the past, a traditional literature review has been confused with a systematic review. This has resulted in criticism about the trustworthiness of systematic reviews (Glanville & Lefebvre, 2000). Grbich (2007) asserts there are four considerations when using qualitative data in developmental activities such as creating models or writing guidelines. As systematic reviews use qualitative data, these considerations would be relevant when applying the findings of a systematic review to develop guidelines. The first consideration is the preselected theoretical stance of the researcher that can slant their focus in favour of certain aspects of the data. The second consideration is the methodological underpinnings of the process followed to gather, process, and utilise the data. The third is the influence of the researcher's choice in selecting the conceptual framework that is used to present findings. The fourth consideration is theory minimisation, which implies that the accuracy of the findings can be increased by limiting the extent to which data is filtered through preselected theoretical lenses.

Considering traditional literature reviews in relation to the four considerations of Grbich (2007), the concerns about the scientific validity of literature reviews can be heightened. A traditional literature review details an exploration of literature on a topic by an author that may

favour a preselected theoretical stance. The motive of some authors may also be to construct an argument by referencing only resources that support that viewpoint. Most literature reviews also do not present an overview of the process through which resources were found, vetted, and analysed (De Vos et al., 2014). The main emphasis is on ‘process’, which is what most traditional literature reviews are, as opposed to ‘research method’, which best describes systematic reviews (Dijkers, 2013).

Systematic reviews undertake a comprehensive and systematic search for available literature and follow a rigorous and accountable process to assess, critically appraise, and synthesise information to provide a summative overview of the topic and provide answers to the research question (Gough, Oliver, & Thomas, 2012; Kyriacou & Issitt, 2008; ten Ham-Baloyi & Jordan, 2016). The systematic review has its roots in clinical practice, being developed by Archie Cochrane in the 1970s to further the desire for evidence-based medicine in the face of limited health funding in the United Kingdom (The Cochrane Collaboration, 2012). An explanation of this research method first appeared in 1975 under the term ‘meta-analysis’ that was coined by G. V. Glass (Evidence for Policy and Practice Information Centre, n.d.).

From its roots in clinical care, the systematic review method gained popularity during the 1990s within the social sciences. In 1995, the Evidence for Policy and Practice Information and Coordinating (EPPI) Centre was created in the Social Science Research Unit of the University of London to develop structured literature review methods in the fields of social science and public policy (Evidence for Policy and Practice Information and Coordinating Centre, n.d.). Similarly, the Campbell Collaboration was established in 1999 to adapt the Cochrane meta-analysis methodology for the behavioural, social science and education disciplines to provide high quality

systematic reviews (Campbell Collaboration, n.d.). The EPPI Centre and Cochrane Collaboration have been instrumental in developing and advancing the systematic review method.

Since 2000, systematic reviews have been employed within the social sciences and are extensively used to inform social reform and government policy (Boland, Cherry, & Dickson, 2014). According to Hemingway and Brereton (2009), systematic review is useful in synthesising non-homogenous qualitative data, especially given the structure and rigour of this research method.

The findings of systematic reviews are linked to the sources from which information is extracted (Venter, 2016). This means that if the data were quantitative, the resultant findings would have a greater quantitative slant. Conversely, if the information from sources were qualitative, the findings of the systematic review would have a greater qualitative focus, usually in the form of themes that emerge from the data analysis component of the systematic review. This is consistent with the purpose of a systematic review as a research procedure focussed on condensing or summarising information. The information used in a systematic review is condensed and reflected in integrated form as the findings of a study.

In line with the goals of transparency of this method and replicability of findings, there are discrete steps in the systematic review process (Glasziou et al., 2001; Torgerson, 2003; Venter, 2016). These are:

1. A research question is formulated and a research proposal is developed to provide a conceptual and motivational background for the proposed study. The protocol contains the inclusion and exclusion criteria for the research outputs that will be selected for the study.

2. Once the study is approved, a search is conducted for all research available on the topic. This includes database searches, hand searching of journal and conference proceedings, and studying the reference section of publications to find possible articles for the study.
3. Each research output is evaluated and selected for the study according to the inclusion and exclusion criteria.
4. Each relevant research output is described and classified. A further refinement or selection may be undertaken to limit the number of articles included in the study.
5. The research outputs are processed using a data extraction sheet in order to assess their quality.
6. The extracted data are summarised in order to form a synthesis of the findings.
7. A report is prepared on the findings and the applicability of the information.

According to Dijkers (2013), systematic reviews are favoured in guideline development as they accurately collate, evaluate and synthesise available information. The quality of any proposed guidelines rely on the quality of information from which they are developed, thereby supporting the use of systematic reviews, given the fundamental strengths of this research method. For the present study the researcher supports the usefulness of a systematic review to summarise information in order to draft guidelines. However, the researcher introduces a caveat to this endorsement. Systematic reviews only summarise available information in a way that highlights gaps in the knowledge base. The systematic review method cannot introduce new information or fill those knowledge gaps. To do this the author of the guidelines has to consider the gaps in the findings of the systematic review and go beyond this research method to fill such gaps either through a search of other related information sources or through experience. In this

study, I used my professional experience and knowledge about test revision to identify the gaps, and completed the guidelines using additional information and personal expertise.

Drafting of guidelines: The third aspect of the guideline development process is the drafting of the guidelines. This process is informed by both the preceding systematic review, as well as consideration of the content the author wants to communicate to the audience. Both considerations are important when drafting guidelines. The guidelines should be based on credible information, which is the purpose of the systematic review. This information can appear to be contradictory, abstract or impractical at times. Guidelines are written however by and for people. This is where the expertise of the role of the author becomes crucial in reframing the findings of the systematic review into a cohesive and usable set of guidelines for the intended audience (Joanna Briggs Institute, 2011).

Assessing the internal and external validity of guidelines: The fourth step of guideline development is to assess their internal and external validity. According to Jaeschke, Jankowski, Brozek, and Antonelli (2009), the process of guideline development should include a practical evaluation of guidelines to underscore the internal validity of the development process, specifically in terms of structure and format, and to assess the external validity or generalisability of the guidelines. As guidelines contain an element of personal judgment, investigation into validity can also be viewed as enquiries into the relevance of the guidelines. Validity can also be understood in terms of accuracy or relevance (Mpasa, 2014). This includes the perceived relevance, as judged by the audience, and the practical relevance as determined through implementation.

According to Petticrew and Roberts (2006), perceived relevance can be investigated by forwarding guidelines to expert reviewers. This can be in the form of a single submission to

reviewers, or during an interactive process. An interactive process, such as the Delphi method, involves a process of submitting guidelines to expert reviewers, amending documentation according to feedback, and resubmitting the document to reviewers in several rounds until consensus is reached (Hasson, Keeney & McKenna, 2000).

The practical relevance of guidelines can be explored through implementation in a research context or by establishing the concurrent validity of the guidelines through other means, such as conceptualisation of the guidelines when applied to case studies (Klepac et al., 2012; Steyn, 2011). In the present study, expert reviewers were used to establish and improve the internal validity of the guidelines. A case study of a recently revised psychological test was used to determine the external validity of the guidelines, as discussed in Objective Two below.

Objective Two: Case study.

For Objective Two, the case study method was utilised with the case of the revision of the Griffiths Scales of Child Development – Third Edition *(Griffiths III)* being the application. Objective Two served dual purposes. The first purpose was to explore the selected psychological test using the test revision guidelines as a reference point. The second purpose was to use the findings of this objective to reflect on the effectiveness of the guidelines for test revision in encapsulating the revision process followed for a revised test, to determine the practical adequacy of the guidelines. As such, the intention of the case study was not purely to judge the revision of a test, as it would be unfair to judge a completed project against guidelines that it did not strive to meet at the time of its completion. The intention was rather to demonstrate how the guidelines could be used by test revision teams to steer future revision projects.

The data for a case study on a revised test would be test materials, supplementary materials produced by the revision team or test publisher, conference presentations and journal

publications. As it would be unlikely that test materials produced by a revision team would be overtly negative about their revised test, it should be anticipated that the outcome of a case study could appear more positive. This being stated, a critical analysis of information that is presented in test materials could flag gaps in information that need to be addressed by the revision team. For the present research, the case study analysis was performed by the researcher, but in practice it would be a helpful activity for revision teams to conduct on their own tests, using unpublished internal reports and process documents.

According to Punch (2014), the value of a case study is the increased level of detail with which a subject is studied, taking into account its complexity and context. Punch identifies three types of case studies:

1. The intrinsic case study that aims to better understand a specific case;
2. The instrumental case study where a case is used to provide insight into an issue or to refine a theory; and
3. The collective case study where several case studies are used to provide greater insight into a population or phenomenon.

For the second objective, an instrumental case study was used to provide insight into the guidelines for test revision (see Table 5 on p.117). In the following section, the steps and procedures of this study are explored.

Steps and Procedure

A series of eleven steps were followed to develop and peer-review the guidelines. In step 12 the guidelines were field-tested.

Objective One.

For Objective One, the operationalisation of the first eleven steps are explored below.

Step one: Information regarding the guideline author.

The author of the guidelines is the researcher in the current study. I am a researcher and lecturer in psychological testing, research methodology, and data analysis. I hold a master's degree in research psychology, and have fifteen years' experience in developing, adapting, and revising educational and psychological tests. As part of my continued professional development, I developed an interest in standards and guidelines within the subdiscipline of psychological testing.

I was invited in 2012 by the Association for Research in Infant and Child Development (ARICD) to assist in the revision of the *Griffiths Mental Development Scales* as psychometric researcher. My role was to steer the revision project according to best practices and international guidelines. I found an absence of comprehensive guidelines that could be used by the revision team for the *Griffiths Scales* revision project. I found this disconcerting as it meant that I had to steer the project in this vacuum of silence from international test organisations. I could draw on my decades of professional knowledge and experience, particularly as a former member and employee of the International Test Commission to guide the revision of the *Griffiths Scales*, but developed a growing concern for revision teams that did not have a similar level of knowledge and experience. I felt that a document for test revision could provide guidance on how test users could be engaged in the process of revising a psychological test, and the active role test users should assume when adopting a revised psychological test. This guideline document could bridge the gap between test users and revision teams to unite all role players in a shared purpose to improve a test by revising it. The *GMDS-ER* revision resulted in the publication of the *Griffiths III* in 2016. After the launch of the test, I noticed the challenges the revision team faced beyond the launch of the *Griffiths III*. This intensified my personal motivation to develop a

guideline document for the revision of psychological tests and the use of revised psychological tests.

Step two: Determine the target audience.

The guidelines are intended for practitioners embarking on a test revision or faced with the choice of adopting a revised psychological test (Adams, 2000; Bush, 2010; Butcher, 2000).

Steps three to seven consisted of the systematic review.

Step three: Determine the review question or scope of the guideline.

The scope of the guidelines is specific to the revision of psychological tests and the use of revised tests. As such, the guidelines were purposefully written to address each step of a comprehensive test revision. The scope of the guidelines would consider the needs of test users with respect to engaging in the revision process and subsequent integration of the revised psychological test into their practice. The guidelines would address the information needs of test users, encourage them to be more engaged in the revision process, and guide them on how to review the documentation and components of revised tests, in order to assess the suitability of a revised test for their clients.

Step four: Perform a literature search.

In step four, the sources of evidence required to develop the guidelines and the process of finding relevant sources were determined. The researcher conducted an extensive search for guidelines, standards and other publications on test revision. Limitations in terms of year of publication and language were set for the searches. The year limitation was set at outputs published between 2000 and 2017. The reason for setting the year limit at 2000 is that the APA journal, *Psychological Assessment*, dedicated an issue in 2000 to the practice of test revision, with some authors commenting that there was a lack of clear guidelines for test revision (Adams,

2000; Butcher, 2000; Silverstein & Nelson, 2000; Strauss, Spreen, & Hunter, 2000). Regarding language, only resources published in English were sourced.

Harden (2010) argues that researchers of systematic reviews sometimes have difficulty in reaching convincing conclusions, as the systematic reviews gathered insufficient evidence. In fairness, determination of the sufficiency of evidence is subjective, and ultimately dependent on each reader. This means that most studies, whether quantitative or qualitative, may not escape such criticism. The best a researcher can do is to make every effort to gather a sufficient quantity of data, at the appropriate depth, to meet the aim and objectives of a study. This must also be followed by a comprehensive process of data verification and analysis. Lastly, a researcher must consider alternative explanations for research findings before committing to an explanation that best fits the available information.

The quality of a systematic review depends on the information sourced. For the present study, a layered approach was employed to ensure that efforts were made to find as many resources that met the inclusion criteria. This included:

- Searches on multiple databases including university-based databases and UPECAT, SEALS, and EBSCOhost. Online journal repositories and publisher-linked database searches, including, Sabinet, Sage, Springer, Taylor & Francis, and Wiley. According to Morrison et al (2009), a comprehensive collection process for a systematic review should include searches on at least two databases, so this requirement was met in the present study.
- Internet searches to find organisations either in the subdiscipline of psychological testing (such as the AERA, Association of Test Publishers, ETS, European Test Publishers Group, and the ITC), or professional organisations with a subspecialisation

in psychological testing (such as the APA, Australian Psychological Society, BPS, EFPA, New Zealand Psychologists Board, and the Professional Board of Psychology in South Africa). This was followed by website searches of such organisations for online content that included relevant guidelines, articles, and other publications.

- International and national psychological testing conference proceedings.
- Scanning the references of found outputs for potential additional resources.

For the electronic searches, combinations of key search terms across four categories were used. These terms are included in Table 6. Each search contained one key term from each of the four categories. The main terms are included in Table 6 with the search terms in brackets being used for databases that allowed for truncation.

Table 6: Systematic review search terms

Category One	Category Two	Category Three	Category Four
'psychological' ('psycholog*')	'testing or 'tests' ('test*')	'revision' ('revis*')	'guideline' ('guide*')
'psychometric' ('psychom*')	'measurement' ('measure*')		'standards' ('standard*')
			'policy' or 'policies'

The main limitation of the search process was language, in that searches were only conducted in English, thereby excluding resources not available in English. However, all resources found through the search process were included for consideration in the review process.

The online database searches yielded a number of hits, but as can be seen in Table 7, only seven resources were found to be relevant for the present study.

Table 7. Database search results

Database	Resources Found	Included in Systematic Review
EBScohost	242	4
Findplus	325	1
Sabinet	94	1
Science Direct	275	1
Springer	577	0
Taylor & Francis	248	0
Wiley Online Library	64	0
Total	1825	7

The researcher's internet searches and scanning of references from the found resources yielded a further 14 resources that were used for the systematic review.

Step five: Critically appraise the literature.

The titles, abstracts, and keywords of resources found through the comprehensive search in step four were scanned. Those resources that included comments, suggestions, guidelines, standards or policies for test revision were reviewed in detail by reading the full document. These included resources that used test revision and test adaptation interchangeably, to verify if the content was applicable to test revision.

The quality of a systematic review hinges on the quality of its resources. The researcher considered the quality of each resource, being critical of the extent of peer-review or support by an organisation of each document. For books, standards, guidelines or other documents that were published by an organisation, the researcher checked if the documents stated that they were peer-reviewed. For articles, publication in peer-reviewed journals was a requirement. Only those documents that met the quality standard of peer-review or institutional support were included in the systematic review. In the end, 21 documents met the inclusion and quality criteria and were included for analysis.

Step six: Extract relevant information.

The primacy of evidence was considered in the sixth step to separate information that would be useful for the guidelines from sources that could be discarded. This process involved the management of sources of information. The main points of the included resources were summarised in classification sheets for further analysis. The classification sheets for the 21 documents that were included in the review are presented in Appendix A.

Step seven: Synthesise information.

In step seven, the framework for the guidelines was constructed. This was achieved by extracting themes and subthemes from the content of the classification sheets. Through this process of content analysis of each research output included in the systematic review, a summarising map of the themes (See Appendix B) was developed (Vaismarodi, Turunen, & Bondas, 2013). This step was conducted to extract the thematic threads that commonly occurred within the research outputs.

Through the content analysis, nine major themes were found, with two themes consisting of two subthemes each and one theme consisting of four subthemes. There appeared to be a strong link between the themes and the generic process of test revision that was developed by the researcher, as discussed in Chapter Three.

During steps eight and nine the guidelines were drafted.

Step eight: Construct the framework for the guidelines.

Given the link between the themes from the content analysis component of the systematic review and the generic test revision process, the researcher decided to use the generic process for test revision developed by the researcher in Chapter Three, as the framework to present the guidelines for the guideline document.

Step nine: Write the guidelines.

In writing the guidelines, the researcher utilised the method employed by the International Test Commission (ITC), American Educational Research Association (AERA), American Psychological Association (APA), and Educational Testing Service (ETS) of writing a short guideline statement, followed by a detailed technical paragraph (AERA, 2014; ETS, 2014; ITC, 2015). Whilst published standards and guidelines from organisations were naturally important, due to their institutional support, the researcher ensured that the insights from authors included in non-organisational publications were also reflected in the guidelines. Guidelines from various sources were scattered throughout sources, and there was a lack of a cohesive and comprehensive set of guidelines on test revision within the available literature. The researcher had to identify therefore the gaps within guidelines as well as between individual guideline statements, and construct newly developed guidelines to present a comprehensive and complete guideline document. This means that the development process was not confined to a synthesis of what was found, but the result of analysis, critique and development on the part of the researcher to create a cohesive document. The researcher also drew on his training, research and experience in test use, test development and test revision, both as a test user, lecturer, trainer, researcher, and data analyst.

The researcher learnt valuable lessons about guideline development during the process of authoring the present guidelines. The first was respect for guideline developers, especially sole authors of guidelines. The process can be isolating and may result in a specific slant across guideline statements. The second was the difficulty in drafting guidelines that not only reflected the resource documents found through the systematic review, or the expertise of the researcher, but to develop guidelines that reflected both resource documents as well as professional

experience. As stated earlier, guidelines have traditionally be developed using either evidence or experience, but in the present study the researcher had the challenging task to utilise both.

The researcher used several techniques to counteract unintentional bias. The researcher consulted the themes from the data analysis to steer the guideline statements. The study promoters also provided a valuable sounding board to comment on initial drafts of the guidelines, in order to promote the impartiality of the guidelines. The peer-reviewers also provided valuable input on the consistency of guidance across guidelines and the consistent use of terminology.

During steps 10 and 11 the guidelines were peer-reviewed and refined.

Step 10: Submit the guideline document for peer-review.

To explore the perceived value of the guidelines for test revision, the guideline document was forwarded to practitioners nationally and internationally who had experience in test revision. Their feedback was used to refine the guidelines. A letter of invitation to participate in the study as reviewers of the guideline document was sent to twelve practitioners (See Appendix C). Once informed consent had been obtained, the guideline document was forwarded to participants. The sampling strategy and sample description is explored in the Participants and Sampling section of this chapter.

Step 11: Refine the guideline document based on the peer-review.

Written feedback on the guideline document was obtained from seven participants. This feedback was collated and considered. Changes were made to the placement of guidelines within the document and the content of the guidelines themselves in line with the feedback received. Reviewers that provided detailed feedback were invited to comment on the changes, in the spirit

of the Delphi method, to ensure that they were satisfied with the final draft (Hasson, Keeney & McKenna, 2000). Reviewer feedback is explored in Chapter Six of this thesis.

Step 12 was used to investigate Objective Two of this study.

Objective Two.

Step 12: Field-test the guideline document.

In step 12 the guidelines for test revision were field-tested, using a case example of a recently revised test. The *Griffiths III* was selected for the instrumental case study. The history of the Griffiths tests are explored in the Measures section of this chapter. During this step, information about the revision of the *Griffiths III* was sourced from the test manuals, the websites of the test owner and the test publisher, and publications as well as presentations about the *Griffiths III*. As research and information continue to be published on the *Griffiths III*, the cut-off date for data collection was set at 31 March 2019. This means that findings from the analysis are subject to future change according to new research and publications. The information on the *Griffiths III* was merged to create an overview of the test, using the framework for the guidelines for test revision developed in Objective One of the present study.

The revision of the *Griffiths III* was explored qualitatively through the lens of the test revision guidelines to facilitate comment on both the test as well as the guidelines. This was achieved by establishing those areas where the process of revising the *Griffiths III* were similar to the guidelines, and those points on which it differed. The purpose of the second objective therefore created a dual opportunity to facilitate insight into both the revision of the *Griffiths III* and the proposed guidelines. As the present researcher is South African, and was involved in the revision of the *Griffiths III*, the case study also presented an opportunity to reflect on the participation of a professional from a developing country in an international test revision,

highlighting how South Africa was reflected in the *Griffiths III*, and to consider what could have been done differently.

Participants and Sampling

Due to the technical focus of this study, twelve participants were selected through purposive sampling to provide feedback on the draft guideline document. Participants were selected for their specific knowledge of psychological testing and experience in test revision. To determine their suitability for the present study, they provided input on their work experience, particularly their work environment, their experience in teaching within the subdiscipline of psychological testing, their use of revised tests, and their history of developing or revising psychological tests. They were approached via email through psychology and test industry networks. An email of invitation to participate in the study was forwarded to potential participants (See Appendix C). Informed consent and feedback was received from seven participants, six female and one male, who constituted the expert review panel. The final sample's years of experience in psychological testing ranged from 18 to 47 years, with a median of 35 years. More specifics about the profession and work experience of the participants are provided in Table 8.

From Table 8 it appears that five participants are psychologists, and two are paediatricians. Three work within a university setting, and four within psychological test organisations. Five participants have experience teaching psychological testing, whilst two do not. All seven participants of the expert review panel have experience in using revised psychological tests, as well as developing and/or revising psychological tests.

Table 8. Work experience of participants

Participant	Profession	Institutional Affiliation	Lecturing in Psychological Testing	Using revised tests	Test Development or revision
1	Psychologist	University	No	Yes	Yes
2	Psychologist	University	Yes	Yes	Yes
3	Psychologist	University	No	Yes	Yes
4	Paediatrician	Test Organisation	Yes	Yes	Yes
5	Paediatrician	Test Organisation	Yes	Yes	Yes
6	Psychologist	Test Organisation	Yes	Yes	Yes
7	Psychologist	Test Organisation	Yes	Yes	Yes

Measures

The *Griffiths III* is the most recent revision of the *Griffiths Mental Development Scales* (GMDS), a test of general development for young children. As a developmental measure, the subject area of the *Griffiths Scales* is within the psychology subdiscipline of developmental psychology. The original Griffiths *Baby Scales* was developed by Dr Ruth Griffiths in 1954 for children from birth to two years of age (Griffiths, 1954). The theoretical framework for its development was Dr Griffiths' theory of child development called *Avenues of Learning* (Griffiths, 1935). The *Baby Scales* assessed development in five areas, namely, Locomotor, Personal-Social, Hearing and Speech, Eye and Hand Coordination, and Performance (Griffiths, 1954). The test was designed to provide a holistic assessment of a child's development and to identify early developmental delays.

The test was expanded in the 1960s with the *Extended Griffiths Scales* assessing children from birth to eight years of age. A sixth subscale, Practical Reasoning, was added in this revision (Griffiths, 1970). The *Baby Scales* were revised in 1996 under the leadership of Dr Michael

Huntley, implementing only such changes as required updating of the scales, including re-norming (Huntley, 1996). The test was revised for children from three to eight years and published in 2006 as the *Griffiths Mental Development Scales – Extended Revised*, under the leadership of Prof Delores Luiz. The 2006 revision focussed on broadening the coverage of the domains, replacing weaker items, updating test equipment, and standardising and re-norming the scales for children from years three to eight (Luiz et al., 2006). Both the 1996 and 2006 editions could be viewed as medium revisions, as the revisions focussed on specific year groups, and included item changes and re-norming, without involving changes to the underpinning theoretical constructs of the test.

The full *Griffiths Scales* that included the editions for children from birth to two years (Huntley, 1996) and children aged three to eight years (Luiz et al., 2006), were revised and merged into a single test published in 2016. This revised edition was named the *Griffiths III*, with Prof Louise Stroud as project leader (Stroud et al., 2016). The *Griffiths III* included a number of changes, such as the updating of theory and the replacement of the original six subscales with five new subscales called Foundations of Learning, Language and Communication, Eye and Hand Coordination, Personal-Social-Emotional, and Gross Motor. Because of this more extensive revision of the test, over 70% of the items in the *Griffiths III* are new. The age range for the test was also narrowed to between birth and six years (Stroud et al., 2016).

The *Griffiths III* was chosen as the focus for Objective Two of this study as a number of the test's components had changed with the 2016 revision, thus representing an extensive revision. This flagged the *Griffiths III* as potentially useful for an instrumental case study to facilitate an evaluation of the revision process for this test through the lens of the guidelines that were developed in the first objective of this study.

Data Analysis

The process of obtaining and analysing the data was discussed earlier in steps one to three of Objective One in the Steps and Procedure section. Data analysis is the process of coming to an understanding of the data in order to uncover patterns and themes contained within a dataset (Mouton, 2001). According to Ryan and Bernard (n.d.), identifying themes is a fundamental task of qualitative research. This can be accomplished by finding the underlying meaning and/or the underpinning framework of content within a text. At the heart of the present study was the systematic review, a technique developed to meet the standards of rigour required for scientific studies (Barnard, Cherry & Dickson, 2014).

Monette, Sullivan and De Jong (2011) distinguish between categorising and contextualising as two outcomes of qualitative data analysis. Categorising relates to generating themes and concepts from data to highlight the structure of a dataset. This is an important precursor to developing a framework for reporting findings and, as such, is connected to the purpose of the systematic review method. Contextualising is a focus on the meaning of data and maintaining a holistic interpretation of the data (Monette, Sullivan & De Jong, 2011).

One criticism of systematic reviews is that, depending on how the data is synthesised, more commonly occurring points of data may lead to overrepresentation of such points over less frequently occurring ones (Cronje, 2009). Another critique is that systematic reviews are observational studies, where the researcher determines the scope of the study, what resources to include, and how findings are synthesised (School of Health and Related Research, n.d.). Researchers therefore have to ensure that a systematic review follows the intended rigour of the method, and that a concerted effort is made to obtain a cross-section of research outputs (Cronje,

2009). The findings of a study should reflect the data that in turn needs to be analysed as objectively as possible by the data analysis tools employed by the researcher.

The researcher of the present study was mindful of the need to reflect the structure of the underlying data, as well as themes that were consistent across the outputs. By combining the outcomes of categorising and contextualising of research findings, the researcher sought to reflect not only the structure and content of the underlying data, but also the meaning intended by the authors of the outputs (Donalek & Soldwisch, 2004; Goliath, 2015). As stated above, the researcher went further, however, to identify and address gaps in the guidelines. By merely reflecting the data from the systematic review, the research product would again have perpetuated the lack of information within the available guidelines, and not address the gaps in the historic body of knowledge. The researcher drew therefore on personal and professional experience in psychological testing, research and data analysis, as well as knowledge obtained from existing guidelines in psychological testing from international and national organisations, such as the APA, AERA, BPS, ETC and ITC, to address identified gaps in order to develop a comprehensive set of guidelines.

As part of the systematic review, the data were analysed using the process of thematic analysis as detailed by Braun and Clarke (2006). The purpose of thematic analysis was to identify patterns in the data, to analyse them in order to delineate each pattern into an emerging theme, and lastly to report themes accurately in terms of the breadth and depth intended by the data.

According to Braun and Clarke (2006), their system of thematic analysis was designed to work with or without a pre-existing theoretical framework. The process of analysis can be inductive, i.e. within theory, or deductive, i.e. outside theory. For the present study, a deductive

process was followed. In data analysis, a researcher can also select to focus on meaning at a semantic or latent level. Semantic analysis focuses on the explicit meaning of research data. In latent analysis, the data are interpreted at a level beyond the apparent, on the underlying conceptualisations or ideas being conveyed. In the present study, data analysis included both levels of meaning, in that the guideline statements summarised the content of the underlying data, whilst the explanatory text that accompanied each guideline expressed the conceptual idea that underpins the guideline statement.

In the present research the six steps outlined by Braun and Clarke (2006) for thematic analysis were applied in the following manner:

1. The research outputs were read repeatedly to allow for immersion. Pertinent points were noted in the Classification Sheets (See Appendix A).
2. The Classification Sheets were consulted to generate initial codes.
3. The codes were collapsed into cogent themes, and all data relevant to each theme were collated.
4. The themes were reviewed and refined, and a Summarising Map (See Appendix B) was generated.
5. A final analysis was performed to define each theme in order to name them.
6. The themes were explored in a scholarly manner as part of the present study.

For Objective Two, the manuals, published research, documents and presentations on the *Griffiths III* were consulted as to how they related to the test's revision process. As the revision process of the *Griffiths III* was analysed from the perspective of the guidelines for test revision, the researcher read each research output to highlight the points of similarity and dissimilarity between the guidelines and the revision process of the *Griffiths III.* The researcher rated the level

at which the revision of the *Griffiths III* met each guideline for test revision using a Likert scale with the categories of 'Sufficiently met', 'Partially met', 'Insufficiently met', 'Insufficient information' and 'Not applicable'. The highest possible rating used in this study was "Sufficiently met", meaning that a guideline had been met completely. This is followed by "Partially met" when some aspects of a guideline were met. The lowest rating for this study was "Insufficiently met", when guideline was not met. The 'Insufficient information' rating was applied when the resources provided insufficient information on which to base a rating, whilst 'Not applicable' was used if the revision did not need to address the specific guideline. These ratings were performed to provide a quantifiable overview of the number of guidelines that were met by the revision team of the *Griffiths III* and the extent that the guidelines were met. The findings from this analysis were presented in table format, to create an overview and facilitate comment.

Trustworthiness

According to Neuman (2011), qualitative studies face a greater burden to substantiate the validity and reliability of their findings, as these concepts are traditionally associated with quantitative studies. The value of qualitative studies rests in the internal validity of the rigorousness of the process followed in data collection and interpretation. Lincoln and Guba (1999) proposed a model of trustworthiness to promote the 'truth value' of qualitative research. The four criteria that underpin this model are credibility (rigour of the research process and congruence of findings), transferability (application of findings to related contexts), dependability (consistency of findings), and conformability (neutrality of interpretation and the extent to which findings can be confirmed by others) (Lincoln & Guba, 1999).

According to Oliver and Peersman (2001), the systematic review method was developed to limit systematic bias and random errors. For the present study, additional mechanisms were employed to promote the quality and neutrality of the findings. For instance, the study employed a structured methodology used by international organisations for developing guidelines (American Academic of Neurology, 2011; Dijkers, 2013; ten Ham-Baloyi & Jordan, 2016).

The quality of the findings of a systematic review is dependent on the evidence that is used during the process. According to Tranfield, Denyer and Smart (2003), this starts with a comprehensive search based on keywords and terms. The researcher made a concerted effort to source materials through various means, including database searches, consulting the references of sourced publications, and reviewing the websites of national and international organisations in the discipline of psychology, and the subdiscipline of psychological testing. The second critical aspect of a systematic review is an accurate summation and critical appraisal of the evidence to be included (Aveyard & Sharp, 2011). The researcher utilised Classification Sheets to facilitate the accurate recording of relevant information. The third consideration when performing a systematic review is how best to synthesise the findings in a way that reflects all the information from the data analysis (Gough, Oliver & Thomas, 2012). The researcher used a Summarising Map (Appendix B) to facilitate the integration of information. Schlosser (2007) recommends the use of structured methods of recording and analysing data, such as Classification Sheets and Summarising Maps, to align research with those internationally recognised quality standards that promote the transparency, rigour, and replicability of a study. A lack of such transparent methods can undermine the trustworthiness of research findings. The researcher further assumed a critical and self-reflective stance during each step of the systematic review, whilst remaining mindful of

the potential pitfalls of the process, such as performing tasks with haste, foreclosing on alternative explanations, and personal bias.

The guideline document was further submitted to an independent expert review panel who verified the guidelines in general, whilst also providing constructive feedback (Petticrew & Roberts, 2006). All feedback was considered to refine the guideline document. The transferability of the refined test revision guidelines was subsequently investigated using a case study, as recommended by Jaeschke, Jankowski, Brozek, and Antonelli (2009). The research process was documented to provide clarification of each research step, thereby promoting the dependability and conformability of the research findings. Finally, each step was completed in consultation with the promoters of the study to promote the credibility of the findings.

Ethical and Legal Considerations

The researcher abided by the ethical rules of the Nelson Mandela University. A research proposal was drafted for the present study and submitted to the Department of Psychology, as well as the Health Sciences Faculty Postgraduate Studies Committee of the Nelson Mandela University for approval. The study commenced on receipt of ethical clearance (Ethical clearance number H15-HEA-PSY-012) (See Appendix D).

As the data collected for the systematic review took the form of published documents within the public domain, there were no ethical issues other than to reflect the content of publications accurately and to reference the publications that were utilised.

For expert review of the guideline document, the researcher adhered to the four basic ethical principles of research. These are beneficence, non-maleficence, justice and respect for the autonomy of persons (De Vos et al., 2014; Terre Blanche, Durrheim & Painter, 2006). Only participants who voluntarily agreed to participate in the study and provided informed consent

were included in the sample. Participants were free to withdraw participation without fear or discrimination. Feedback was obtained from reviewers in writing and within a period with which they were comfortable. In terms of anonymity, the research complied with the Protection of Personal Information Act (2013). All feedback from reviewers was treated as confidential and was anonymised by using codes. No participant names were used in the writing of this thesis. All feedback was treated as equally important and played an important part when the guideline document was refined. Participants were also given access to the findings of the study, particularly the final guideline document.

For the case study on the *Griffiths III*, the researcher had a non-maleficent intention and strove to respect the work product of the test revision team. Performing a test revision is sufficiently challenging without fearing criticism based on standards and guidelines that were either not in existence at the time, or not easily identifiable. The review of the *Griffiths III* was not performed therefore to highlight potential failings of the revision project, but was a dual review of the *Griffiths III* revision process and of the guidelines for test revision. This stated, the analysis was performed however in a scientific, honest and unbiased manner, to create an example of analysis for future revision teams to follow.

The researcher stored copies of all data on a password-protected computer, to be kept for five years for verification purposes. The above measures are consistent with international guidelines for research ethics (American Psychological Association, 2017; National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1978).

Concluding Remarks

This chapter explored the research methodology employed to conduct the present study. The process by which the guidelines were developed was presented, including the methods used

to verify the accuracy of the guidelines, specifically the expert review panel and the case instrumental case study. The process of sampling participants for the review panel was detailed, and the history of the *Griffiths III* was provided as it pertained to the case study. The procedural steps of the study, including the data analysis, were outlined. The chapter also considered the issues of trustworthiness, and the measures that were taken by the researcher to promote the credibility, transferability, dependability, and conformability of the research findings. Finally, the importance of research ethics, together with the steps that were taken to address these concerns, were explained. In the next chapter, the findings of the study are presented according to the sequence that they emerged during the study.

Chapter 6: Findings and Discussion

Introduction

In this chapter, the findings of the study are presented according to the two research objectives. The findings of each objective are followed by a discussion of specific research findings to explore their connection with the literature reviewed in earlier chapters of this thesis.

Objective One

For objective one, information from various literature sources was synthesised, reflected on, and added to from related resources as well as the researcher's relevant knowledge and experience to develop a set of guidelines for the revision of psychological tests and the use of revised psychological tests. As stated in Chapter Five, the ten generic phases of test revision (see Chapter 3 page 63) were used as a framework to structure the guidelines. Themes were identified from the literature sources to act as indicators for the aspects emphasised in the guidelines. The guidelines were then developed within the ten-phase process, utilising the themes as a framework for emphasis. Table 9 reflects the documents reviewed and included as primary sources for the development of the guideline statements. These are cross-referenced with the phases they contributed to in the drafting of the guidelines for test revision.

Test revision themes

The researcher derived themes from the documents listed in Table 9. The purpose of the themes was to reveal the recurring messages contained across documents, in order to ensure that these messages would be reflected in the test revision guidelines produced. Nine major themes and eight subthemes were identified, as presented in Table 10. This table is followed by an explanation of the content of each theme and subtheme.

Table 9. Contribution of primary sources to the themes

Document Number	Document	Phase									
		1	2	3	4	5	6	7	8	9	10
1	Adams, K. M. (2000). *Practical and ethical issues pertaining to test revisions.*	X	X	X	X					X	X
2	American Educational Research Association. (2014). *Standards for educational and psychological testing.*	X	X		X		X		X	X	X
3	Bush, S. S. (2010). *Determining whether or when to adopt new versions of psychological and neuropsychological tests: Ethical and professional considerations.*	X							X	X	X
4	Butcher, J. N. (2000). *Revising psychological tests: Lessons learned from the revisions of the MMPI.*	X	X	X			X	X	X	X	X
5	Camara, W. J. (2007). *Standards for educational and psychological testing: Influence in assessment development and use.*		X	X	X		X		X		
6	European Federation of Psychologists' Associations. (2013a). *EFPA review model for the description and evaluation of psychological and educational tests: Test review form and notes for reviewers (Version 4.2.6).*		X				X	X	X	X	
7	European Federation of Psychologists' Associations (EFPA). (2013b). *Performance requirements, context definitions and knowledge & skill specifications for the three EFPA levels of qualifications in psychological assessment.*									X	X
8	Education Testing Service (ETS). (2014). *ETS standards for quality and fairness.*	X	X	X	X		X				X
9	Education Testing Service (ETS). (2009). *ETS international principles for fairness review of assessments: A manual for developing locally appropriate fairness review guidelines in various countries.*	X		X	X	X					
10	Foxcroft, C. D. (2004). *Planning a psychological test in the multicultural South African context.*	X			X	X					
11	Geisinger, K. F. (Ed.). (2013). *APA handbook of testing and assessment in school psychology and education (Volume 3).*				X	X		X			
12	International Test Commission. (2016). *ITC Guidelines for translating and*	X			X	X	X	X		X	

	adapting tests (2nd Ed.).										
13	International Test Commission. (2015). *Guidelines for practitioner use of test revisions, obsolete tests, and test disposal.*								X	X	X
14	International Test Commission. (2013a). *ITC guidelines on test use.*				X	X		X	X	X	X
15	International Test Commission. (2013b). *ITC Guidelines on quality control in scoring, test analysis, and reporting of test scores.*	X	X	X			X			X	X
16	King, M. C. (2006). *Adopting revised versions of psychological tests.*		X								X
17	Liu, J., & Dorans, N. J. (2013). *Assessing a critical aspect of construct continuity when test specifications change or test forms deviate from specifications.*		X		X		X		X	X	
18	Mattern, K. D., Kobrin, J. L., & Camara, W. J. (2012). *Promoting rigorous validation practice: An applied perspective.*								X		X
19	Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). *Psychological Testing on the Internet: New Problems, Old Issues.*		X	X					X	X	X
20	Oliveri, M. E., Lawless, R., & Young, J. W. (2015). *A validity framework for the use and development of exported assessments.*				X		X		X		
21	Strauss, E., Spreen, O., & Hunter, M. (2000). *Implications of test revisions for research.*	X	X			X			X		X

Table 10. Summarising map of themes for test revision

Theme		Document Number																				
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	Reasons for revising a test																					
1.1	• Factors internal to the test	X		X					X													X
1.2	• Factors external to the test	X	X		X	X	X		X						X		X	X		X		X
2	Role players in test revision	X	X		X					X	X					X						
3	Revision planning																					
3.1	• Revision scope and process	X			X	X					X					X		X		X		
3.2	• Post-launch activities				X											X			X			
4	Relationship between test editions	X			X							X		X		X		X				X
5	Test item development		X		X		X		X				X								X	
6	Norm development approach		X		X		X					X	X		X						X	
7	Test validity and reliability		X	X	X	X	X						X	X	X							X
8	Fairness of test results across different groups		X			X			X	X	X	X	X		X						X	
9	Test Users																					
9.1	• Information to test users		X	X			X		X					X		X				X		
9.2	• Test user feedback	X			X				X													
9.3	• Test user responsibility		X					X						X	X							
9.4	• Adopting revised tests	X		X	X									X			X					X

Theme one: Reasons for revising a test.

There are many factors that contribute to the need for a test revision. The main factors relate to test components (internal) and the external environment (external) within which a test functions.

Subtheme 1.1: Factors internal to the test.

Multiple reasons related to test components may lead to the need for it to be revised. These include outdated test materials, psychological test properties and normative data that are no longer applicable to test takers, a desire to include new test populations (such as age or language groups), as well as advances in psychological testing and test administration techniques, such as extending to other modes of testing (for instance, online or computerised adaptive testing) (Adams, 2000; Bush, 2010; Strauss, Spreen, & Hunter, 2000). These factors necessitate periodic reviews of tests for appropriateness, including specifications, active items, and materials (ETS, 2014).

Subtheme 1.2: Factors external to the test.

Tests operate in an environment that can affect the need for them to be revised. From the publisher's perspective, this includes economic concerns such as marketability and sales (Adams, 2000). Tests further interface with practitioners, researchers, test classification bodies, professional bodies, and the broader society. For practitioners and researchers, tests are the products of academic enquiry and applied research. Changes in the academic definition of test constructs and new research findings may necessitate a re-evaluation of the underpinning theory, or components of a test that in turn could signal the need for it to be revised (AERA, 2014; King, 2006). The passage of time can affect the accuracy of norms, which is an important consideration for test classification bodies (EFPA, 2013a). Professional bodies may also change the technical

standards for test use by practitioners that can impact on the ability of test users to comply with new guidelines when using certain tests (Camara, 2007; ETS, 2014; Naglieri et al., 2004). The broader society is not static, and changes within test populations, public perceptions of a test, changing schooling and child-rearing practices, as well as changes in consensus of acceptable or desirable levels of behaviour, can result in a need for a revised test (ITC, 2013a; Liu & Dorans, 2013; Strauss, Spreen, & Hunter, 2000). Test publishers should monitor the varying demands from multiple external role players and provide a comprehensive motivation for the revision of a test (AERA, 2014; Butcher, 2000).

Theme two: Role players in test revision.

Psychological tests engage with multiple role players that have an interest in the success of the assessment process. During a test revision, input is required therefore from different sources, including experts in psychological testing, subject specialists, test users, financial managers, sales representatives, statisticians, and test clients (Adams, 2000; AERA, 2014). These stakeholders should represent a multicultural cross-section of the population for whom the test is intended (ETS, 2009; Foxcroft, 2004). Care should be taken however that such sources have relevant skills or sufficient experience with a test to provide meaningful feedback. A core team steers the revision process, and their roles and responsibilities should be delineated and coordinated. Decisions regarding credit for work and final decision-making authority should be settled prior to commencement of the revision (Butcher, 2000; ITC, 2013b).

Theme three: Revision planning.

Test revision is a complex process that requires adequate planning and execution. This extends to the revision process and the post-launch activities.

Subtheme 3.1: Revision scope and process.

The goals and extent of a revision should be carefully delineated to avoid unnecessary detours (Butcher, 2000). Test revision can take a considerable time to complete, and certain deadlines should be established to maintain momentum on the project (Naglieri, et al., 2004). This being said, developers should be prepared to be responsive to the process and to deviate from the explicit blueprint, as a dynamic interplay exists between test components, meaning that a change in one aspect may necessitate unanticipated changes to other components (Foxcroft, 2004). Care should be taken to avoid taking shortcuts as a result of financial or time constraints, as errors can be prevented with a more reasonable schedule and sufficient quality controls (Adams, 2000; Camara, 2007). All steps in the revision process should be documented and reasons recorded for key decisions made during the journey (ITC, 2013b; Liu & Dorans, 2013).

Subtheme 3.2: Post-launch activities.

Test publishers should construct a reasonable changeover process for the transition between old and new test editions (Butcher, 2000). Training should be prioritised to allow test users to adopt a revised test. Test developers should pursue an active research agenda that promotes academic enquiry into a revised test, using a variety of studies on different populations. This may include advising and supporting independent researchers (ITC, 2013b; Mattern, Kobrin, & Camara, 2012).

Theme four: Relationship between test editions.

A revised test forms part of an existing legacy of a test and, as such, the connection between different editions must be established to avoid inappropriate use or misleading results (Butcher, 2000; Strauss, Spreen, & Hunter, 2000). A revised test may include, for instance, some items, assessment formats, and scoring methods of its predecessor (Adams, 2000). One method

of demonstrating the link between test editions would be a score-equating method, specifically aimed at test norms and interpretation of results (ITC, 2015; Liu & Dorans, 2013). This may also be supported by a set of strong anchor items that are common between old and new versions (Geisinger, 2013). Invariably, some changes will occur over time, and the shift needs to be accounted for in test documentation (ITC, 2013b). Test users are advised not to use a revised test until they have familiarised themselves with how it differs to its predecessor. At times, changes may not be obvious, so a careful study of test manuals and communication from test publishers is required (Strauss, Spreen, & Hunter, 2000).

Theme five: Test item development.

Item development should be performed with great care, including sufficient field-testing and piloting of items on multicultural samples that represent the intended population (Butcher, 2000; ETS, 2014). The purpose of extensive pretesting is to conduct small-scale validity, reliability, and cross-cultural fairness studies to assist in the refinement of items (ITC, 2016; Olivieri, Lawless & Young, 2015). The method used to establish the properties of items (such as classical test theory or item response theory) should be thorough and results should be accurately documented (AERA, 2014; EFPA, 2013a).

Theme six: Norm development approach.

When developing norms, the statistical techniques used and the samples tested are primary concerns. Test revision teams should select the most accurate and generalisable normative approach for the test and its intended purpose (Butcher, 2000). The selection of participants, standardised test conditions, size of norm groups, and representativeness of samples for generalisation to the broader population should be described in detail (EFPA, 2013a; ITC, 2016; Olivieri, Lawless & Young, 2015). Attempts should be made to include a sufficient number of

participants from all intended populations in test samples to allow for empirical analysis (AERA, 2014; Geisinger, 2013). This being stated, it may prove to be an unfeasible target in some instances. If, for example, a test is used with special populations, it may not be possible to source an adequate range of participants from each population group to establish individual norms. In such instances, appropriate comparison groups may be used, with quantitative results being supplemented with in-depth clinical case studies (Geisinger, 2013; ITC, 2013a).

Theme seven: Test validity and reliability.

Test developers should provide comprehensive documentation on steps taken during the revision process to promote the validity and reliability of the measure. A range of evidence should be provided on different forms of test validity, including face, content, construct, and criterion validity (AERA, 2014; Butcher, 2000; Camara, 2007; EFPA, 2013a). The exploration of test validity for all intended populations does not cease on publication of the test, but is a continuing journey as studies on certain populations may only become feasible years after publication (Bush, 2010; ITC, 2015). Evidence should include comparison to other similar tests, including empirical quantitative studies that extend beyond the apparent similarity of tests at face value (Strauss, Spreen, & Hunter, 2000). The reliability of a revised test should be explored in various ways, including test-retest, different modes of testing, interrater, split-half reliability, and evidence of measurement error that could affect the accuracy of scores (ITC, 2013a). The findings of such studies should be supported with data analyses that are appropriate to the method (AERA, 2014; ITC, 2016).

Theme eight: Fairness of test results across different groups.

Test developers should provide evidence of methods employed in the revision process to promote the fair interpretation of results for test takers from all intended cultural and language

groups (AERA, 2014). In addition, steps to reduce construct-irrelevant sources of variance in test performance that prevent accurate measurement of test constructs for test takers from certain populations should be reported (Camara, 2007; ETS, 2009). This includes item formats, presentation of items, and required responses (Foxcroft, 2004). Test developers should provide a plan to promote the fair use of their test that includes administration, scoring systems, interpretation and reporting of results, aimed at reducing unintended negative consequences for test takers from certain populations (ETS, 2014). Test developers should demonstrate a commitment to fairness by providing a future plan for different versions of tests that include consideration of the effect of language and culture in producing nuances in psychological constructs. Amongst these considerations is test adaptation for different cultures and translation into various languages, aimed at actively reducing sources of construct-irrelevant variance in test scores (Geisinger, 2013; ITC, 2016). Adaptation would include removing biased items that are culturally loaded, and that contain a secondary level of assessment, such as acculturation, that is outside the scope of the intended construct (Olivieri, Lawless & Young, 2015). As such, the experts involved in translation or adaptation should be familiar with both the source and target languages and cultures of a test, to allow for adaptations that are sufficiently similar or comparable to the original source version (Olivieri, Lawless & Young, 2015). The purpose of fairness in testing is to allow for accurate assessment and meaningful interpretation of test results (ITC, 2013).

Theme nine: Test users.

Test users emerged as a comprehensive theme. A test is viewed as a product that is designed for test users. Test users should therefore be considered by developers and publishers. Similarly, there are expectations that test users should comply with when they utilise tests. The

four subthemes for test users related to information to users, feedback from users, the responsibilities of test users, and adoption of revised tests.

Subtheme 9.1: Information to test users.

Test publishers should communicate openly and actively with test users regarding changes that affect them, and the motivation for such decisions (ETS, 2014). This includes information regarding revised tests, especially adjustments to score scales, underpinning constructs, and comparability of previous and revised test editions (AERA, 2014; Bush, 2010). Test users should be informed of the justification for a revised test, and a description of the scope of the revision (ITC, 2015). If publishers become aware of common errors in the use of a revised test, they should notify test users in writing and, if required, through special meetings dedicated to error prevention (AERA, 2014; ITC, 2013b). Publishers should also update the information contained within their online presence to prevent errors or misconceptions amongst test users (Naglieri et al., 2004).

Subtheme 9.2: Test user feedback.

Publishers should harness the expertise of test users and request feedback and input from them as part of the revision process (Adams, 2000; Butcher, 2000). Test users can also be invited to participate in field-testing, as their feedback and advice would be helpful in refining the test, as well as gauging the likelihood that a revised test will be received positively by the market (Adams, 2000; ETS, 2014).

Subtheme 9.3: Test user responsibility.

Test users should only offer services for which they are qualified (AERA, 2014). In terms of testing, this implies that test users should only use tests they have been trained in and certified as qualified to use (EFPA, 2013b). Test users should verify that the constructs of the tests they

select are relevant for the assessment needs of their clients (ITC, 2015). Test selection should also be based on consideration of its scientific merit and the quality of the test components. This requires a review of test materials through the lenses of legislation and professional standards (ITC, 2103a). To facilitate such decision-making test users should remain current on changes within psychological testing and participate in appropriate training on a revised measure prior to including a product in their test battery (ITC, 2015).

Subtheme 9.4: Adopting revised tests.

Test users need to remain current with the tests they use. This includes budgeting to purchase revised tests and to undergo required training as soon as possible after the launch of a revised test (Adams, 2000; ITC, 2015). Test users should consider their motivation for delaying adoption of a revised test. Invalid reasons include change resistance and personal attachment to a previous edition, delaying tactics to finish old test booklets and protocols, and an unwillingness to be trained on a revised test (ITC, 2015; King, 2006). The industry standard is for test users to switch over to a revised test within one year post-launch, especially if evidence exists of a normative shift over time, such as the Flynn effect, that may affect the accuracy of the previous test edition (Bush, 2010; Butcher, 2000; Strauss, Spreen, & Hunter, 2000). A valid reason for using an older test with clients is where validity or appropriate norms have not been established on the revised edition, but are available for the old test (King, 2006).

Concluding comments from the themes

There appeared to be congruence between the themes and the ten phases of test revision formulated by the researcher. The following emerged as important aspects for test revision:

- to monitor the test and the environment within which it is used;
- to include input from diverse expert sources in the revision;

- to develop a clear framework and plan for the revision;
- the strong advancement of validity, reliability, and fairness concerns;
- communication between test developers and test users; and
- to obtain early buy-in from test users for a revised test.

There was agreement between the reviewed sources on these themes. The themes therefore offered clear direction for the development of guidelines. In the next section the guidelines are presented.

Guidelines for the revision and use of revised psychological tests

Thirty guidelines were developed within the framework of the ten phases of test revision, and according to the themes extracted from the source materials. Table 11 presents an integration of the guidelines within the ten phases. The table is followed by an explanatory paragraph of each guideline. The complete guideline document contains an introduction, body, and concluding remarks (See Appendix E).

Table 11. Test revision guideline statements

	Phase	Guidelines
1	Pre-Planning	1.1 Test revisions should endeavour to improve the quality, utility, accuracy and fairness of a test. 1.2 Revision teams should consist of a mix of internal and external stakeholders of the test.
2	Initial Investigation	2.1 Test publishers are responsible for monitoring the context within which tests operate, including the use of and feedback about tests, and the industry requirements for psychological tests, as this information may inform the decisions of revision teams. 2.2 A test should be revised or withdrawn when new research data, significant changes in the test domain, or altered conditions of test use may affect the validity of test score interpretations. 2.3 During a test revision, feedback should be obtained from diverse internal and external sources, including test users and test takers.
3	Project Planning	3.1 Test developers should provide a plan to address fairness in the design, development, administration, and use of a revised test. 3.2 The rationale, goals, scope, and process of a test revision should be planned, followed and documented. 3.3 Revision teams should consider constraints in terms of time, cost, and resources when designing a test revision.

4	Academic Enquiry	4.1 The conceptualisation and operationalisation of components of revised tests should be reviewed and appropriately revised to minimise construct-irrelevant sources of score variance. 4.2 Revision teams should balance the needs of test users and the domain measured, when deciding on test items and the nature of tasks required from test takers of a revised test. 4.3 Utilising careful analysis, optimally functioning components of a test should be considered for inclusion in a revised test to act as anchor items, or to foster a sense of brand familiarity between different test editions.
5	Item Development	5.1 The development of test items should consider multicultural contexts, and the possibility that revised tests may be used eventually in settings for which they were not initially intended. 5.2 When authoring item content and test instructions, revision teams should anticipate translation of a revised test into other languages in the future.
6	Test Piloting	6.1 Test items and equipment must be piloted sufficiently using samples that represent the intended population for the revised test. 6.2 Revision teams should select a balanced mix of items for a revised test to ensure that all intended underpinning constructs are adequately assessed at various ability levels.
7	Test Standardisation	7.1 Revision teams should give due consideration to the representativeness and size of standardisation samples, in order to develop normative information for a revised test that is applicable to intended test takers. 7.2 Revised tests should be accompanied at launch with adequate norms and standardisation information.
8	Conduct Supporting Research	8.1 Revision teams shall prioritise research into all target populations of a revised test, including clinical and non-clinical samples. 8.2 Multiple methods should be employed to investigate the relationship between previous and revised editions of a test. 8.3 Research should be conducted into the validity and reliability of a revised test.
9	Test Product Assembly and Launch	9.1 The extent of a revision should be communicated in the product description of a test. 9.2 When tests are revised, users should be informed of the changes to the specifications, underlying constructs, and changes to the scoring method. 9.3 Test users should be clearly informed of the comparability and relationship between the previous and revised editions of a test. 9.4 Documentation for revised tests should be amended and dated to keep information for test users current. 9.5 Test publishers should consider the economic circumstances of test users when determining the cost of a revised test.
10	Post-Launch Activities	10.1 Test publishers and users share a joint responsibility to engage with each other regarding revised tests. 10.2 Test publishers should develop a reasonable strategy to assist test users to switch to a revised test edition. 10.3 Test publishers should offer comprehensive training to promote the level of competence with which test users employ revised tests. 10.4 Test users shall guard against resistance to change; keep current with changes to tests; and strive to adopt a revised test as soon as possible, with due consideration for the best interests of their clients. 10.5 Revision teams should develop a comprehensive post-launch research strategy and encourage the dissemination of independent research studies.

Phase one: Pre-planning.

1.1 **Test revisions should endeavour to improve the quality, utility, accuracy and fairness of a test.** During a test revision, revision teams have to take cognisance of preceding versions. This can be both a benefit and a challenge. The benefit is that from previous versions a body of research evidence and feedback from test users as well as the test market in general is available to provide insight into areas of the test that may be improved on. This includes expansion in the use of the test beyond the original test taker market, and the need to consider new markets in the content, materials, and standardisation sample and norms of the revised test.

There are different aspects that can be revised in a test. If a considerable period has lapsed between revisions, improvements may include refinements to the underpinning construct of the test, the relevance of stimuli, and normative information (Bush, 2010). Expanded reasons for revision may include an extension of the age range of the test population, broadening of the intended test population in terms of ethnic, cultural or language groups, improved accuracy of the test, and alternative forms of administration, scoring and reporting, including fixed computer-based or internet delivered modes (Strauss, Spreen, & Hunter, 2000).

One challenge in the existence of a previous test version is that it creates an immediate benchmark against which the revised version will be measured. Revision teams need to take cognisance of both the benefit and challenge of having previous test versions. Ultimately, whatever changes are made during a revision must be shown to be a clear improvement on earlier versions of a test. A revised test that fails to demonstrate this advancement will likely be poorly received by the market and consequently be unsuccessful (Butcher, 2000). The primary aim should always be to deliver the highest quality test for the ultimate benefit of test takers.

1.2 Revision teams should consist of a mix of internal and external stakeholders of the test. Test revision should be steered by a dedicated, core project committee, called a revision team. As such, the revision team will reflect the existing academic and economic aspects of a test. The revision team would consist therefore of multiple role players, including experts in the subject matter, the field of psychological testing and statistics, as well as representatives from the test publisher involved in test marketing and financial management (Adams, 2000). It is also important that revision teams reflect the rich mix of subgroups within the test's intended population, including different racial, ethnic, and language groups, as well as gender (ETS, 2009; Foxcroft, 2004). These representatives should have expertise that will allow them to represent the linguistic and cultural differences of the intended test population (ITC, 2016). When aspects of the revision rest on the opinions or decisions of experts, the process of selecting those experts, their relevant areas of expertise and experience, together with their qualifications should be documented (AERA, 2014; ETS, 2014).

A revision team should create a vision and mission for the project at the outset, against which all project decisions can be measured (Butcher, 2000). To maintain forward movement on the project it is also important that the roles and responsibilities of revision team members are agreed on in advance (ITC, 2013b). The revision team should agree at the outset on who has final responsibility for completing different tasks, and who has final authority on project decisions. If decision-making power resides with the revision team, then it should be stipulated by which majority vote decisions will carry. Issues related to credit for different components of the revision, and arrangements regarding financial matters, such as salaries, stipends, or royalties should be finalised before the project commences, as these may become sources of conflict at later stages (Butcher, 2000). Above all, the main criteria for project members are their expertise

and willingness to perform tasks, their commitment to the test, their decision-making ability, and a commitment towards engaging in a collaborative effort intended to promote the welfare of test takers (Butcher, 2000).

Phase two: Initial investigation.

2.1 Test publishers are responsible for monitoring the context within which tests operate, including the use of and feedback about tests, and the industry requirements for psychological tests, as this information may inform the decisions of revision teams. Tests operate in a dynamic and changing environment. Test publishers have a responsibility to monitor changes in test conditions and the use of their test products (AERA, 2014). They should be proactive and maintain a responsive attitude (Liu & Dorans, 2013). This includes a variety of actions. Publishers should familiarise themselves with the professional and technical industry standards that apply to tests (Camara, 2007). Test specifications, items, materials, and publications should be reviewed periodically to ensure that their products meet the required standards (ETS, 2014). Changes to industry standards may require a publisher to revise a test to align it with the updated standards.

If significant test information or content has been published within the public domain, it may challenge test validity, which will require test revision earlier than anticipated. Publishers should protect test security therefore by enforcing their copyright of test materials. As part of this process, the internet must be scanned for complete or partial test components that require professional training and registration, as test use by unauthorised persons may diminish test security and cause harm to test takers (Naglieri et al., 2004).

Publishers should be proactive in seeking feedback from test users and researchers by inviting comments through trade publications and test user forums (Adams, 2000). In the event

that any changes to the use of a test are made, test users should be informed of the changes that affect them (ETS, 2014). This includes the intention to embark on a new revision, as this may affect test users at a future date.

2.2 A test should be revised or withdrawn when new research data, significant changes in the test domain, or altered conditions of test use may affect the validity of test score interpretations. It can be challenging to choose the correct moment to revise a test. Important cues can come from research findings, changes in the domain of the test, and amended conditions to the use of tests that are implemented by external bodies.

An important cue is when critical test components have become outdated (Adams, 2000). A key indicator that this has occurred is changes to the theoretical framework that underpins the test. In addition to this, advances in measurement theory, psychological testing practice, and norm development are also important considerations (King, 2006). Changes in the intended test population over time may also necessitate a change. For some tests, a shift in popular culture may date some test items. For tests that challenge, and that rely on the difficulty of individual items, changes in test performance, such as the Flynn effect, may reduce the overall difficulty of tests. Improved nutrition, health care, child-rearing practices, and education have been mentioned as possible causes for the Flynn effect (Strauss, Spreen, & Hunter, 2000). This is especially pertinent for developing countries. Concern about the effects of time on the validity of interpretations of test results is evident in industry publications. For instance, the EFPA label a test as inadequate if the normative and standardisation information is 20 years or older (2013a). To obtain a rating of ‘Excellent’, such information should be less than a decade old. Requirements such as these are intended to inform test users, as well as publishers who should

remain cognisant of changes to important industry standards and benchmark their products against them.

Furthermore, changes in test taker behaviour and performance over time should be scrutinised, as these would affect the perceived value of the product by potential users, and the face validity of tests by test takers (Liu & Dorans, 2013). Sometimes, a test that has reached its expiration date due to outdated norms may still be used meaningfully in already existing longitudinal research. Such a test may also be preferable for certain clinical groups if it is supported by adequate research. However the proviso is that such a test should not be used for decision-making based on its norms, but rather for its qualitative value as one part of a comprehensive portfolio of evidence (AERA, 2014). Revision teams should define the period for which outdated tests may still be used for these purposes and inform test users.

The above aspects highlight the intricate climate of time, culture, and context within which tests are used. Revision teams should take heed of the impact of different factors to determine the correct moment to embark on a test revision.

2.3 During a test revision, feedback should be obtained from diverse internal and external sources, including test users and test takers. It is important to gather feedback from test users and researchers early in the project regarding changes that are required in the test. This should be a broad consultation of both users of the test as well as those who do not use the test, information from experts in the test's subject matter, and experts in the theory and practice of psychological testing. The purpose of such consultations is to update the test revision team on current knowledge within the subject field, and opinions from the test user market (Butcher, 2000). Requesting input serves multiple functions. Firstly, it recognises and values the experience of test users and makes them feel included in the revision. Secondly, it allows for

identification of latent experts on the test, who may be drawn during later phases of the revision project (ITC, 2013b). Thirdly, it creates a sense of collaboration between the revision committee and test users. Finally, it creates a database of interested users and researchers who may be approached later to review the revised test and to provide feedback on the likely acceptance of the product by the broader market (Adams, 2000; ETS, 2014).

Phase three: Project planning.

3.1 Revision teams should provide a plan to address fairness in the design, development, administration, and use of a revised test. A psychological test is a controlled, standardised observation. Its ultimate goal is to measure a construct or set of constructs accurately and fairly, without any interference from sources that are not integrally linked to the construct(s). Revision teams should consider measures to promote the fairness of their tests. The intended changes of a test revision should include therefore plans to improve fairness and accuracy (ETS, 2009). A suggested starting point would be to document historical actions of previous test versions to address fairness (ETS, 2014). Creating such an overview will set a trajectory for the current test revision, by highlighting strengths and potential gaps in previous test editions. It will similarly provide insights into the specific actions that were more successful in increasing fairness, as well as those that were less helpful. It may also assist in constructing future fairness plans. For a current test revision, the measures taken to improve fairness, validity and reliability, including the analyses used and results thereof should be documented (AERA, 2014). Revision teams are ethically obliged to represent the level of fairness in a revised test accurately, including its potential shortcomings in this regard for specific populations.

3.2 The rationale, goals, scope, and process of a test revision should be planned, followed and documented. Test revision takes considerable planning and effort. Without

appropriate management, it has the potential to veer off course (Butcher, 2000). It is essential to delineate the goals and scope of the project at the outset to act as a compass. Each step in the process should be documented to demonstrate how technical quality has been achieved (ITC, 2013b). The rationale for major decisions about the current test revision should also be explained in detail, as these will be important for existing users of previous versions of the test, as well as for future revisions of the test (ETS, 2014).

3.3 Revision teams should consider constraints in terms of time, cost, and resources when designing a test revision. Test revision can be an expensive process that may take years to complete (Naglieri et al., 2004). Projects are limited however by the extent of available resources. The timeframe and budget for the test revision should be considered early in the project, as these will have implications as to the extent of the project and the quality of the final product (Adams, 2000). The demands of the eventual project may outstrip the resources initially allocated, which can result in errors that could have been avoided (Camara, 2007). It is therefore advisable that revision teams have a realistic concept regarding these resources in order to plan their allocation throughout the test revision project from the outset, and to set aside a contingency fund and additional time for unforeseen problems.

Phase four: Academic enquiry.

4.1 The conceptualisation and operationalisation of components of revised tests should be reviewed and appropriately revised to minimise construct-irrelevant sources of score variance. Tests should strive to measure constructs accurately, without interference from factors that are outside the scope of the construct. The variance in test scores should be linked directly to variance in the assessed construct, and not because of construct-irrelevant sources (Camara, 2007). As such, performance should provide valid evidence of the test construct for test

takers from all populations for whom the test was designed (Oliveri, Lawless & Young, 2015). Revision teams should conduct research to determine the extent of construct-irrelevant interference in test scores, as such interference may affect the recommendations that are based on test scores (ETS, 2014). Different sources of interference in accurate measurement can all undermine a test user's ability to make valid inferences about test takers (Oliveri, Lawless, & Young, 2015).

Culture and language are important considerations in this regard. Specific words or phrases may have different meanings for people from different countries or contexts. Cultural differences may also affect how a construct should best be assessed (ITC, 2016). Test items with high cultural loads require test takers to have specific knowledge of the mainstream culture of a specific country. This can unfairly discriminate against test takers from other countries or cultures that differ from the perceived mainstream as such items may measure acculturation as an irrelevant secondary construct, in addition to the intended primary construct of the test (Oliveri, Lawless, & Young, 2015). It is important that a revised test measures the construct accurately for all intended populations (ITC, 2013a).

Test takers from different backgrounds may be less familiar with certain item formats or modes of testing, which may negatively affect their performance. For instance, in developing countries such as South Africa, test takers from underprivileged communities may not be familiar with computers or tablet-based technologies such as keyboards and touch screens. Using these modes of testing on test takers from these communities would negatively affect their test performance (Foxcroft, 2004). Revision teams should periodically review any sources of construct-irrelevant interference, to identify invalid components that may unfairly prevent test takers of certain groups from demonstrating their abilities in the intended test construct (ETS,

2009). Such components should be revised as far as possible, or removed during a test revision (ETS, 2014).

Revision teams should take cognisance therefore of the effects of language and culture on how constructs are assessed and take proactive steps to counteract these sources of measurement interference in revised tests.

4.2 Revision teams should balance the needs of test users and the domain measured, when deciding on test items and the nature of tasks required from test takers of a revised test. As part of the academic enquiry of a test revision, revision teams should familiarise themselves with the needs of practitioners who use the test. It is important to understand the contexts in which a test is used, as well as for what purpose. The nature of tasks included in a revised test should be informed by the contexts in which the test is utilised, as well as by the test takers. As there may have been changes in the contexts and the tasks test takers may be expected to perform since the launch of the previous version of a test, these changes should be considered in the types of items included in the revised test. These needs would be particularly important if a test is expected to interface with other forms of assessment in formal settings, including hospitals or educational systems, for diagnostic purposes or to track the effectiveness of remediation or intervention strategies (Liu & Dorans, 2013).

4.3 Utilising careful analysis, optimally functioning components of a test should be considered for inclusion in a revised test to act as anchor items, or to foster a sense of brand familiarity between different test editions. Major test revisions face heightened scrutiny from existing test users. The product of a revision that reflects a shift in underpinning constructs, test questions, target populations, as well as scoring or norming methods, can create a sense of disconnect between the revised and previous test versions. Steps to address this potential lack of

connection are to include items from previous versions in a revised test. Such legacy items would need to be selected after expert and statistical vetting as being useful for the revised edition. These items would create an anchor block, which can assist in establishing the link in test difficulty between different versions (Geisinger, 2013). Another strategy is for a revised test to utilise the same system of item formats or scoring as previous versions to minimise errors in administration and scoring by users of the previous version (Adams, 2000). All decisions in this regard would be guided however by expert opinion, user feedback, and statistical analysis.

Phase five: Item development.

5.1 The development of test items should consider multicultural contexts, and the possibility that revised tests may be used eventually in settings for which they were not initially intended. A popular test may be used eventually for applications for which it was not originally designed. This is particularly prevalent in the global environment, where tests have been developed for a specific country but are eventually used in other countries. Revision teams need to be aware of this possibility and develop items that either are applicable for a global audience or easily adapted for other cultures (Foxcroft, 2004). The benefit for revision teams is that they may already have some insight of the global exposure of the previous test, which will inform the test items that are included in the revised test.

Test publishers should periodically review the changes in contexts in which their tests are used, as well as in the test populations, as this may suggest possible aspects to be addressed in test revision (ITC, 2013a). Another trend in psychological testing is the conversion of standard tests to computer-based or online tests. These modes of testing require special consideration and adaptation. The equivalence of traditional and technological versions of a revised test would be improved if revision teams were mindful of such future developments, and if they created test

items from the outset that could be extended to other modes of testing (Strauss, Hunter & Spreen, 2000).

5.2 When authoring item content and test instructions, revision teams should anticipate translation of a revised test into other languages in the future. The issues of culture and language remain amongst the most challenging aspects in the accuracy of test results as they directly affect test content (ITC, 2013a). A popular practice in the test industry is to translate tests into other languages to extend the test user market and for cross-cultural research. Multiple-language tests are not only desirable, but also often necessary to reduce bias and promote accurate and fair testing in international settings (Geisinger, 2013). Translation from the original source language to a new target language without accounting for cultural differences can be an important source of construct-irrelevant interference.

In some cases, it may be impossible to create a direct translation of items, due to non-existent words in the target language. This would necessitate adaptation of items for the target language (ETS, 2009). Revision teams should generate a comprehensive vision for a revised test, utilising feedback from the review of the countries and populations using the previous test version. By knowing the initial countries where a revised test will be used, test instructions and items for a revised test must be authored in such a way as to simplify future translation and adaptation, as these practices can affect the success of multilingual international tests (Oliveri, Lawless, & Young, 2015).

If materials are developed in multiple languages as part of the test revision, these should minimise language bias as a nuisance variable. This would require knowledge of the source and target languages, and experts should be selected who are familiar with the different languages and regional nuances within a single language, as well as knowledgeable of all intended test

populations. Attention must be paid to attaining equivalent difficulty in texts for different languages and populations. Revision teams should provide evidence of the similarity in meaning for all intended populations for a revised test (ITC, 2016; Oliveri, Lawless, & Young, 2015).

Phase six: Test piloting.

6.1 Test items and equipment must be field-tested and piloted sufficiently using samples that represent the intended population for the revised test. Field-testing and piloting of potential test items form an important aspect of test development. In test revision, there is a chance that the final item mix in a test will consist of newly developed items, intact items from the previous version, as well as items from the previous version that have been updated or refined. Revision teams should not rely on assumptions about item content, construct or difficulty as a basis for final item selection and placement in the revised test. All decisions should be informed by field-testing and adequate pilot studies (Butcher, 2000). The purpose of field-testing is to obtain qualitative feedback from test takers and users, which can be utilised to refine items. It also assists in quality control by detecting errors in the administration, content, and scoring of items (Camara, 2007). Piloting is used mainly to collect quantitative data on a pool of potential test items, to allow for item analysis and to assist in the selection of items for the final revised test (ITC, 2016). Data from field-testing and piloting are indispensable in promoting the quality of the final test and, as such, thorough field-testing and piloting should be conducted. Rushing through such testing will invariably lead to spurious results, which will undermine the vision and mission of the test revision project (Camara, 2007).

It is advisable that samples for field-testing and piloting closely resemble the intended test population (AERA, 2014). The sampling process and characteristics of the pilot sample

should be documented with as much care as the final standardisation sample and should be included in manuals of the revised test (ETS, 2014).

The analysis of field-test data is important to the revision process. Analyses will assist in identifying items that may contain cultural or linguistic construct-irrelevance. For tests intended for multiple populations, differential item function (DIF) analyses would provide further insight into how individual items operate for different populations and establish if items perform equally well for all intended test takers (AERA, 2014). Field-test results will inform therefore the selection of final items for a revised test (Oliveri, Lawless, & Young, 2015). As cross-cultural fairness is an important criterion against which tests are judged, the methods used to analyse qualitative and quantitative field-test data should be documented in the test manuals and communicated to role players (ETS, 2014).

6.2 Revision teams should select a balanced mix of items for a revised test to ensure that all intended underpinning constructs are adequately assessed at various ability levels. Selecting appropriate items for inclusion in a revised test is crucial. Consideration should be given to user needs, test length, and coverage of underlying constructs at all intended levels of difficulty. The number of test items will depend on the focus of the test, as screening tests may require fewer items per construct than diagnostic tests. Different tests will require preferences for specific item types or a focus on certain sub-constructs. If these become specific features of a revised test, they should be stipulated in marketing and publication material (Liu & Dorans, 2013). For revised tests that provide broader assessment of a construct, evidence should be provided to prove even coverage of the test construct and its ability to assess the knowledge, skills, and abilities of test takers (Oliveri, Lawless, & Young, 2015). Thoroughness in the item analysis and documentation of results from pilot testing will be indispensable (EFPA, 2013a).

Revision teams should demonstrate a commitment to ongoing item analysis and accurate item selection by performing research periodically, even if with partial data on a revised test, to identify sources of error quickly (ITC, 2013b).

Phase seven: Test standardisation.

7.1 Revision teams should give due consideration to the representativeness and size of standardisation samples, in order to develop normative information for a revised test that is applicable to intended test takers. Test norms are an important consideration during test development and test revision processes. Norms are the interface between client-focussed test performance and the external macro-system that surrounds the test taker and within which they function. Norms assist in transforming potentially meaningless test scores to an objective peer-informed interpretation of test behaviour. It is important therefore for a revision team to design a strategy to develop norms to maximise generalisability and usability, whilst keeping costs within acceptable parameters (Butcher, 2000).

The important questions about norm samples relate to who to include in the sample, and how many test takers are required. The norm sample should consist of participants that are relevant for the intended test populations. Tests that are used for diverse populations require a more complex sampling strategy. In the event that the norm sample cannot consist of sufficient representation from all groups, research should be conducted to demonstrate the equivalence in performance of different groups on a revised test. Test norms may not be used for populations where adequate norms or research evidence of equivalence is lacking (ITC, 2013a, 2016). All information about size, composition, and source of norm groups, including their representativeness, should be provided in test manuals (EFPA, 2013a).

The size of samples used to develop norms for a revised test is an important consideration. Revision teams should familiarise themselves with the guidelines of test classification agencies when planning the test standardisation phase. The EFPA, for instance, is prescriptive in how sample size affects the score and subsequent classification of a test. The organisation distinguishes between traditional norms, which view a norm sample group as individual strata, and the increasingly popular continuous norming approaches that divide sample sets into overlapping subgroups that would allow for a more seamless norming matrix for all groups (EFPA, 2013a). Table 12 reflects the minimum EFPA (2013a) sample requirements for a low-stakes test (which is not a primary source of evidence when making life-changing decisions) and the associated qualitative classification of different samples.

Table 12. Minimum EFPA sample size requirements for low-stakes test classification

Classification	Traditional Norming	Continuous Norming
Inadequate (1)	Below 200	Less than 8 subgroups (maximum group size of 69)
Adequate (2)	200-299	8 subgroups with 70-99 participants each
Good (3)	300-999	8 subgroups with 100-149 participants each
Excellent (4)	1000 and above	8 subgroups with at least 150 participants each

From Table 12 it appears that the EFPA would only consider a traditional norming sample as adequate if there are at least 200 participants in each normed group. Due to the seamless norming approach used in continuous norming using overlapping samples, the minimum sample size of 70 is applied for each normed group (EFPA, 2013a). Table 12 demonstrates the level of scrutiny imposed by test classification bodies. Revision teams should familiarise themselves with relevant criteria from the bodies they would approach for test classification, as a lack of compliance can be expensive and time-consuming to correct retrospectively.

7.2 Revised tests should be accompanied at launch with adequate norms and standardisation information. Revised tests should be published with the relevant documentation and information that would allow test users to determine the suitability of a test for their clients. The standard information required includes evidence to support the norms, and the validity and reliability of the revised test for the intended populations (ITC, 2016).

The main forms of validity are face, content, construct, and criterion-related validity. Face validity is the superficial appearance and presentation of a test for test takers, whilst content validity refers to coverage of the construct by test items, as judged by experts. As such, face and content validity rely on qualitative judgments that are embedded in the test development and revision process. Construct validity assesses if a test measures what it is intended for, or whether there are unintended underlying constructs embedded in the test that impact on accurate measurement. Factor analysis is a popular method for investigating construct validity. It would also be important to investigate differential item functioning for test takers from different language or cultural groups to determine the cohesive functioning of the test and its ability to measure accurately across different test populations (EFPA, 2013a).

Construct validity can also be investigated by comparing performance on different tests that measure a similar construct. Such investigations into concurrent or convergent validity aim to underscore the validity of a revised test by comparing it to an established test with proven validity. A correlation coefficient of 0.6 between test performances would provide adequate evidence of concurrent or convergent validity (EFPA, 2013a). Criterion-related validity can be investigated using postdictive (ability to predict former behaviour), concurrent (ability to predict current behaviour), and predictive (ability to predict future behaviour) studies. In this context, concurrent studies could refer to the ability of a revised test to predict test performance on other

similar tests or on present real-world behaviour. Findings from a range of studies with samples exceeding 200 would be considered excellent (EFPA, 2013a).

Reliability is the evidence of consistency of measurement of a revised test over time, test versions, internal consistency and test administrator. A test-retest correlation coefficient of below 0.6 would be inadequate for a reliable measure, whilst a coefficient exceeding 0.8 would be an excellent result (EFPA, 2013a). A range of coefficients can be used to measure the internal consistency of a test or subtests, including Cronbach's alpha, Kuder-Richardson 20 or 21, Lambda-2.factor analysis (omega or theta), and greatest lower bound estimate. According to EFPA (2013a), coefficients of below 0.7 would be inadequate, and higher than 0.9 would be excellent. Correlation studies between different forms of a revised test or multiple scorer ratings can also offer valuable insights into reliability.

Item-response theory (IRT) may be used to offer insight into item discrimination, item difficulty, and guessing of answers. The use of IRT can extend to developing a theoretical model of the test and estimates of a revised test's ability to measure underlying trait factors. Such studies are largely dependent on adequate sample size, with a suggested minimum guide of 200 participants for a one-factor (discrimination), 500 for a two-factor (discrimination and difficulty), and 700 for a three-factor (discrimination, difficulty, and guessing) studies (EFPA, 2013a). Revision teams should investigate different forms of reliability in a range of studies, using adequate and applicable samples, to form a clear picture of test reliability. An indication should be supplied of error of measurement in a revised test as well as measures implemented by the revision team in the norms and interpretations to overcome such artificial lowering or raising of scores (ITC, 2013a).

Some tests are used to assist in the diagnosis of certain disorders or illnesses, and to monitor the effectiveness of treatment for clients. With the fragmentation of traditional diagnoses into ever-widening and deepening layers, producing norms or research relevant to each category has become unfeasible. Revision teams should therefore provide at least some information in the manuals of revised tests about the scores of test takers from certain clinical groups, compared with matched samples from non-clinical samples. Such information could qualitatively guide test users about potential uses of a revised test for clinical populations until additional research is published (Geisinger, 2013).

Phase eight: Conduct supporting research.

8.1 Revision teams should prioritise research into all target populations of a revised test, including clinical and non-clinical samples. It may take years after publication for research to be conducted with a revised test on clinical populations. This being said, revision teams should identify key populations and conduct research to guide the use of the test for such populations, for inclusion in the test manuals and training materials, together with a communication that such research serves as a starting point for ongoing research on different populations.

Research should draw on samples from various clinical and non-clinical populations, and effort should be made to produce research that will maximise the usability and generalisability of findings (Oliveri, Lawless, & Young, 2015). Revision teams should prioritise such research and inform test users of research results on revised tests as soon as possible. This may include releasing pertinent information early through presentations and bulletins to test users, prior to dissemination of the full results in professional publications (EFPA, 2013a). Research should be based on sufficient data, as it would not be in the best interests of test users to rely on the

perceived similarity of certain populations or levels of test performance (Strauss, Spreen, & Hunter, 2000). Users of revised tests should request research information on clinical populations from test publishers, and consider contributing to such projects if it is within their field of interest or expertise (Bush, 2010).

8.2 Multiple methods should be employed to investigate the relationship between previous and revised editions of a test. It is important for test users to understand how a revised test compares to its predecessors. Failure to do so would lead to misleading results, and result in unintended and inappropriate use of a revised test (Strauss, Spreen, & Hunter, 2000). This information includes a comparison of the validity and reliability of the previous and revised editions, differences in the intended populations, conditions for test use, administration and scoring guidelines, and how norm tables should be used and results interpreted.

As the performance of similar test populations may change over time due to anomalies, such as the Flynn Effect, test users should be informed that test performance on different test editions may not be used interchangeably or be directly equated (Strauss, Spreen, & Hunter, 2000). Revision teams should investigate test performance on different test editions, including score equity assessments aimed at analysing construct continuity and equivalence of scores (Liu & Dorans, 2013). If, during test revision, changes have been made to the underlying constructs of the test, it will restrict comparison of global scores, as the factor structures of the different editions will be dissimilar. One method of providing some insight into performance includes converting scores from different versions to a common comparable scale, such as T-scores, that have been corrected for biographical variables such as age, gender, and culture (Strauss, Spreen, & Hunter, 2000).

8.3 Research should be conducted into the validity and reliability of a revised test.

Revision teams have a responsibility to provide comprehensive evidence of the test validity and reliability of a revised test (Butcher, 2000). This information should include technical documentation that highlights different types of validity and reliability (Camara, 2007). The evaluation of tests by classification organisations requires comprehensive presentation of validity at statistically significant levels, as evidenced by a range of studies (EFPA, 2013a). Professional bodies similarly require test users to base test selection on reviews of validity and reliability, as a minimum best practice requirement (ITC, 2015). Revision teams should not rely exclusively on the validity and reliability evidence of previous editions of a test but should fully investigate these areas on the revised test. This evidence should be supplied in the test manuals of the revised test and be expressed clearly with statistical information appropriate to the methods used (AERA, 2014).

It is of concern that some test users employ tests without adequate training or tests that have been adapted, translated, or revised without adequate supporting research or validation (Naglieri et al., 2004). Test users are advised by registration bodies to avoid using tests that have inadequate or unclear technical documentation, an oversight that can be directly linked to failings on the part of revision teams and publishers (ITC, 2013a). Although revision teams may strive to meet these professional standards, they often have to balance competing demands and tight deadlines. This may result in taking shortcuts in the phases just prior to the launch of a revised test, most notably research on test validity and reliability (Mattern, Kobrin, & Camara, 2012). Common sources of an impoverished validation practice are a lack of staff resources or capacity, monetary or business concerns, and a lack of research data on a revised test.

Various aspects may be of interest to test users. This includes technical information regarding the construct representation of the test (i.e., content validity) (Camara, 2007). Within the competitive test publishing market test users would also take note of how test takers perform on a revised measure as compared to other similar products (i.e., concurrent validity) (Strauss, Spreen, & Hunter, 2000). Test revision teams should note the different forms of validity (including content, construct, concurrent, and criterion validity) and reliability (such as test-retest, split-half, inter-rater, and intra/inter-scale reliability). Research is ever expanding in these fields, but revision teams should focus on tried-and-tested methods that communicate the strengths and weaknesses of a revised test in a clear and unbiased fashion (Mattern, Kobrin, & Camara, 2012).

Phase nine: Test product assembly and launch.

9.1 The extent of a revision should be communicated in the product description of a test.

Butcher (2000) identifies ‘light’, ‘medium’ and ‘extensive’ as three types of test revision. A ‘light’ revision entails changes made mostly to the test manual. Aspects that could fall within this type are minor updates to item wording or editorial changes. A ‘medium’ revision is more intensive and includes changes to or replacing non-performing items, and updating the norms of a test. An ‘extensive’ revision involves a complete reanalysis and reconstruction of the test. This could include re-examining the theoretical foundation of the test and major changes to items or subscales, together with a new set of test instructions. An ‘extensive’ revision would also include new norm data, as well as validity and reliability studies (Butcher, 2000).The term ‘revised’ should only be attached to tests that have been updated in significant ways, such as in ‘medium’ and ‘extensive’ revisions.

If the test has not been changed significantly after a ‘light’ revision, the test should rather be marketed as containing minor changes or updates. The extent of these changes should be clearly communicated to existing and new test users.

9.2 When tests are revised, users should be informed of the changes to the specifications, underlying constructs, and changes to the scoring method. Revision teams should present any changes to a revised test in comprehensive technical documentation. Documentation should also focus on how the revised test differs from its predecessor (ITC, 2016).

The theoretical foundations for updates to constructs should be supplied (EFPA, 2013a). Any differences in target populations, methods of norm development, and the correspondence between norms from previous and revised test editions and their potential impact should be unpacked (ITC, 2015). Differences and similarities in the techniques used to convert raw scores to standardised scores must be explained to avoid confusion amongst test users (ITC, 2013b). Documents should do more than reflect on different editions of a test and should go further in justifying the need for a revised version (ITC, 2015). Emphasis should be placed on evidence regarding how the revised test builds or improves on its predecessor, as it would be unethical to develop a test that cannot at least be held to the standards of its predecessor (Naglieri et al., 2004).

9.3 Test users should be clearly informed of the comparability and relationship between the previous and revised editions of a test. There are many reasons why the ties between the previous and revised editions of a measure should be clearly established. The first is that a revision team may face change resistance from established test users, who may make unfounded claims that the revised test is too different from its predecessor or, more likely, so

similar that the expense to purchase the revised test is unnecessary (Butcher, 2000). The second reason is that test users conduct an assessment based on the construct in question. It would be important for test users to be aware of the comparability of the constructs being measured by previous and revised test versions to assess the relevance of the revised test for the assessment need (EFPA, 2013b). A third motivation is that, despite following explicit blueprints in test revision, changes may occur over time (ITC, 2013b). A revised test may draw on content from its predecessor, but there may also be new test questions. With regard to the latter point, items in some content areas are more difficult to replace than others, which may result in marked differences between previous and revised test editions.

For challenge tests that include items with a range of difficulty, items with difficulty levels at the extreme low and high ends are more difficult to develop, clone or replicate. This may affect the overall difficulty of the revised test, which will affect how its test scores compare to a previous version (Liu & Dorans, 2013). Any changes to the difficulty level of a test between different versions should be clearly explained, as well as the comparability of test scores from different editions (AERA, 2014).

9.4 Documentation for revised tests should be amended and dated to keep information for test users current. Any substantial changes to a test should be reflected in its updated documentation, and with supplementary information to existing test users. This includes general information as well as cautions regarding test use (AERA, 2014). The focus should be on the adequacy of information for test users, including administration guidelines, technical information, and norm supplements (EFPA, 2013a). The main purpose of this information is to enable evaluation of the revised measure for its use on individual test takers, as well as certain populations (ITC, 2013a).

9.5 Test publishers should consider the economic circumstances of test users when determining the cost of a revised test. The cost of revised tests has continued to escalate. Test users have their practices in a variety of settings, which affects the availability of funds to purchase new tests. Those in private practice may need to budget for a revised test, while those in institutional settings may have to apply for funds from their employers. Some test users may be from developing countries, where the availability of finance is lower (Adams, 2000). These financial considerations will influence the rate at which a revised test is adopted (ITC, 2015). Test publishers should consider their test user market and price revised tests accordingly. Test users should also be informed as early as possible what the price range for a revised test would be, to enable them to plan for this expense, and to engage with test publishers.

Phase ten: Post-launch activities.

10.1 Test publishers and users share a joint responsibility to engage with each other regarding revised tests. The quality of the psychological testing services is informed by the relationship between test publishers and test users (ITC, 2015). This requires concerted effort from both stakeholder groups to improve the dialogue concerning psychological tests, including revised test editions (Bush, 2010). One area in which test publishers can improve this relationship is by communicating openly and accurately with test users regarding revised tests. This would include their online presence and the information provided on publisher websites. Test publishers should remember that existing and potential test users consult these online resources, and the relationship can be supported by providing accurate and updated information regarding tests (Naglieri et al., 2004). Test users should connect with the publishers of the tests they utilise and engage with the information provided by test publishers.

10.2 Test publishers should develop a reasonable strategy to assist test users to switch to a revised test edition. Financial considerations play an important role in the speed with which test users will adopt a revised test (ITC, 2015). Test publishers become frustrated by the perceived unwillingness of test users to invest in revised tests (Adams, 2000). Many test users will also adopt a waiting strategy to evaluate new research after the launch of a revised test. This may not necessarily be due to change resistance, but the need to be convinced that a revised test represents a tangible improvement over the previous version (King, 2006). The result is that it can take time for a revised test to gain acceptance in the professional community. Test publishers should assist test users by developing a reasonable strategy to transition to the revised test edition. This may include financial assistance, such as a reduced pre-launch order price. Test publishers need to decide on and advertise an end date of use for the previous version, and remain steadfast in their resolve, whilst assisting test users to adopt the revised test (Butcher, 2000).

10.3 Test publishers should offer comprehensive training to promote the level of competence with which test users employ revised tests. Test users are required to remain current with changes and advances in tests, and to only offer services for which they are qualified (EFPA, 2013b; ITC, 2013a). Test publishers can assist test users to achieve this practice standard by offering training programmes and practical workshops in the months leading up to and subsequent to the publication of a revised test. This will enable test users to adopt a revised instrument faster (Butcher, 2000). In reality, some common mistakes in the use of a revised test will surface. Test publishers should advise users of these common errors in a timely manner through a variety of ways, such as in writing or through special error prevention

meetings. Publishers must document and disseminate information on common errors in the use of a revised test, as well as how to prevent such mistakes (ITC, 2013b).

10.4 Test users should guard against resistance to change, keep current with changes to tests, and strive to adopt a revised test as soon as possible, with due consideration for the best interests of their clients. Test practitioners rely on the psychological tests they employ. Over time, this reliance can become ingrained, which can result in attachment to a specific test edition that is outdated (Butcher, 2000). Attachment to a previous test version is not an acceptable justification for not adopting a revised test (ITC, 2015; King, 2006). Test users must accept responsibility for the tests they use, and the accuracy of the recommendations they make (EFPA, 2013b).

The industry standard is for test users to transition to a revised test within six months to a year post-launch (Bush, 2010). This decision should be informed by the relevance of the test for each test taker and the purposes of the test user. Users should have an unbiased approach to revised tests and review the scientific merits of revised tests before reaching a decision. Despite the cost implications of adopting a revised test, economic considerations should not be the primary basis for decisions about test selection. The merits of the test in facilitating an accurate assessment of a test taker should be the most important criteria (ITC, 2015).

A revised test should be adopted as soon as possible if evidence exists that there has been a shift in norms from the previous test edition (such as a large Flynn effect), or if there have been updates to the conceptualisation or measurement of the test constructs (Strauss, Spreen, & Hunter, 2000). Previous test versions may still be used for research purposes, and for test takers assessed from groups (such as language, culture, age, or specific disability) for whom there is an

absence of appropriate test norms or validity studies on the revised test, but which is available on the previous edition (King, 2006).

Test users should only offer assessment services they are qualified to render (AERA, 2014). This requires that they update their knowledge when switching over to a revised test, by studying the test materials and by undergoing training (EFPA, 2013b). Test users should not assume that the method of test administration, scoring, and interpretation used for a previous version would still apply to a revised edition. Time should be taken to learn how to competently use and interpret a revised test (King, 2006; Strauss, Spreen, & Hunter, 2000). In addition, test users should be committed to lifelong learning by refreshing their knowledge about the tests they employ, through follow-up seminars and experiential training (AERA, 2014). They should also remain current in their knowledge of legislation, policy and psychological testing practice. This includes advice, warnings, and guidelines from their professional bodies and their employers (ITC, 2013a).

10.5 Revision teams should develop a comprehensive post-launch research strategy and encourage the dissemination of independent research studies. At launch, a revised test is accompanied by the research performed during its development. As it is adopted by test users, a revised test is used in many contexts with test takers from different backgrounds. Each test session is unique and provides an opportunity for research and learning. Revision teams should spearhead ongoing research into a revised test, and engage with researchers and test users internationally. They should develop a list for test users and researchers that highlights the evidence required to validate a revised test for use on different populations (Mattern, Kobrin, & Camara, 2012). In addition, revision teams should encourage independent research aimed at replicating the validity and reliability claimed in test materials (ITC, 2013b). Minor deviation

from the claimed statistics is acceptable, but significant differences beyond expected patterns of performance should be noted, published, and researched further. Test users should be open to participating in research studies, and lend their expertise to assist in data collection, providing relevant anonymised test data, and sharing interesting test experiences on a revised test (ITC, 2013a).

In Sum: The 30 Test Development Guidelines

These 30 guidelines provide guidance for stakeholders of test revision, including revision teams, test publishers, and test users. It is worth noting that a third of the guidelines are within the last two phases, given that the extant standards and guidelines on test revision, from notable organisations such as the AERA, APA, ETS, and ITC, provide fewer guidelines for post-launch activities (AERA, 2014; ETS, 2009, 2014; ITC, 2013a, 2013b, 2015, 2016). The relative silence of these organisations on the responsibilities of revision teams and publishers after a test is launched may add to a misconception amongst test users and less experienced revision teams that a revision journey ends with the revised test's launch. The present guidelines highlight however that a revision can be viewed as a precursor to the work that follows the launch. The success of a revised test depends on the effort that goes into the marketing, training and follow-up that occurs after it enters the test market (Geisinger, 2013). At the point of launch, a revised test enters the test user market. Some questions and issues will initially surface in practical daily test sessions between test users and test takers (Silverstein & Nelson, 2000). This will necessitate communication with test publishers and the refinement of some revised test components by revision teams.

The development of the guidelines was guided by the themes that emerged during the preceding thematic analysis. Table 13 displays the coverage of the themes within the guidelines.

Table 13. Summarising map of themes for test revision

Theme	Guideline Number																														
	1.1	1.2	2.1	2.2	2.3	3.1	3.2	3.3	4.1	4.2	4.3	5.1	5.2	6.1	6.2	7.1	7.2	8.1	8.2	8.3	9.1	9.2	9.3	9.4	9.5	10.1	10.2	10.3	10.4	10.5
1. Reasons for revising a test																														
Factors internal to the test				X								X	X	X	X															
Factors external to the test			X						X			X	X																	
2. Role players in test revision		X						X																						
3. Revision planning																														
Revision scope and process		X				X	X	X						X		X				X					X					
Post-launch activities																		X								X	X	X	X	X
4. Relationship between test editions	X										X								X			X	X						
5. Test item development									X		X	X	X		X															
6. Norm development approach				X												X														
7. Test validity and reliability				X		X			X					X	X	X	X			X										X
8. Fairness of test results	X					X			X			X		X		X														
9. Test Users																														
Information to test users			X				X							X			X	X	X	X	X	X	X	X		X		X		
Test user feedback			X		X					X	X															X				X
Test user responsibility																				X						X			X	
Adopting revised tests																									X		X	X	X	

From Table 13, there appears to be good coverage of the themes within the guidelines, with the cross-tabulation indicating 80 cross-ticks between themes and guidelines. Information to test users featured in 13 guidelines, making it the most prevalent theme across the guidelines. Test reliability and validity was the second most prominent, featuring in nine guidelines. This was followed closely by planning of the revision scope and process which is mentioned in eight guidelines. This underscores that, particularly for revised tests, there should be a continuous exchange of information between developers and test users, to promote the overall quality of the revised test as well as the adoption of the revised test (ETS, 2014; ITC, 2015).

Planning the scope and process of a test revision can also be assisted through constant communication between revision teams and test users, as the specific needs of test users can be considered in changes to a revised test, and test users can develop a sense of familiarity with an upcoming test edition even during its development process. Test revision teams should also consider the different uses of a test and the contexts within which a test is employed by test users, as this will affect the specific forms of validity and reliability that would be most appropriate to test users (Bush, 2010; ITC, 2016).

To field-test the relevance of the proposed guidelines, these were investigated with the case study of the *Griffiths III* (Stroud et al., 2016), the 2016 revision of the Griffiths Scales of Child Development. The findings of this investigation are presented in objective two.

Objective Two

The guidelines developed in objective one of this study were investigated further through their application to an instrumental case study utilising the *Griffiths III*. The findings are presented in tables, with explanatory texts below each table. The tables include a rating of the revision process of the *Griffiths III* to highlight the extent to which the test's revision mirrored

the phrasing and spirit of the guidelines for test revision. The reader will notice that work and research on the *Griffiths III* has continued since the launch of the test in 2016, meaning that the ratings are based on the evidence available by April 2019. This means that some of the ratings may improve over time as more research is produced on the *Griffiths III*, but equally may deteriorate if research and work on the test ceases.

Findings and discussion of the revision process of the *Griffiths III*

The revision process of the *Griffiths III* was explored in terms of how it related to the guidelines for test revision. As mentioned earlier, the highest possible rating used in this study was "Sufficiently met", meaning that a guideline had been met. This is followed by "Partially met" when only some aspects of a guideline were met. The lowest rating for this study was "Insufficiently met", when the guideline was not met. The findings of this investigation are presented according to the ten phases of test revision in the tables that follow. Each table is followed by a discussion of the findings from each phase.

Table 14. Exploration of Phase 1: Pre-planning of the Griffiths III revision

Guideline	Findings from the *Griffiths III*	Rating
1.1 Test revisions should endeavour to improve the quality, utility, accuracy and fairness of a test.	- Creation of a seamless scale. - Improvement on diagnostic assessment in clinical, educational, neuropsychological, forensic / psycho-legal and research contexts. - Upgraded equipment to meet industry safety standards.	Sufficiently met
1.2 Revision teams should consist of a mix of internal and external stakeholders of the test.	- International test development team focused on cultural fairness. - Revision team members not culturally diverse. - Inclusion of multidisciplinary advisory teams. - Involvement of test publisher. - Division of labour, assigning credit, and resolution of disagreements unclear.	Insufficiently met

The predecessors of the *Griffiths III* were the Baby Scales for children from birth to two years (Huntley, 1996), and the *Griffiths Mental Development Scales – Extended Revised* (GMDS-ER) for children from years 3-8 (Luiz et al., 2006). These two tests were developed a decade apart and had different scoring and norming procedures that created a disconnect when assessing children who were at a developmental age where an accurate assessment of their abilities required them to complete items from both sets of tests. The revision team decided to merge the age ranges of the two previous tests and create a new test, called the *Griffiths III*. This revision included expanded uses for the test, particularly in the field of neurodevelopment, as well as updated equipment (Stroud et al., 2016).

The intention of the *Griffiths III* revision was in agreement with Bush (2010), in that the revision team intended to improve the quality of the test by addressing aspects that were not covered by previous editions, such as updating the underpinning constructs and creating a seamless test from the two separate previous scales for babies and older children. The revision team also had a similar intention to Strauss, Spreen, and Hunter (2000) in creating a more culturally fair test, particularly by considering universal contexts of child development as well as by removing any culturally loaded test questions that may have affected the global suitability of the *GMDS-ER*. According to Stroud et al (2016) many of the pictures in the *GMDS-ER* were outdated and did not reflect cultural diversity. This meant that specifically non-white children might not relate to the pictures. The *GMDS-ER* has been used in South Africa where the majority of the population is black. It was therefore important that the pictures reflected greater ethnic diversity. Importantly, the revision team appointed a South African psychologist to create new pictures for the *Griffiths III*. The resulting pictures are more reflective of different ethnic groups.

As the revision team of the *Griffiths III* sought to improve and update several aspects of the test, Guideline 1.1 is rated as 'Sufficiently met'.

Although the test is owned by a United Kingdom (UK) charity and Learned Society, the Association for Research in Infant and Child Development (ARICD), the revision team consisted of a multidisciplinary team from the UK, Republic of Ireland, and South Africa. Input was also obtained from international advisors who reflected diverse cultures. It is important to note that, whilst the *Griffiths III* is used in several countries, the revision team who decided on the test's constructs, developed test questions, and made all decisions pertaining to the revision process, constituted only white professionals, with some team members, including the present researcher, from South Africa. The revision also had creative input from the German-based publisher, Hogrefe, who played an important role in sourcing and designing the test equipment and materials (Stroud et al., 2016).

From the test manuals, it is unclear how the revision team operated in terms of division of labour, and whether decisions were made about assigning credit for work. It is also unclear how decisions were made, and how differences of opinion within the revision team were settled. If differences were based on majority vote, the extent of such majority is also not discussed. From a guideline perspective, the revision team did consist of a multi-disciplinary team that included the test publisher (Adams, 2000). The issue of cultural representativeness of the revision team is not addressed by the test manuals, which can lead to questions about the cultural diversity of the team (ETS, 2009; Foxcroft, 2004; ITC, 2016). For these reasons, Guideline 1.2 is 'Insufficiently met' as, on balance, there are too many aspects of this specific guideline that are unmet or not explained clearly.

Table 15. Exploration of Phase 2: Initial investigation of the Griffiths III revision

Guideline	Findings from the *Griffiths III*	Rating
2.1 Test publishers are responsible for monitoring the context within which tests operate, including the use of and feedback about tests, and the industry requirements for psychological tests, as this information may inform the decisions of revision teams.	- Periodic review of a test is considered good practice to ensure its continued ethical soundness. - Other major tests had started using a more global approach to test development. - Exploring alternative ways of assessing children.	Partially met
2.2 A test should be revised or withdrawn when new research data, significant changes in the test domain, or altered conditions of test use may affect the validity of test score interpretations.	- A review of the Griffiths to ensure that it taps relevant domains and constructs.	Partially met
2.3 During a test revision, feedback should be obtained from diverse internal and external sources, including test users and test takers.	- 'Avenues of Learning' workshop; interviews with experts; feedback from practitioners who use the Griffiths, as well as those who do not. - Feedback from administrators of pilot test and standardisation sample.	Partially met

The *Griffiths III* revision team acknowledges that it is good practice to review a test periodically (Stroud et al., 2016). The reasons for revising the *Griffiths III* were to ensure that the test reflected the developmental stages for children in a contemporary context. The test can be used in schools to track the development of children. It is often used to detect developmental delays in children and to track the development of children with special needs. For severely handicapped children developmental progression is sometimes measured less in general terms than in the attainment of specific skills, so the test was designed to facilitate the measurement of such small steps for children in special needs schools (Stroud, 2016). Whilst this is a positive step in the assessment of children, it could be argued that in comparing the development of children with special needs to the theoretical conceptualisation of childhood development, not enough emphasis is placed on the uniqueness of each child. This is further highlighted by the fact

that the standardisation and norm sample of the *Griffiths III* only included normally developing children. This raises questions about the revision team's perception of developmentally delayed or severely handicapped children. It would be important for the revision team to provide guidance on how the test should be used with such children.

The revision team explored the underpinning constructs of the test and found theoretical and research evidence that would affect the construct validity of the test (Stroud et al., 2016). It is unclear however from the test manuals whether a review was performed of the criteria of test accreditation bodies for psychological tests, and how the previous version of the Griffiths Scales would be rated against such criteria. Despite the other theoretical and research reviews performed by the revision team, the impact of professional standards (ETS, 2014), which for tests would include test accreditation bodies (EPFA, 2013a), would be important to sufficiently meet Guidelines 2.1 and 2.2. The test manuals also do not adequately address the revision team's exploration of advances in measurement theory (King, 2006), apart from the application of continuous norming. For this reason, these two guidelines were 'Partially met'.

As part of the revision, the ARICD and Hogrefe reviewed other developmental tests for children to ascertain current trends in test development, norm development, modes of testing, and the uses of tests. This review, conducted during the first of six phases of the revision of the *Griffiths III*, was extended to obtaining feedback on the test and its domains. The feedback process included users of the Griffiths by means of workshops and questionnaires, interviews with experts, and questionnaire responses from non-users of the Griffiths (Stroud, Green, Bloomfield, & McAlinden, 2017). These sources provided the revision team with information to frame the scope and extent of the *Griffiths III*. During the later stages, feedback was obtained from test administrators who collected pilot item information and standardisation data, to assist

with item refinement and selection, and to correct minor issues with item instructions and scoring criteria (Stroud et al., 2016). This broad consultative process aligns with the guidance from Butcher (2000), but one of the elements of feedback is to identify latent experts (ITC, 2013b) who can advise later on the likely acceptance of the test by the user market (Adams, 2000; ETS, 2014). It is unclear whether the revision team performed this component of the guideline therefore Guideline 2.3 was 'Partially met'.

Table 16. Exploration of Phase 3: Project planning of the Griffiths III revision

Guideline	Findings from the *Griffiths III*	Rating
3.1 Revision teams should provide a plan to address fairness in the design, development, administration, and use of a revised test.	- Decision to develop *Griffiths III* as a criterion-referenced test. - International team of item developers and administrators. Pilot data collected in South Africa.	Partially met
3.2 The rationale, goals, scope, and process of a test revision should be planned, followed and documented.	- Six phases of revision for *Griffiths III* planned, and documented within test manuals.	Sufficiently met
3.3 Revision teams should consider constraints in terms of time, cost, and resources when designing a test revision.	- Multidisciplinary team of developers and volunteers. - Financial contribution and infrastructure support from Hogrefe.	Sufficiently met

The revision team decided to revise the *Griffiths III* as a criterion-referenced test, based on benchmarks of child development supported by research internationally (Stroud et al., 2016). This allowed for future projects aimed at developing local norms, whilst limiting the extent of cultural bias of items for children outside the UK and Republic of Ireland. The revision team also consisted of members from the UK, Republic of Ireland, and South Africa. Although the revision team consisted of white, English-speaking professionals, key sources of information for item refinement were the multicultural, multilingual team of South African test administrators that collected the pilot and standardisation data. The test administrators flagged potential sources of

item bias, and suggested culturally reduced alternatives (Stroud et al., 2017). This partially addresses the ETS (2009) comments about implementing a plan to address test fairness, but evidence of the cross-cultural analyses that were performed to establish fairness (AERA, 2014; Foxcroft, 2004) are not presented in the test manuals. Guideline 3.1 has therefore only been 'Partially met'.

The revision team documented all phases of the revision, as well as motivations for key decisions in the test manuals (Stroud et al., 2016), which aligns with advice from the ITC (2013b) and ETS (2014). Initially the revision team anticipated a medium revision, but as the revision phases unfolded, the committee became convinced that an extensive revision would be in the best interests of the test, test users, and children. It was decided that an extensive revision would address the feedback received from test users and researchers about the underpinning constructs, outdated test items and materials, and the cultural fairness of the *GMDS-ER* (Stroud, 2013). According to Butcher (2000), such a decision can become disruptive to the timeframe of the revision project. Adams (2000) also mentions the implications of such changes on resources and finances in particular. The test publisher of the *Griffiths III* however became similarly convinced of the increased scope of the revision and dedicated infrastructure and financial support to ensure that the revision extended to cover the required changes (Stroud et al., 2016). As the documentation regarding the revision steps are documented clearly in the test manuals, and the test publisher provided additional capital and resources to extend the scope of the revision, Guidelines 3.2 and 3.3 are 'Sufficiently met'.

Table 17. Exploration of Phase 4: Academic enquiry of the Griffiths III revision

Guideline	Findings from the *Griffiths III*	Rating
4.1 The conceptualisation and operationalisation of components of revised tests should be reviewed and appropriately revised to minimise construct-irrelevant sources of score variance.	- Reverse reengineering of complete test, including construct domains, test items and instructions, scoring criteria and equipment.	Partially met
4.2 Revision teams should balance the needs of test users and the domain measured when deciding on test items and the nature of tasks required from test takers of a revised test.	- Feedback from test users. - Criterion-referenced test.	Sufficiently met
4.3 Utilising careful analysis, optimally functioning components of a test should be considered for inclusion in a revised test to act as anchor items, or to foster a sense of brand familiarity between different test editions.	- 30% of items from previous editions retained in *Griffiths III*.	Sufficiently met

The academic enquiry of the revision project included a complete reverse reengineering process. All test components were dissected, analysed for relevance, and updated as needed. Following a comprehensive literature search of child development and input from experts on child development theory, the theoretical constructs that underpin the subscales of the test were updated (Stroud, Foxcroft & Marais, 2012). This, in turn, necessitated the development of new test questions for the *Griffiths III* (Stroud, Foxcroft, Cronje, & Marais, 2014). Exploration of the constructs was comprehensive, but research to determine that variance in scores was not due to construct-irrelevant sources (Camara, 2007), such as the influence of culture in international populations (Oliveri, Lawless, & Young, 2015), is not presented in the test manuals. This may invite concern about the accurate measurement of the test constructs for all intended populations.

This is however a broader critique of psychological tests, as cross-cultural equivalence is a theoretical goal that tests can only strive for (ITC, 2013a). On balance, Guideline 4.1 has thus been ‘Partially met’.

The revision team was mindful of the challenges test users face when changing from a test they rely on to a revised version (Stroud, 2013). Feedback was obtained from test users about how they used the Griffiths, what they liked and disliked about it, and the changes they would prefer to be made to the test (Stroud, Foxcroft, & Marais, 2012). This feedback was obtained to meet the needs of existing users with the revision, to make them feel part of the journey, and to connect them with the project. The *Griffiths III* was designed as a criterion-referenced test, with item selection being guided by the underpinning constructs of the test. This allows the test to be used in different settings, and particularly to track the effectiveness of remediation on a child over time (Liu & Dorans, 2013).

Item analysis was performed on items in the *GMDS-ER*, and about 30% of items were retained for the *Griffiths III*, including legacy items, such as form boards, brick boxes, and the original ‘Animal’s Day’ book developed by Dr Ruth Griffiths for the test in the 1950s. Such legacy items were given new prominence within the test to support their status as cornerstones of a *Griffiths III* assessment (Stroud et al., 2016). This is in line with the recommendation by Geisinger (2013) for anchor items across tests that could establish a link between different versions. For the importance placed on test user feedback, the retention of some previous items, and the ultimate importance of underpinning constructs of items, Guidelines 4.2 and 4.3 are ‘Sufficiently met’.

Table 18. Exploration of Phase 5: Item development of the Griffiths III revision

Guideline	Findings from the *Griffiths III*	Rating
5.1 The development of test items should consider multicultural contexts, and the possibility that revised tests may be used eventually in settings for which they were not initially intended.	- Removal of culturally sensitive items. - Inclusion of activities and toys that are common to different countries. - Consideration of future tablet-based version.	Sufficiently met
5.2 When authoring item content and test instructions, revision teams should anticipate translation of a revised test into other languages in the future.	- Consideration of translation of items.	Sufficiently met

The revision team removed a number of culture-specific or outdated items that appeared in the *GMDS-ER*, including the ability to ride a bicycle, get on and off a bus unaided, and go to a shop alone. Items that included stimuli that were more relevant for specific countries or cultures, such as birthday parties or pictures featuring only white children, were also exchanged for material that was more common across different ethnic groups, thereby focussing on universal contexts of child development (Stroud, 2016; Stroud et al., 2016). The South African members of the revision team contributed to item development, and developed the first subscale, Foundations of Learning, of the test, focussing on universal and unique contexts of child development. Foundations of Learning is also the only new subscale in the *Griffiths III*, thereby creating an important platform for the South African team members to showcase their abilities and contribution to the test. The ITC (2013a) similarly recommends that the contexts where tests are used be reviewed periodically, and that changes are made to be more inclusive of different contexts. Guideline 5.1 was thus ‘Sufficiently met’.

The revision team emphasised future expansion of the test, including translation into other languages, and for adaptation to tablet-based testing. The language subscale was extensively

revised to minimise construct-irrelevant variance across cultures and to enable future translation of the *Griffiths III* (Stroud et al., 2014). The revision team encouraged research on the adaptation of the *Griffiths III* for tablet-based testing using gamification and a storyline to encourage the participation of children (Marais, Stroud, Foxcroft, & Cronje, 2017). Such considerations of adaptations to other modes of testing and for other language groups have been highlighted by a number of authors (Foxcroft, 2004; Geisinger, 2013; ITC, 2013a; Oliveri, Lawless, & Young, 2015; Strauss, Spreen, & Hunter, 2000). Guideline 5.2 has therefore also been 'Sufficiently met'.

Table 19. Exploration of Phase 6: Test piloting of the Griffiths III revision

Guideline	Findings from the *Griffiths III*	Rating
6.1 Test items and equipment must be field-tested and piloted sufficiently using samples that represent the intended population for the revised test.	- Field-tested and pilot tested in the UK, Republic of Ireland, and South Africa only.	Insufficiently met
6.2 Revision teams should select a balanced mix of items for a revised test to ensure that all intended underpinning constructs are adequately assessed at various ability levels.	- Item selection based on underpinning constructs and a range of item difficulties.	Sufficiently met

Test items were field-tested repeatedly after each refinement in the UK and Republic of Ireland. The mix of viable items was pilot tested on South African children from different ethnic and socioeconomic groups. The item performance of children from these diverse populations informed the selection of items for the *Griffiths III* (Stroud et al., 2016). For an international test however, it could be argued that pilot testing should have occurred in other regions that represent some of the markets of the Griffiths, such as the European mainland, South America, Middle East, Oceania and Asia. The AERA (2014) recommend that pilot test samples should closely resemble the intended test population, whilst the ETS (2014) state that information about the

performance of pilot samples be explored in detail in the test manuals. It appears that the revision team of the *Griffiths III* did not perform test piloting to this standard and for this reason Guideline 6.1 was 'Insufficiently met'.

Feedback from users of the *GMDS-ER* was that the test, with 504 items, took too long to administer (Samuel, 2014). This was, in part, because the *GMDS-ER* consisted of two components, the *Baby Scales* for years one and two, and the *Extended Griffiths Scales* for children for years three to eight. There were 429 items piloted for the *Griffiths III*, and 321 were selected for inclusion in the final test (Cronje, 2016). This means that the *Griffiths III* has 36% fewer items than the *GMDS-ER*. In fairness, the age ceiling of the test was also dropped from eight years for the *GMDS-ER* to six years for the *Griffiths III*, which would have reduced the number of items. This reduction in number of test items concurs with Liu and Dorans (2013) that test length depends on the purpose of a test, but that revision teams should consider test clients and practitioners, and guard against producing a test that over-assesses a construct. The *Griffiths III* revision team selected items that covered the theoretical constructs at different ability and age levels (Stroud et al., 2016). In the test manual the underpinning construct for each test question is provided, enabling test users to determine whether test takers have a particular strength or weakness in specific constructs. This aligns with the recommendation of Oliveri, Lawless and Young (2015) that test manuals detail the extent to which test questions saturate the underpinning construct. Guideline 6.2 was thus 'Sufficiently met'.

Table 20. Exploration of Phase 7: Test standardisation of the Griffiths III revision

Guideline	Findings from the *Griffiths III*	Rating
7.1 Revision teams should give due consideration to the representativeness and size of standardisation samples, in order to develop normative information for a revised test that is applicable to intended test takers.	- Standardisation sample from UK and Republic of Ireland. - Sample describes gender, age, location, and socioeconomic status. - Sample size calculated for continuous norming.	Insufficiently met
7.2 Revised tests should be accompanied at launch with adequate norms and standardisation information.	- Comprehensive norm tables for complete age range of the test.	Insufficiently met

The standardisation sample of the *Griffiths III* consisted of children from the UK and the Republic of Ireland, who were selected for having an uneventful medical history and no concerns regarding developmental milestones (Stroud et al., 2016).

Although the test was also piloted in South Africa, there may be some concern about the applicability of the norms for children outside the UK and the Republic of Ireland. The test manual explores the breakdown of gender, age, urban versus rural representation, and socioeconomic status of the sample (Stroud et al., 2016). The ethnic breakdown of the samples for the South African pilot testing and the standardisation in the UK and Republic of Ireland is not presented, which may raise questions about the cultural validity of the standardisation and norm information. This is of concern as the ITC (2013a; 2016) advises that test norms should not be used for populations where research evidence is lacking about the suitability of the norms. In light of the above information, the revision team of the *Griffiths III* could have done more to include a more diverse sample in terms of international representation, ethnic diversity and different ability levels.

Using the EFPA (2013a) criteria for continuous norming, the sample sizes for years three (n=69), four (n=64) and six (n=57) fall in the 'inadequate' category, with the samples for years one (n=88), two (n=77) and five (n=73) meeting the criteria for 'adequate'. These sample sizes could have been greater if the revision team had used an international sample. The norm tables are comprehensive and the standardisation information for the *Griffiths III* is presented clearly (Foxcroft et al., 2016). The test manuals, however, do not provide adequate information on the international suitability of the norms. For the above reasons, Guidelines 7.1 and 7.2 have been 'Insufficiently met'.

Table 21. Exploration of Phase 8: Conduct supporting research of the Griffiths III revision

Guideline	Findings from the *Griffiths III*	Rating
8.1 Revision teams should prioritise research into all target populations of a revised test, including clinical and non-clinical samples.	- ARICD decision to focus on clinical case studies, instead of norms for clinical samples. - Research on children with autism. - International studies and field work.	Partially met
8.2 Multiple methods should be employed to investigate the relationship between previous and revised editions of a test.	- Correlation between *Griffiths III* and *GMDS-ER*.	Insufficiently met
8.3 Research should be conducted into the validity and reliability of a revised test.	- Different types of validity and reliability have been established.	Partially met

The revision team has prioritised research on the *Griffiths III* for clinical samples. Some of this work has been conducted in conjunction with international researchers, with an emphasis on children along the autism spectrum (Ezhilmangai, 2017). The ARICD has further drafted three documents, specific to children with special needs. The first is a statement on how to use the test with children that score at a development quotient (which is used similarly to an intelligence quotient), of below 50 (ARICD, 2018a). The second communication is a motivation for not developing norms for children with special needs (ARICD, 2018c). The main reason provided is

that the ever-expanding spectrum and varieties of syndromes and disorders would limit the applicability of such norms for many children. The ARICD instead endorses using the *Griffiths III* for atypically developing children as a qualitative tool to determine areas of relative strength and weakness for the child. The purpose of such an assessment is to focus solely on the child being tested and to develop a clear picture of their current skills set, in order to inform the development of an individual support program tailored for the child. In order to assist test users in flagging children for future follow-up, the ARICD (2018d) also produced a document of items in the Gross Motor subscale that are key items to indicate potential delays in children. Whilst these efforts are helpful, the ARICD statements lack clear guidance about how to use the test optimally with children from clinical populations. This still leaves a knowledge-gap for practitioners (Bush et al., 2018).

Subsequent to publication of the revised scales, research studies have also detailed the use of the *Griffiths III* in different countries, including Kenya (Watters, 2017), Israel (Posener, 2017), and South Africa (Jansen, 2017). There is however a need for more international research. As practitioners from South Africa were involved in the test revision, more research on South African samples could be expected from these professionals in the three years since the launch of the test. The body of research on the *Griffiths III* continues to grow, but at the time of drafting the present study, the ARICD still needed to produce more evidence-based research on the utility of the test for children with special needs, as recommended by Oliveri, Lawless and Young (2013). For this reason, Guideline 8.1 has been partially met thus far.

In terms of the relationship between old and new test editions, the test manual details the findings of a quantitative study that established a good relationship between the *Griffiths III* and the *GMDS-ER* (Stroud et al., 2016). Guideline 8.2 calls however for investigations using

multiple methods (Liu & Dorans, 2013; Strauss, Spreen, & Hunter, 2000). A qualitative exploration of the similarities and differences in the underpinning test constructs, and comparative quantitative research of the factor structures of the *Griffiths III* and *GMDS-ER* would be examples of additional research. Given the lack of variety of studies exploring the relationship between the old and new test editions Guideline 8.2 has been 'Insufficiently met'.

The manual explores internal consistency as a function of reliability, and construct delineation and coverage as a function of validity (Stroud et al., 2016). Subsequent to publication, the ARICD added further research into the concurrent validity and stability reliability of the *Griffiths III* on a sample of children from the United Kingdom and Republic of Ireland (Cronje, Green, & Venter, 2017). The concurrent validity included the relationship between the *Griffiths III* and two other tests, namely the *Ages & Stages Questionnaires (3rd Edition)* (Squires & Bricker, 2009) and the *Wechsler Preschool and Primary Scales of Intelligence (4th Edition) – United Kingdom* (WPPSI-IV-UK) (Wechsler, 2012). The stability reliability study incorporated test-retest and interrater reliability components. This follows the criteria of the ITC (2015) that test selection should be based on adequate validity and reliability evidence. Whilst the validity and reliability of the *Griffiths III* has been investigated in different studies, sampling has remained within the countries of the original standardisation. Mattern, Kobrin and Camara (2012) indicate that validity and reliability studies are continuing research priorities beyond the launch of a test. This being said, further studies from other countries are required to add to the body of validity and reliability evidence, and therefore, at present, Guideline 8.3 has been 'Partially met'.

Table 22. Exploration of Phase 9: Test product assembly and launch of the Griffiths III revision

Guideline	Findings from the *Griffiths III*	Rating
9.1 The extent of a revision should be communicated in the product description of a test.	- Revision journey is detailed in test manual.	Sufficiently met
9.2 When tests are revised, users should be informed of the changes to the specifications, underlying constructs, and changes to the scoring method.	- Process of developing new subscale constructs are explained in manual, together with supporting references. - Scoring method is explained, together with practical examples.	Partially met
9.3 Test users should be clearly informed of the comparability and relationship between the previous and revised editions of a test.	- Statistical correlations between *Griffiths III* and *GMDS-ER* are explained.	Insufficiently met
9.4 Documentation for revised tests should be amended and dated to keep information for test users current.	- Manuals and record books have been updated with minor edits since publication, and test users have been informed through communiques. - Manuals are not dated	Partially met
9.5 Test publishers should consider the economic circumstances of test users when determining the cost of a revised test.	- The test is expensive, but at a similar price to main competitors. - Record books, drawing books, and manuals are at a reduced cost to African countries.	Partially met

The manuals of the *Griffiths III* present the extent of the revision, and the motivation for decisions taken during the process. The constructs of each subscale are explored in detail, together with the sources of information that informed the subscale definitions (Stroud et al., 2016). The scoring method is explained in detail, together with practical examples (Green et al., 2016). The extent of explanations provided in the manuals is consistent with the standard published by the AERA (2014), and for this reason Guideline 9.1 has been ‘Sufficiently met’.

Test users are advised in the manuals of the changes to the *Griffiths III* and its relationship to the *GMDS-ER*. A list of items that were retained from the *GMDS-ER*, as well as new items, has been detailed in table format within Part II of the manuals (Green et al., 2016). The statistical

relationship between the *Griffiths III* and *GMDS-ER* is presented in Part I of the manuals (Stroud et al., 2016). These fit the suggestions offered by Liu and Dorans (2013). The manuals do not reflect, however, how the changes to the test may have affected the equivalence of the test's norms between the different editions, as stipulated by the ITC (2013b). The specific changes to the underpinning constructs and how the *Griffiths III* compares at construct level to the *GMDS-ER*, are not explored in the test manuals, as advised by the ITC (2016). This may lead to some confusion for users migrating from the old edition to the newer one. For this reason Guideline 9.2 has been 'Partially met' and Guideline 9.3 is 'Insufficiently met'.

Since the publication of the *Griffiths III* in 2016, the manuals, record book and drawing book have been updated. The reasons for updates included clarifications to individual item instructions, minor editorial changes, and inclusion of more information on the record book and drawing book to assist test users during administration. Test users were advised of changes through the ARICD website, and updated materials have been made available to existing test users. The manuals do not have a system to date them, however, such as a specific month and year of issue, or a sequential numbering system, which is a standard from the AERA (2014). This means that users who purchased the test shortly after launch, and who may have missed updates from the publisher, may be working with outdated manuals. At present, there have not been major changes to the manuals, and test users do have a responsibility to remain updated on communiques from the publisher. Guideline 9.4 has therefore only been 'Partially met'.

The purchase price in 2018 for a *Griffiths III* kit is expensive at around £1605. It is, however, comparable to the purchase price of other popular tests of child development. The *Bayley Scales of Infant and Toddler Development (3rd Edition)* (Bayley, 2006) kit costs around £1,401, whilst the *Wechsler Intelligence Scale for Children (5th Edition) – United Kingdom*

(Wechsler, 2014) costs around £1,299 in 2019. The only accommodation in price made for *Griffiths III* users from some developing countries is the reduced price of replacement manuals, record books and drawing booklets. This is a small price concession for some developing countries, but as a test user will need to replenish record books and drawing booklets, this saving can add up over time. This concession somewhat addresses the comment of Adams (2000) regarding product pricing for developing countries. The *Griffiths III* is therefore priced in line with some of its main competitors, but a general comment would be that these tests are still expensive for test users from developing countries (Gilmore et al., 2015; van Dulm, 2013) which, according to the ITC (2015) can influence the rate that a test is adopted by users in such economies. Therefore, Guideline 9.5 is ‘Partially met’.

Table 23. Exploration of Phase 10: Post-launch activities of the Griffiths III revision

Guideline	Findings from the *Griffiths III*	Rating
10.1 Test publishers and users share a joint responsibility to engage with each other regarding revised tests.	- ARICD website is updated with new information. - Test users can directly contact ARICD and test publisher via email. - Annual ARICD international conference for test users that encourages presentations from webinar delegates.	Sufficiently met
10.2 Test publishers should develop a reasonable strategy to assist test users to switch to a revised test edition.	- Test publisher stopped selling kits of previous editions but continues to sell manuals and record books.	Sufficiently met
10.3 Test publishers should offer comprehensive training to promote the level of competence with which test users employ revised tests.	- Two-step training for new users. - Online course for users of previous edition. - Registration with ARICD required before being allowed to buy *Griffiths III*. - Observational protocol. - ARICD statement on age equivalence. - Error prevention meeting.	Sufficiently met
10.4 Test users should guard against resistance to change, keep current with changes to tests, and strive to adopt a revised test as soon as possible, with due consideration for the best interests of their clients.	- Face-to-face training offered to users of previous edition. - Slower migration to *Griffiths III*.	Partially met
10.5 Revision teams should develop a comprehensive post-launch research strategy and encourage the dissemination of independent research studies.	- Dedicated research committee at ARICD to advise researchers, collaborate, and assist with funding and expertise. - ARICD hosts International Scientific Meeting, to promote presentation of independent studies.	Partially met

Griffiths III users have several avenues to contact the test developer and publisher. They can email the ARICD, the Hogrefe publishers, their local *Griffiths III* distributor, as well as the tutors that conducted the training course they attended. The publisher and ARICD encourage open and active communication between all parties, with contact details of specific areas, such as

training, registration, research and management, being listed on the ARICD and Hogrefe sites. This provides the communication avenues suggested by other authors (Bush, 2010; ITC, 2015; Naglieri et al., 2004). For providing various engagement avenues between test users and the revision team, Guideline 10.1 is 'Sufficiently met'.

Hogrefe has managed the changeover from the *GMDS-ER* and the *Baby Scales* to the *Griffiths III* by discontinuing the sale of kits for previous editions at the launch of the *Griffiths III*. At the launch of the *Griffiths III*, test kits were also sold at a discounted rate on pre-order to encourage the changeover to the new edition. The ITC (2015) agrees with this practice of offering revised tests at a reduced price to encourage test users to adopt the revision edition. The only materials available for purchase for the *GMDS-ER* are record books and manuals that would be important support for pre-existing longitudinal studies. This again aligns with Butcher's (2000) comment for test publishers to have a firm end date for the sale of previous test editions on the launch of a revised test. Guideline 10.2 is therefore 'Sufficiently met'.

In terms of training, new users have to complete an e-learning module through the ARICD and obtain a score of at least 80%. This is followed by an intensive face-to-face three-day training course with a registered *Griffiths III* tutor. During this course, attendees have to be signed off by the tutor as competent in their familiarity of the kit, scoring, report writing, and test administration on children. Only after successful completion of the two-step training is a new user registered with the ARICD, which is a requirement to purchase a *Griffiths III* kit. Registered users of the *GMDS-ER* have to complete an e-learning module in order to be registered as a *Griffiths III* user. They also have the option of a one-day follow-up training with a registered tutor should they feel the need for a face-to-face conversion course. The present researcher and three practitioners from South Africa who contributed to the revision of the *Griffiths III* have

been accredited as tutors for the test. These tutors present about three courses for new users and two conversion courses per year in South Africa. My experience as a tutor is that attendees feel heartened that South Africa was represented in the revision team, but they query the suitability of the norms for South African children and request additional research on South African samples. The training platform has sparked some ongoing research collaborations between the tutors and attendees, specifically researching the *Griffiths III* with developmentally delayed children and children on the autism spectrum.

To assist *Griffiths III* users in using the test accurately, the ARICD hosted a professional development day (ARICD, 2018) that allowed test users to interact with the revision team, observe test administrations, ask questions about test scoring and interpretation, and explore the use of the *Griffiths III* using complex case studies. The *Griffiths III* employs a different method of establishing age equivalence for children, and to clarify the appropriate method to establish the age equivalence of a child's test performance the ARICD (2018b) authored a document for test users. These actions are consistent with recommendations by the EFPA (2013b) and ITC (2013a).

As a test session further affords the test user an opportunity to gather qualitative observation of the child's test behaviour, the ARICD has also encouraged the development of an observational protocol for test users (Currin, 2017). Guideline 10.3 is thus 'Sufficiently met'.

From McAlinden and Bloomfield (2018) it appears that test users have been slow in migrating from the *GMDS-ER* to the *Griffiths III*. In May 2016, there were 8412 registered *GMDS-ER* users. By May 2018, two years after the release of the *Griffiths III*, there were 1076 registered *Griffiths III* users. This figure is 87% lower than for the *GMDS-ER*, although it does raise the question of how many registered *GMDS-ER* users were still actively using the test. As

Bush (2010) indicates that the industry standard is for test users to migrate to a new version within six months, the revision team should investigate possible sources of change resistance or non-adoption of the *Griffiths III*, including the possible effect of the cost of the training and the test kit. Although the ARICD offers face-to-face conversion training on the *Griffiths III* for *GMDS-ER* users, the actual rate of migration is unclear, and for this reason Guideline 10.4 is only 'Partially met'.

The ARICD encourages independent research, by hosting international scientific meetings in the UK, to highlight presentations from international researchers. Test users outside the UK can also register to participate in these meetings online, and even present their findings to the international audience from remote sites. The ARICD encourages researchers to inform them of research projects, as the ARICD is prepared to share expertise, resources, and even funding to facilitate continued research. To this effect, the ARICD is providing support to projects aimed at translating and validating the *Griffiths III* in Italy, Portugal, and Brazil, and has drafted guidelines for researchers involved in *Griffiths III* adaptation studies (McAlinden & Bloomfield, 2018). This agrees with the guidelines of the ITC (2013a) for forums that allow open communication and active debate on a revised test by users and researchers. Despite this, there is still a need for more peer-reviewed research on the *Griffiths III* in academic journals, as suggested by Mattern, Kobrin and Camara (2012). At present, Guideline 10.5 has therefore been 'Partially met'.

In Sum: An Analysis of the Guidelines and the *Griffiths III* Revision Process

The analysis of the *Griffiths III* revision process from the perspective of the guidelines for test revision revealed different levels of similarity between the test's revision and the guideline statements. Twelve guidelines (40%) were sufficiently met by the test's revision process that

indicated a high level of agreement in those areas. Twelve guidelines (40%) were partially met, indicating a moderate level of agreement between the guidelines and the test. Six guidelines (20%) were insufficiently met, indicating a low level of agreement. The percentage of guidelines that were sufficiently met is good, taking into account that the guideline document did not exist at the time of the revision of the *Griffiths III.*

Although some guidelines were insufficiently met, the truth is that any complex process could face valid critiques. Therefore, in fairness, many of the areas, such as the adequacy of pilot testing or norm-sampling, cross-cultural fairness, and elements of bias against certain population groups, can be raised about the revision projects of many tests used internationally in psychological testing. A retrospective review through the lens of the proposed guidelines can inform the future research agenda of a psychological test, to produce the evidence that is currently lacking, yet needed by test users to use the test on different populations.

The first guideline of test revision states that a test revision should strive to improve on its predecessors. The review of the *Griffiths III* demonstrates that this journey of improvement is indeed a process that continues for the lifespan of a test, raising unexpected challenges for revision teams along the way. The case study of the *Griffiths III* has highlighted that, as much as the guidelines for test revision exist as a benchmark for the revision of psychological tests, the implementation of guidelines should be viewed as a continuous journey. This supports the sentiments expressed by van der Linden (2005) that, despite the most rigorous test specifications, developers will struggle in the process of test revision to produce a test that completely meets all of their original project goals. Ultimately, the level of adherence to guidelines is moderated by the resources available at the time, events that occur during the process, and the efforts of practitioners to strive for best practice.

In Chapter Four, the concern was raised that guidelines could infringe on the autonomy of practitioners (Weisz et al., 2007). Some test revision projects may have a smaller scope than the entirety of the 30 guidelines covered in the guidelines document, and therefore revision teams need to decide what guidelines are more important for the goals of their revision project. Whilst this might sound contentious, guidelines do not exist to supervise the work of practitioners, but to offer a suggested route to those practitioners who choose to follow the guidelines (CISA, 2011).

From the perspective of a practitioner from a developing country, South Africa, in an international test revision, I have learnt how test revisions work practically with an international team. The international team seemed to appreciate the contribution of team members from South Africa and increasingly relied on our expertise in cross-cultural test administration in multilingual contexts. This positive regard is supported by the move of the ARICD to appoint a number of South Africans in key roles within the revision of the *Griffiths III.* The roles included tasks such as project management, development of test items and test equipment, authoring test manuals and materials, statistical analysis, and project researchers. Whilst the above appointments are notable accolades, working with international teams can be complex, as final decisions are made at a higher level than the individual contributor. This is highlighted by the fact that the norm sample only consisted of children in the UK and Republic of Ireland which, in retrospect, can be viewed as an oversight by the revision team of this international test, as well as a missed opportunity for the South African members of the revision to contribute South African data to the test's norms.

Concluding Remarks

In this chapter, the themes related to test revision were discussed. The themes were used to illustrate the sentiments contained within each of the 30 guidelines developed for the revision of psychological tests. The guidelines were field-tested using the instrumental case study of the *Griffiths III* to highlight how the guidelines worked in practice for a revised test, and how the guidelines could be useful retrospectively to flag potential concerns in a revised test to guide the ongoing agenda of a revision team. Finally, the case study highlighted that a revision journey extends beyond the publication of a revised test, connecting the revision team to a test beyond its launch. The next chapter concludes the study with considerations of the limitations of the study and suggestions for future research.

Chapter 7: Conclusions, Strengths, Limitations and Recommendations

In this chapter, the conclusions of the study are presented according to the two objectives of the study. This is followed by a reflection on the strengths and limitations of the study. Finally, recommendations are offered for future research, as well as practice recommendations for researchers who intend to follow a similar method to develop guidelines.

Objective One

The process of developing guidelines for the revision of psychological tests and the use of revised psychological tests has impressed on the researcher how important such guidelines can be. It also became apparent how especially difficult the process of developing guidelines can be for individual authors. Guidelines should manage the information from data with the anticipated needs of the intended audience (Jaeschke, Jankowski, Brozek, & Antonelli, 2009; Kish, 2001). The researcher found that solely relying on published research to author guidelines would present a pitfall if there were knowledge gaps in available publications. Such spaces would need to be filled with sound advice that can only come from expertise. Through the present process, the researcher has become mindful of several key components when developing guidelines.

The first component is the importance of having a structured method at the outset (Dijkers, 2013). Different aspects of a guideline development process can raise unique challenges that can slow the project down. By having a structured method in place, it becomes easier to focus on forthcoming steps of the process, to facilitate planning and create forward momentum. This being said, each development process is unique, given the different mix of challenges that emerge. Guideline developers should therefore also remain flexible and open to the issues arising from each challenge, as careful management of different sources of information can enrich the final product. The present study performed a focussed search for literature on test revision and

found limited results. By considering the practice of test revision and the confusion that exists within the literature between such practices as test development and adaptation, the researcher was able to locate sources that contributed to the understanding of test revision. This means that guideline developers should have a thorough grounding in their discipline, as valuable resources and insights may come from specific topics that may be eliminated from keyword searches. Regardless of the information sources, the final guidelines are always written by an author(s) and this requires a level of expertise to ensure a careful balance that reflects both evidence and sound advice. Guidelines that only reflect evidence are at risk of sounding too clinical and not connecting with readers. On the opposite side, guidelines based purely on experience could be incomplete, not aligned to research evidence, or come across as judgemental, thus alienating some readers. As knowledge changes over time, it is also important that guidelines are not written in too rigid terms, to extend the applicability of the guidelines even when some of the supporting information changes.

The second component was the importance of sentiment or meaning in guideline statements (Guédon & Savard, 2000). By including a thematic review of the documents used to develop the guidelines, the researcher was able to determine the common threads between documents. This created a set of messages that should be embedded in the guidelines apart from the specific content that needed to be included in individual guidelines. This included the importance of communication between test publishers, revision teams and test users. By actively encouraging dialogue common ground can be established that may enhance the working relationship between these role players. Other important messages concerned validity, reliability, and fairness as crucial considerations for psychological testing, and therefore test revision.

The third important component is the value of feedback from research promoters, outside experts and critical readers to draft versions of guidelines. In Chapter Four it was stated that guidelines can be created by individual authors without peer-review and that this can result in guidelines of questionable quality and practicality. Authors may fear exposing their work to readers at a stage when there may be room for refinement, but including such a feedback mechanism to penultimate work allows for input from other voices. This creates a product that is less reflective of the interpretations of the author, but more reflective of general expert consensus. Such a document would be more aligned to the ultimate goal of guidelines, which is general expert advice instead of directives from one expert. It further addresses the concerns expressed by authors about the quality of guidelines (Siering, Eikermann, Hausner, Hoffmann-Eßer, & Neugebauer, 2013; Woolf, Schünemann, Eccles, Grimshaw & Shekelle, 2012). For the present guidelines, the researcher utilised peer-reviewers to ensure that the guidelines were practical and did not only reflect the opinions of the researcher, but also those of a panel of experts.

The final component is the importance of keeping the reader in mind when creating guidelines (Jaeschke, Jankowski, Brozek, & Antonelli, 2009). This creates an interesting challenge for guideline developers, as each member of the target audience is at their own level of expertise and familiarity with concepts. Similarly, guidelines should offer best practice advice that is not overly prescriptive, as rigid statements can be seen as undermining the professional judgement of readers. Additionally, a guideline that merely summarises available evidence may be less helpful to some readers, particularly newer members of a profession. Guideline developers should add therefore a more descriptive level of information, including examples, to assist readers in understanding the meaning of a guideline (Donalek & Soldwisch, 2004; Goliath,

2015). In the present study, the researcher followed this process of expounding on guideline statements in an explanatory section below each guideline that has been a valuable contribution to the understanding of the text.

In all, 30 guidelines were developed for test revision. The question arises whether there are more, whether some guidelines should have been split further. In answering, depending on each reader, there may always be room for additional guidelines or increased nuances to a guideline document. For these reasons, guidelines are seen as living documents that are periodically reviewed to reflect new information. For the present study, the 30 guidelines were found to reflect an important starting point of the process of developing guidelines for test revision. The guidelines highlight the need for test revision to be viewed as a unique and important part of the psychological testing industry, which makes test revision worthy of a separate set of guidelines.

Objective Two

The guidelines for test revision were found to be helpful in analysing the revision of a psychological test. The willingness of the revision team of the *Griffiths III*, the ARICD, and the publishers Hogrefe to dismantle, investigate, and redesign a test with a longstanding history is praiseworthy. The process of managing change resistance amongst test users by engaging them for feedback and input throughout the revision process is an example of how test publishers can use a revision project to reconnect with the test market, and enliven an existing test brand. The structured way that each phase of the revision process is documented in the *Griffiths III* manuals also deserves mention. The manuals provide a clear historical record of the revision of the *Griffiths III* that will assist future revisions of the test. In particular, the clear representation of the test manuals allows readers to determine the suitability of the test for specific clients.

Although the guidelines endorsed much of the work and process followed in the revision of the *Griffiths III*, the guidelines also highlighted important aspects that would have enhanced the test's revision project. The most prominent of these related to cross-cultural fairness and the need for more empirical evidence on the cross-cultural suitability of the test, and the fairness of test results. Although the revision team piloted the *Griffiths III* on a multi-cultural sample in South Africa, their results of such fairness studies were not reported in the test manuals. Some test users may also argue that the world is greater than one country. A test that is sold internationally should consider therefore culture on a global scale and report findings from more countries in its review of fairness. Some of these areas cannot be attended to by the present test. This includes greater diversity in the cultural representation of the revision team. Another is the effect of extensive piloting on cross-cultural samples from different countries to inform item development and final item selection. Other aspects may be addressed in future, such as research with clinical samples, and testing children from other countries to create international norms for the *Griffiths III.*

The process of analysing a team's work is always complex, as the quality of the analysis rests on the amount of information the researcher is able to obtain. In the present study the ARICD was open to the review and provided a number of documents and presentations that the researcher would not have been aware of or able to access without compliance from the revision team. This has created an atmosphere where feedback can be given to the ARICD in the confidence that the analysis of the *Griffiths III* produced insights that would stimulate future action from the revision team.

As a member of the revision team of the *Griffiths III,* I valued my participation in the project. I felt my international colleagues valued my ideas and that I was able to contribute

meaningfully to the test. My relationship with the ARICD and the test has continued beyond the launch of the *Griffiths III* in 2016. I have been fortunate to be invited to present research at two International Scientific Meetings of the ARICD, and to meet with the revision team for writing retreats. As such, my professional practice has been enriched by the project. My experience also supports the concept that the lifespan of a revision extends beyond the launch of the test. Comparing my contribution on the *Griffiths III* revision to the guidelines I developed in the present study, has also been helpful for my continuing professional development. In light of the guidelines, I regret that I could have done more to further the cross-cultural validity of the test, particularly through South African research. I also question whether my exposure in international test revision has impacted on my work in South Africa. For instance, I am still concerned about the lack of progress in test development and revision in my country and I am still coming to terms with what I can do to contribute meaningfully in the subdiscipline of psychological testing in South Africa. This stated however, I feel that my contribution to the *Griffiths III* would have been enhanced if I had access to the test revision guidelines that were developed in the present study, and therefore consider these guidelines as valuable for other practitioners. I also value the guideline development experience I gained through this study and feel confident in my ability to contribute to much needed guidelines in psychological testing in South Africa.

Strengths of the Study

An important strength of the present study is the knowledge base of the researcher and the study promoters. As the researcher, I had experience in test development, adaptation, translation, and revision and was therefore able to identify the separate challenges faced in each of these processes, whilst also being mindful of common elements. One study promoter, Professor Watson, had experience in test development, with the level of skill in editing academic journals

that are indispensable in writing guidelines. The other promoter, Professor Stroud, was experienced in test revision as revision leader for the *Griffiths III*. The importance of supervisory feedback in a project of this scope cannot be overemphasised, as it allows researchers room to explore knowledge, develop insights, and perform research in a supportive environment within which they are challenged.

A second strength of the study is that resources were obtained through multiple means, including database searches, mining the references of found documents, reviewing of relevant international conferences, and investigation of organisational websites. This ensured that every effort was made to find all relevant resources for inclusion in the systematic review.

A third strength was the use of a structured methodology, most notably the systematic review process that is reputed for its scientific rigour.

The final strength was the importance of obtaining feedback on the test revision guidelines both through expert reviewers and through application to the case study. Having multiple opportunities for feedback on the guidelines allowed the researcher to reflect critically on the messages conveyed through the guidelines.

Limitations of the Study

The first limitation was that literature searches were limited to English. Although references of found documents were scanned for resources in other languages, the possibility exists that further resources in other languages may exist that were not included in the systematic review.

The second limitation was the absence of a generic process in the subdiscipline of psychological testing to the development of guidelines. The researcher had to tailor a process from the medical professions to create a transparent method to develop guidelines for test

revision. Although this appeared as an initial obstacle, it became an unanticipated strength as it required the researcher to clarify the meaning and purpose of guidelines, and to develop a robust method.

Another potential limitation of the study was that, although practice guidelines are found in the subdiscipline of psychological testing, these documents do not always have the prominence they deserve in daily practice. From experience, many practitioners appear to implement practice guidelines without considering the origins of guidelines, or the needs that underpin them. This creates different levels of compliance with respect to guidelines that may lead to apathy amongst some practitioners towards the importance or relevance of such documents in their work. The guidelines for test revision may be important therefore within themselves, but they may only be of value to those who choose to implement them, although this criticism can similarly apply to other guideline documents.

Recommendations

Recommendations are offered for future research on test revision by other researchers interested in following a similar methodology, for those involved in test revision, and for the *Griffiths III*.

In terms of future research in test revision, the guidelines developed in the present study serve as a starting point for further development and refinement:

- As guidelines are living documents it is recommended that the guideline document developed in this study be updated periodically to reflect advances in knowledge and practice.

- It is recommended that organisations with an interest in test revision read the guidelines of the present study, and either adopt them or tailor the guidelines to their specific needs.
- The guidelines for test revision can be used to investigate the completed revision project of other tests, to further field-test the applicability of the guidelines.
- The guidelines can be used by revision teams embarking on test revision, as the guidelines have undergone a peer-review process and been trialled through application to the case of a completed revision project.
- For guideline documents, developers should be more explicit about the process that was followed to develop the guidelines, as this information is lacking in many guideline and standards documents from prominent national as well as international organisations. Readers should not be expected to accept the validity of guidelines based on the publishing organisation, but should reach a decision based on the process that was followed and sources used to produce the guidelines.
- Researchers interested in following a similar methodology are advised to consider using the guideline development process used in this study, or a similar, proven method.
- Researchers are advised to work closely with experienced supervisors and to submit their guidelines for peer-review, as these aspects are important in enhancing the eventual quality of the guideline statements.

The research also offers recommendations for test users:

- Test users are advised to consider their reasons for test selection, and to remain mindful of the needs of their clients.

- Test users should be active consumers of psychological tests and remain connected to test publishers and the wider community of test users.
- Test users should create an informed opinion of revised tests and adopt them as soon as possible, with due consideration of requirements from professional bodies and the best interests of each client.
- Test users should also forward their opinions and experiences of psychological tests to publishers and be willing to be part of revision projects, either as project members, field-testers, or critical reviewers.

Some recommendations are offered for the ARICD and the revision team of the *Griffiths III:*

- It is recommended that more research be conducted internationally. A recently launched test affords many opportunities for research, as little research would be available on the use of the test in various contexts. This opportunity allows researchers to be the first to contribute to the body of knowledge concerning a test.
- Given that the *Griffiths III* still has a relatively small user base, and that the ARICD encourages researchers to inform the association of research projects, it would be helpful for researchers to engage with the ARICD to determine the uniqueness of their research topic. This would allow researchers to shape their research to contribute new knowledge about the *Griffiths III* or to link with other researchers who are studying a similar topic.
- Research that would be particularly relevant for the *Griffiths III* would be cross-cultural validation, research using clinical samples, and translation or adaptation for different languages or countries.

- For future revisions, the ARICD should be more specific regarding the cultural mix of norm samples, and provide evidence of the cross-cultural validity within the test manuals.

Reflection on the Contributions of the Present Study

As stated in Chapter 1, the present study represents unique contributions, particularly in six areas. The first contribution is the development of a framework that encapsulates the ten phases of the process of test revision. Whilst 'moderate' or 'extensive' revision projects would use all the phases of the framework those involved in 'light' revisions are encouraged to analyse the project against all ten phases of the framework in order to ensure the exhaustiveness of the project and the attention to detail that may be easily overlooked.

The second contribution is that the present researcher developed a comprehensive process to develop guidelines in the discipline of psychology that would be useful for other projects. The practice of merging data with expert input to author guidelines also adds valuable insight for other guideline authors.

The development of 30 guidelines that cover the lifespan of a revision project is the third contribution. These guidelines go beyond the extent of generic guidelines available from psychological testing organisations. The guidelines would be useful to all practitioners either involved in the revision of tests or faced with the transition between previous and revised test editions.

The analysis of the *Griffiths III* revision from the perspective of the guidelines developed by the present researcher is the fourth contribution. This analysis is the first of its kind, and serves as a practical example for revision teams to follow as they analyse the revision projects of their own tests.

The fifth contribution is an applied example of the developing field of mixed-method research using a combination of qualitative research methods. Many studies in psychology are qualitative in nature and the broadening of the mixed-methods design from its traditional roots as a combination of quantitative and qualitative methods is an important development in research design for the social sciences. In the present study, the researcher had to follow a qualitative method in Objective One to develop the guidelines as the topic did not lend itself to a quantitative approach. The added concern was that even if a quantitative approach could be used, there were too many gaps in the available data. These missing pieces would also be reflected in quantitative data analysis. The researcher needed to use a qualitative method to reflect on the available data, consider the gaps in knowledge and address these with new information. Similarly, the case study in Objective Two was best undertaken from a qualitative perspective, to again reflect on the revision of the *Griffiths III* from the perspective of the proposed guidelines, and to highlight the extent of similarity or dissimilarity between the test's revision and the guidelines. This method created a basis to critique the actions of the revision team for the *Griffiths III* and to inform future research and development agendas for the test.

The sixth contribution is the present researcher's reflection of his involvement in the revision of the *Griffiths III*. Being from a developing country, South Africa, I reflected on my contribution within an international test revision project, in terms of how this contributed to my professional development as well as how I could have contributed more to the project to further cross-cultural research on the test during its development. Apart from the benefits I have gained from the *Griffiths III* revision, I am more convinced that my contribution to the project would have benefitted if I had had access to the test revision guidelines I developed in this study. Test

revisions are complex and at times it becomes difficult to decide on the best way forward. The guidelines developed in this study will assist myself and others in future test revisions.

Concluding Remarks

In the information age, practitioners have access to a body of discipline-specific knowledge. It can be difficult however to consider all opinions, weigh the available evidence, and decide on a course of action, especially within time sensitive work environments. The benefit of guidelines is that they provide practitioners with advice that is supported by a process of sourcing evidence, data analysis, and feedback from experts in the relevant discipline. The success of guidelines depends however on the exhaustiveness of the data considered in their development, the quality of the process that was followed to develop the guidelines, and the clarity and definitiveness with which the guidelines are written.

The present study found a lack of clear and comprehensive guidelines from organisations in the subdiscipline of psychological testing, about the revision of psychological tests, and how test users should engage with revised tests. The study sought therefore to develop such guidelines to inform revision teams and test users. The 30 guidelines developed describe the ten phases of test revision, which cover the lifespan of a revision process. The ten phases were envisaged by the researcher and ordered as a stepwise framework for the revision of psychological tests and, as such, are a unique contribution of the study.

Importantly the 30 guidelines highlighted that the revision of a psychological test differs from the development of a new test, and that a revision process does not end on publication of the revised test. The journey continues post-launch, which requires a commitment from revision teams, publishers, researchers and test users to continue engaging with each other to ensure the success of the test. The primary consideration in psychological testing should not be about the

preferences of practitioners or project agendas, but about the best interests of test clients. A test may just be a created product, but their results carry meaning. A test session creates an opportunity for test users to interact meaningfully with clients in a process that can have a positive impact on the lives of clients. It is therefore important that those professional role players in psychological testing, including publishers, revision teams and test users, ensure that revised tests meet the highest standards of quality and fairness.

References

Adams, K. M. (2000). Practical and ethical issues pertaining to test revisions. *Psychological Assessment, 12,* 281-286.

Altmann, T., & Roth, M. (2018). The Self-esteem Stability Scale (SESS) for cross-sectional direct assessment of self-esteem stability. *Frontiers in Psychology, 9*(91). doi:10.3389/fpsyg.2018.00091

American Academy of Neurology. (2011). *Clinical practice guideline process manual*. St. Paul, MN: The American Academy of Neurology.

American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

American Educational Research Association. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

American Psychological Association (2017). *Ethical principles of psychologists and code of conduct.* Retrieved from http://www.apa.org/ethics/code/

American Psychological Association, (n.d.). *Understanding psychological testing and assessment.* Retrieved from http://www.apa.org/helpcenter/assessment.aspx

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

Anderson, L. W. & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* New York: Longman.

Assessment Oversight and the Personnel Psychology Centre. (2009). *Structured Interviewing: How to design and conduct structured interviews for an appointment process.* Retrieved

from https://www.canada.ca/content/dam/canada/public-service-commission/migration/plcy-pltq/guides/structured-structuree/rpt-eng.pdf

Association for Research in Infant and Child Development. (2018a). *ARICD statement on use of Griffiths III Scales of Child Development for children functioning below a developmental quotient of 50.* London: Author.

Association for Research in Infant and Child Development. (2018b). *ARICD statement on the appropriate use of developmental age equivalents in Griffiths III*. London: Author.

Association for Research in Infant and Child Development (2018c). *ARICD statement on normative scoring.* London: Author.

Association for Research in Infant and Child Development (2018d). *Griffiths III Subscale E Gross Motor 'red flag' items*. London: Author.

Association for Research in Infant and Child Development (2018, October 12). Professional development day for Griffiths users. Royal College of Paediatrics and Child Health, London.

Australian Psychological Society. (2018). *Ethical guidelines for psychological assessment and the use of psychological tests.* Melbourne: Author.

Aveyard, H., & Sharp, P. (2011). *A beginner's guide to evidence-based practice in health and social care.* Glasgow: McGraw Open Press.

Aylward, G. P. (2009). Developmental screening and assessment: What are we thinking? *Journal of Developmental and Behavioral Pediatrics, 30*(2), 169-173.

Aylward, G. P., & Aylward, B. (2011). The changing yardstick in measurement of cognitive abilities in infancy. *Journal of Developmental and Behavioral Pediatrics, 32*(6), 465-8.

Babbie, E. (2016). The practice of social research (14th ed.). Boston, MA: Cengage Learning.

Barnes, B. R. (2012). Using mixed methods in South African psychological research. *South African Journal of Psychology, 42*(4), 463-475.

Battacherjee, A. (2012). *Social science research: Principles, methods, and practices* (2nd ed.). Retrieved from http://scholarcommons.usf.edu/oa_textbooks/3

Bayley, N. (2006). *Bayley scales of infant and toddler development: Administration manual.* San Antonio, TX: Harcourt Assessment.

Bechger, T., Hemker, B., & Maris, G. (2009). *Over het gebruik van continue normering [On the use of continuous norming]*. Arnhem, The Netherlands: Cito.

Bers, M. (2012). Designing digital experiences for positive youth development. N.Y. Oxford.

Betehenner, D. W. (2009). *Growth, standards and accountability.* Center for Educational Assessment. Retrieved from http://www.nciea.org/cgi-bin/pubspage.cgi

Blech, E. (2007). *Thriving through change: A leader's practical guide to change mastery.* Alexandria, VA: ASTD Press.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals*. New York: David McKay Company.

Boland, A., Cherry, M. G., & Dickson, R. (2014). *Doing a systematic review: A student's guide.* London: SAGE.

Brannigan, G. G., & Decker, S. L. (2006). The Bender-Gestalt II. *American Journal of Orthopsychiatry, 70,* 10-12.

British Psychological Society. (2017). *Psychological testing: A test user's guide.* Retrieved from https://ptc.bps.org.uk/sites/ptc.bps.org.uk/files/guidance_documents/ptc02_test_users_guide_2017_web.pdf

Bryman, A. (2016). *Social research methods.* Oxford: Oxford University Press.

Bryman, A., & Bell, E. (2014). *Research methodology: Business and management contexts.* Cape Town: Oxford University Press.

Burghardt, G. M., Bartmess-LeVasseur, J. N., Browning, S. A., Morrison, K. E., Stec, C. L., Zachau, C. E., & Freeberg, T. M. (2012). Perspectives – minimizing observer bias in behavioral studies: A review and recommendations. *Ethology, 118*(6). Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1439-0310.2012.02040.x/epdf

Burke, A., Austin, T-L., Bezuidenhout, C., Botha, K., Du Plessis, E., Jordaan, E., … Vorster, A. (in press). *Understanding psychopathology: South African perspectives* (3rd ed.). Cape Town: Oxford University Press.

Bush, S. S. (2010). Determining whether or when to adopt new versions of psychological and neuropsychological tests: Ethical and professional considerations. *The Clinical Neuropsychologist, 24,* 7-16.

Bush, S. S., Sweet, J. J., Bianchini, K. J., Johnson-Greene, D., Dean, P. M., & Schoenberg, M. R. (2018). Deciding to adopt revised and new psychological and neuropsychological tests: an inter-organisational position paper. *The Clinical Neuropsychologist, 32*(3), 319-325.

Butcher, J. N. (2000). Revising psychological tests: Lessons learned from the revisions of the MMPI. *Psychological Assessment, 12,* 263-271.

Butcher, J. N. (2009). *Oxford handbook of personality assessment*. New York: Oxford University Press.

Butcher, J. N., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Manual for the restandardized Minnesota Multiphasic Personality Inventory: MMPI-2*. Minneapolis: University of Minnesota Press.

Camara, W. J. (2007). *Standards for educational and psychological testing: Influence in assessment development and use.* Retrieved from http://www.teststandards.org/files/standards%20-%20influence%202007.pdf

Campbell Collaboration. (n.d.). *The Campbell Collaboration: Background.* Retrieved from http://www.campbellcollaboration.org/background/ index.php

Carretero-Dios, H., & Perez, C. (2007). Standards for the development and review of instrumental studies: Considerations about test selection in psychological research. *Intemational Joumal of Clinical and Health Psychology, 7,* 863-882.

Certified Information Systems Auditor (CISA). (2011). *Understanding policies, standards, guidelines, and procedures.* Retrieved from http://cisacertified.blogspot.co.za/2011/04/understanding-policies-standards.html

Coaley, K. (2009). *An introduction to psychological assessment and psychometrics.* London: SAGE.

Chau, C. L. (2014). *Positive technological development for young children in the context of children's mobile apps.* Unpublished doctoral thesis, Tufts University, Medford, United States of America.

Chilemba, W., van Wyk, N. C., & Leech, R. (2014). Development of guidelines for the assessment of abuse in women living with HIV/AIDS in Malawi. *African Journal for Physical, Health Education, Recreation and Dance, 20*(3:2), 1189-1201.

Cohen, R. J., & Swerdlik, M. E. (2018). *Psychological testing and assessment: An introduction to tests and measurement* (9th ed.). New York: McGraw Hill.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative and mixed methods approaches.* London: SAGE.

Creswell, J. W., Hanson, W. E., Plano Clark, V. L., & Morales, A. (2007). Qualitative research designs: Selection and implementation. *The Counseling Psychologist, 35*(2), 236-264.

Cronje, J. (2016, May 6). *Griffiths III: Norm sample, standardisation and comparison with GMDS-ER*. Presentation at the launch of the Griffiths III, London, United Kingdom

Cronje, J. (2009). *A systematic review of higher education admissions testing practices in Israel: Implications for South Africa.* Unpublished master's dissertation, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa.

Cronje, J., Green, E., & Venter, D. (2017). *Griffiths III psychometrics: ongoing reliability and validity studies*. Paper presented at the 17th International Scientific Meeting of the Association for Research in Infant and Child Development, Birmingham, United Kingdom.

Currin, L. (2017). *Towards the development of an observation protocol for the Griffiths III*. Paper presented at the 17th International Scientific Meeting of the Association for Research in Infant and Child Development, Birmingham, United Kingdom.

Daley, T. C., Whaley, S. E., Sigman, M. D., Espinosa, M. P. & Neumann, C. (2003). IQ on the rise: The Flynn effect in rural Kenyan children. *Psychological Science, 14,* 214-219.

Daly, J. (2005). *Evidence-based medicine and the search for a science of clinical care.* Berkeley: University of California Press/New York: Milbank Memorial Fund.

De Vos, A. S., Strydom, H., Fouche, C. B., & Delport, C. S. L. (2014). *Research at grassroots level* (4th ed.). Pretoria, South Africa: Van Schaik Publishers.

Dijkers, M. (2013). *Introducing GRADE: A systematic approach to rating evidence in systematic reviews and to guideline development.* Retrieved from http://www.ktdrr.org/products/update/v1n5/

Domino, M. L. (2006). *Psychological testing: An introduction* (2nd ed.). New York: Cambridge University Press.

Donalek, J. G., & Soldwisch, S. (2004). An introduction to qualitative research methods. *Urologic Nursing, 24*(4), 354-356.

Dunbar-Krige, H., Venter, R., Nel, M., & Mavuso, M. (2015). *Guidelines for assessment adaptation.* Pretoria, South Africa: Van Schaik Publishers.

Education Testing Service (ETS). (2014). *ETS standards for quality and fairness.* Retrieved from www.ets.org.

Education Testing Service (ETS). (2009). *ETS international principles for fairness review of assessments: A manual for developing locally appropriate fairness review guidelines in various countries.* Retrieved from www.ets.org.

Evidence for Policy and Practice Information (EPPI) Centre. (n.d.). *The role and work of the EPPI-centre.* Retrieved from http://eppi.ioe.ac.uk/cms/ Default.aspx?tabid=63

European Federation of Psychologists' Associations (EFPA). (2013a). *EFPA review model for the description and evaluation of psychological and educational tests: Test review form and notes for reviewers (Version 4.2.6).* Retrieved from http://www.efpa.eu/download/650d0d4ecd407a51139ca44ee704fda4

European Federation of Psychologists' Associations (EFPA). (2013b). *Performance requirements, context definitions and knowledge & skill specifications for the three EFPA levels of qualifications in psychological assessment.* Retrieved from http://www.efpa.eu/download/1b272a998e297c248413fbb761134697

Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R. R. (2010). The Dutch review process for evaluating the quality of psychological tests: History, procedure, and results. *International Journal of Testing, 10,* 295-317.

Eyde, L. D., Robertson, G. J., & Krug, S. E. (2010). *Responsible test use: Case studies for assessing human behavior* (2nd ed.). Washington, DC: American Psychological Association.

Ezhilmangai, R.P. (2017). *The impact of Griffiths Mental Development Scale in autism spectrum disorder*. Paper presented at the 17th International Scientific Meeting of the Association for Research in Infant and Child Development, Birmingham, United Kingdom.

Ferreira, R. (2016). *Psychological assessment: Thinking innovatively in contexts of diversity*. Cape Town: Juta.

Field, M. J., & Lohr, K. N. (1990). *Clinical practice guidelines: directions of a new program.* Washington, DC: National Academy Press.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101,* 171–191.

Foster, D., & Miller, J. L. (2010). *Global test security issues and ethical challenges.* Paper presented at the International Test Commission Conference, Hong Kong.

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

Foxcroft, C., Paterson, H., le Roux, N., & Herbst, D. (2004). *Psychological assessment in South Africa: A needs analysis.* Retrieved from http://www.hsrc.ac.za/en/research-outputs/ktree-doc/1186

Foxcroft, C. D. (2011). Ethical issues related to psychological testing in Africa: What I have learned (so far). *Online Readings in Psychology and Culture, 2*(2). Retrieved from http://dx.doi.org/10.9707/2307-0919.1022

Foxcroft, C. D., & Roodt, G. (2013). *Introduction to psychological assessment in the South African context* (4th ed.). Cape Town: Oxford University Press.

Fried, E. I., & Flake, J. K. (2018). Measurement matters. *Observer, 31*(3). Retrieved from https://www.psychologicalscience.org/observer/measurement-matters

Fröhner, J. H., Teckentrup, V., Smolka, M. N., & Kroemer, N. B. (2017). *Addressing the reliability fallacy: Similar group effects may arise from unreliable individual effects.* Retrieved from https://www.biorxiv.org/content/early/2017/11/07/215053

Frost, N., Nolas, S. M., Brooks-Gordon, B., Esin, C., Holt, A., Mehdizadeh, L., & Shinebourne, B. (2010). Pluralism in qualitative research: The impact of different researchers and qualitative approaches on the analysis of qualitative data. *Qualitative Research 10*(4), 441-460.

Geisinger, K. F. (Ed.). (2013). *APA handbook of testing and assessment in school psychology and education* (Volume 3). Washington, DC: APA.

Gilmore, L., Islam, S., Su, H., & Younesian, S. (2015). A global perspective on psycho-educational assessment. *Journal of Psychologists and Counsellors in Schools, 25*(1), 66-74.

Glanville, J., & Lefebvre, C. (2000). Identifying systematic reviews: Key resources. *Evidence-Based Medicine, 5,* 68-69.

Glasziou, P., Irwig, L., Bain, C., & Colditz, G. (2001). *Systematic reviews in health care: A practical guide.* Cambridge: Cambridge University Press.

Gough, D., Oliver, S., & Thomas, J. (2012). *An introduction to systematic reviews.* London: SAGE.

Gough, D., Thomas, J., & Oliver, S. (2012). Clarifying differences between review designs and methods. *Systematic Reviews,* 1(28). doi:10.1186/2046-4053-1-28

Graham, J. R. (2012). *MMPI-2: Assessing personality and psychopathology* (5th ed.). New York: Oxford University Press.

Grbich, C. (2007). *Qualitative data analysis: An introduction.* Thousand Oaks, CA: SAGE.

Gregory, R. J. (2015). *Psychological testing: History, principles, and applications* (7th ed.). Harlow, UK: Pearson

Griffiths, R. (1935). *A study of imagination in early childhood and its function in mental development*. London, UK: Routledge.

Griffiths, R. (1954). *The abilities of babies.* New York, NY: McGraw-Hill.

Griffiths, R. (1970). *The abilities of young children: A comprehensive system of mental measurement for the first eight years of life*. London: Association for Research in Infant and Child Development (ARICD) & University of London Press.

Gronseth, G. S., Woodroffe, L. M., & Getchius, T. S. (2011). Clinical practice guideline process manual. *St. Paul, MN: American Academy of Neurology*.

Grotth-Marnat, G. (2009). *Handbook of psychological assessment* (5th ed.). Hoboken, N.J.: John Wiley & Sons.

Guédon, M.-C., & Savard, R. (2000). *Tests à l'appui: Pour une intervention intégrée de la psychométrie en counseling d'orientation.* Sainte-Foy: Septembre.

Hambleton, R. K., & Xing, D. (2006). Optimal and monoptimal computer-based test designs for making pass–fail decisions. *Applied Measurement in Education, 19*(3), 221-239.

Harden, A. (2010). Mixed-methods systematic reviews: Integrating quantitative and qualitative findings. *Focus Technical Brief (25)*. Retrieved from http://ktdrr.org/ktlibrary/articles_pubs/ncddrwork/focus/focus25/Focus25.pdf

Hasson, F., Keeney, S. & McKenna, H. (2000). Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing, 32*(4), 1008-1015.

Hathaway, S. R., & McKinley J. C. (1942). *Manual for the Minnesota Multiphasic Personality Inventory*. Minneapolis: University of Minnesota Press.

Hemingway, P., & Brereton, N. (2009). What is a systematic review? *European Journal of Oral Implantology, 1*, 174–175.

Heuchert, J. W. P., Parker, W. D., Stumpf, H., & Myburgh, C. P. H. (2000). The Five-Factor Model of personality in South African college students. *American Behavioral Scientist, 44*(1), 112–125.

Hogan, T. P. (2013). *Psychological testing: A practical introduction* (3rd ed.). Hoboken, N.J.: John Wiley & Sons.

Holman, L, Head, M.L, Lanfear, R., & Jennions, M.D. (2015). Evidence of experimental bias in the life sciences: Why we need blind data recording. *PLoS Biology 13*(7). Retrieved from http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002190

Huntley, M. (1996). *Griffiths Mental Development Scales from birth to 2 years – Manual*. Oxford, UK: ARICD.

International Test Commission. (2016). *ITC Guidelines for translating and adapting tests* (2nd ed.). Retrieved from http://www.intestcom.org/

International Test Commission. (2015). *Guidelines for practitioner use of test revisions, obsolete tests, and test disposal.* Retrieved from https://www.intestcom.org/files/guideline_test_disposal.pdf

International Test Commission. (2013a). *ITC guidelines on test use*. Retrieved from http://www.intestcom.org/files/guideline_test_use.pdf

International Test Commission. (2013b). *ITC Guidelines on quality control in scoring, test analysis, and reporting of test scores*. Retrieved from https://www.intestcom.org/files/guideline_quality_control.pdf

International Test Commission. (2005). *ITC Guidelines on computer-based and internet delivered testing.* Retrieved from https://www.intestcom.org/files/guideline_computer_based_testing.pdf

Jaeschke, R., Jankowski, M., Brozek, J., & Antonelli, M. (2009). How to develop guidelines for clinical practice. *Minerva Anestesiologica, 75,* 504-508.

Jansen, J. (2017). *Implementing assessment and intervention programmes in a university community clinic situated in a rural area in Port Elizabeth, South Africa*. Paper presented at the 17th International Scientific Meeting of the Association for Research in Infant and Child Development, Birmingham, United Kingdom.

Joanna Briggs Institute (2011). *Joanna Briggs reviewers' manual*. Adelaide, Australia: Joanna Briggs Institute.

Kaplan, R. M., & Saccuzzo, D. P. (2013). *Psychological assessment and theory: Creating and using psychological tests* (8th ed.). Toronto, Canada: Wadsworth, Cengage Learning.

Kamens, D. H., & McNeely, C. L. (2010). Globalization and the growth of international educational testing and national assessment. *Comparative Education Review, 54,* 5-25.

Kassin, S. M., Dror, I. E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition, 2*(1), 42-52.

King, M. C. (2006). Adopting revised versions of psychological tests. *The CAP Monitor, 23*, 6-7.

Kish, M. A. (2001). Guide to development of practice guidelines. *Clinical Infectious Diseases, 32*(6), 851-854.

Klepac, R. K., Ronan, G. F., Andrasik, F., Arnold, K. D., Belar, C. D., Berry, S. L., …Strauman, T. J. (2012). Guidelines for cognitive behavioral training within doctoral psychology programs in the United States: Report of the Inter-Organisational Task Force on Cognitive and Behavioral Psychology Doctoral Education. *Behavior Therapy, 43*, 687–697.

Kyriacou, C., & Issitt, J. (2008). *What characterizes effective teacher-initiated teacher-pupil dialogue to promote conceptual understanding in mathematics lesson in England in key stages 2 and 3: A systematic review.* Evidence for Policy and Practice Information Centre. London, England: University of London.

Koch, S. E. (2005). *Evaluating the equivalence, across language groups, of a reading comprehension test used for admissions purposes.* Unpublished doctoral thesis, Nelson Mandela Metropolitan University, South Africa.

Laher, S., & Cockcroft, K. (2013). *Psychological assessment in South Africa: Research and applications.* Johannesburg: Wits University Press.

Laher, S., Fynn, A., & Kramer, S. (2019). *Transforming research methods in the social sciences: Case studies from South Africa.* Johannesburg: Wits University Press.

Liu, J., & Dorans, N. J. (2013). Assessing a critical aspect of construct continuity when test specifications change or test forms deviate from specifications. *Educational Measurement: Issues and Practice, 32,* 15-22.

Liu, J., & Walker, M. E. (2007). Score linking issues related to test content changes. In N. J. Dorans, M. Pommerich, & P. W. Holland, (Eds.), *Linking and aligning scores and scales*. (pp. 109–134). New York, NY: Springer-Verlag.

Lincoln, Y., & Guba, E. (1999). Establishing trustworthiness. In Bryman, A., & Burgess, R.G. (Eds.). *Qualitative research Vol III.* London: SAGE.

Luiz, D., Faragher, B., Barnard, A., Knoesen, N., Kotras, N., Burns, L. E., …, O'Connell, R. (2006). *Griffiths Mental Development Scales – Extended Revised (GMDS-ER)*. Oxford, UK: ARICD.

Marais, R., Stroud, L., Foxcroft, C., & Cronje, J. (2017). *Connecting items of the Griffiths III using tablet-based gamification and a storyline*. Paper presented at the 17th International Scientific Meeting of the Association for Research in Infant and Child Development, Birmingham, United Kingdom.

Marimwe, C., & Dowse, R. (2017). Development of an item bank of health literacy questions appropriate for limited literacy public sector patients in South Africa. *Journal of Communication in Healthcare, 10*(4), 273-284.

Mattern, K. D., Kobrin, J. L., & Camara, W. J. (2012). Promoting rigorous validation practice: An applied perspective. *Measurement: Interdisciplinary Research and Perspectives, 10*(1-2), 88-92.

Maul, A. (2017) Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives, 15*(2), 51-69.

McAlinden, P., & Bloomfield, S. (2018). *The use of the Griffiths Scales of Child Development around the world.* Paper presented at the first Griffiths Assessment Scales Scientific Meeting, Shanghai, China.

McCauley, R. J., & Strand, E. A. (2008). A review of standardized tests of nonverbal oral and speech motor performance in children. *American Journal of Speech-Language Pathology, 17,* 81-91.

McCrae, R. R. (2018). Method biases in single-source personality assessments. *Psychological Assessment, 30*(9), 1160-1173.

McDonald, A., Newton, P., Whetton, C., & Benefield, P. (2001). *Aptitude testing for university entrance: A literature review.* National Foundation for Educational Review. Slough, England: NFER.

Merriam, S. B., Caffarella, R. S., & Baumgartner, L. M. (2007). *Learning in adulthood: A comprehensive guide* (3rd ed.). San Francisco: Jossey-Bass.

Meyer, G. J., & Eblin, J. J. (2012). An overview of the Rorschach Performance Assessment System (R-PAS). *Psychological Injury and Law, 5*, 107-121.

Moerdyk, A. (2015). *The principles and practice of psychological assessment* (2nd ed.). Pretoria: Van Schaik Publishers.

Monette, D. R., Sullivan, T. J., & De Jong, C. R. (2011). Applied Social Research: A tool for the human services. Belmont, CA: Brooks/Cole Cengage Learning.

Morrison, A., Moulton, K., Clark, M., Polisena, J., Fiander, M., Mierzwinski-Urban, M., … & Hutton, B. (2009). English-language restriction when conducting systematic review-based meta-analyses: Systematic review of published studies. *Canadian Agency for Drugs and Technologies in Health*. Retrieved from

https://www.cadth.ca/media/pdf/H0478_Language_Restriction_Systematic_Review_Pub_Studies_e.pdf

Mouton, J. (2001). *How to succeed in your masters and doctoral studies: A South African guide and resource book.* Pretoria: Van Schaik.

Mpasa, F. (2014). *Strategies for the implementation of clinical practice guidelines in the intensive care: A systematic review.* Unpublished master's dissertation, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa.

Murphy, K. R., & Davidshofer, C. O. (2004). *Psychological testing: Principles & applications* (6th ed.). Englewood Cliffs, New Jersey: Prentice-Hall.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the internet: New problems, old issues. *American Psychologist, 59*(3), 150-162.

National Academy of Sciences. (2009). *Strengthening forensic science in the United States: A path forward.* Washington, DC: National Academies Press.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1978). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research.* Bethesda, Md.: The Commission.

Neuman, W. L. (2011). *Social research methods. Qualitative and quantitative approaches* (7th ed.). Boston, MA: Pearson International.

Noor, K. B. M. (2008). Case study: A strategic research methodology. *Applied Science, 5*, 1602-1604.

Nunnally, J. C., & Bernstein, I. H. (1993). *Psychometric theory*. New York: McGraw-Hill.

Oliver, S., & Peersman, G. (2001). *Using research for effective health promotion.* Buckingham, England: Open University Press.

Piaw, C. Y. (2012). Replacing paper-based testing with computer-based testing in assessment: Are we doing wrong? *Procedia - Social and Behavioral Sciences, 64,* 655-664.

Oliveri, M. E., Lawless, R., & Young, J. W. (2015). *A validity framework for the use and development of exported assessments.* Princeton, NJ: Educational Testing Service.

Paterson, H., & Uys. K. (2005). Critical issues in psychological test use in the South African workplace. *South African Journal of Industrial Psychology, 2005, 31*(3), 12-22.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Malden, MA: Blackwell Publishing.

Posener, E. (2017). *The Israeli experience with Griffiths III: First research, pitfalls and tips*. Paper presented at the 17th International Scientific Meeting of the Association for Research in Infant and Child Development, Birmingham, United Kingdom.

Proctor, E. K., & Staudt, M. (2003). Targets of change and interventions in social work: An empirically based prototype for developing practice guidelines. *Research on Social Work Practice*, *13*(2), 208-233.

Protection of Personal Information Act, no. 4 of 2013. Retrieved from http://www.justice.gov.za/legislation/acts/2013-004.pdf

Proyer, R. T., & Häusler, J. (2007). Assessing behavior in standardized settings: The role of objective personality tests. *International Journal of Clinical and Health Psychology, 7*(2), 537-546.

Punch, K. F. (2014). *Introduction to social research: Quantitative and qualitative approaches*. London: Sage.

Rhoades, K. & Madaus, G. (May, 2003). *Errors in standardized testing: A systematic problem.* Retrieved from http://www.bc.edu/research/nbetpp/statements/M1N4.pdf

Rorschach, H. (1927). *Rorschach test – psychodiagnostic plates.* Cambridge, MA: Hogrefe.

Ryan, G. W., & Bernard, H. R. (n.d.). *Techniques to identify themes in qualitative data.* Retrieved from http://www.analytictech.com/mb870/readings/ryan-bernard_techniques_to_identify_themes_in.htm

Samuel, C. (2014). *Practitioners' views of the Griffiths Scales: Informing the revision process*. Unpublished master's treatise, Nelson Mandela Metropolitan University, South Africa

Sampson, J. P. (1999). Integrating internet-based distance guidance with services provided in career centers. *The Career Development Quarterly, 47*(3), 243-254.

Savard, R., Gingras, M., & Turcotte, M. (2002). Delivery of career development information in the context of information computer technology. *International Journal for Educational and Vocational Guidance, 2*(2), 173-191.

Schlosser, R. W. (2007). Appraising the quality of systematic reviews. *Focus Technical Brief (17)*. Retrieved from http://www.ktdrr.org/ktlibrary/articles_pubs/ncddrwork/focus/focus17/Focus17.pdf

School of Health and Related Research. (n.d.). *Systematic review: What they are and why they are useful?* Retrieved from http://www.shef.ac.uk/scharr/ir/ units/systrev/index.htm

Schwarz, N. (1999). Self-reports: how the questions shape the answers. *American Psychologist, 54*, 93-105.

Shenton, A. K. (2004). Strategies for ensuring trustworthiness in qualitative research. London, UK: IOS Press.

Siering, U., Eikermann, M., Hausner, E., Hoffmann-Eßer, W., & Neugebauer, E. A. (2013). Appraisal tools for clinical practice guidelines: A systematic review. *Plos One, 8*(12), e82915.

Silverman, D. (2011). *Interpreting qualitative data: A guide to the principles of qualitative research* (4th ed.). London: SAGE.

Silverstein, M. L., & Nelson, L. D. (2000). Clinical and research implications of revising psychological tests. *Psychological Assessment, 12,* 293-303.

Simpson, S. H. (2011). Demystifying the research process: Mixed methods. *Pediatric Nursing, 37*(1), 28-29.

Squires, J., & Bricker, D. (2009). *Ages & Stages Questionnaires (3rd Edition).* Baltimore: Paul H. Brookes Publishing Co.

Steyn, G. M. (2011). Determining guidelines for professional development: A qualitative study. *Journal of Educational Studies*, *10*(1), 2-30.

Strauss, E., Spreen, O., & Hunter, M. (2000). Implications of test revisions for research. *Psychological Assessment, 12,* 237-244.

Stroud, L. (2016, May 6). *Griffiths III*. Presentation at the launch of the Griffiths III, London, United Kingdom.

Stroud, L. (2013). *The Griffiths III: Revising the scales from the inside out*. Paper presented at the 15th International Scientific Meeting of the Association for Research in Infant and Child Development, Birmingham, United Kingdom.

Stroud, L., Green, E., Bloomfield, S., & McAlinden, P. (2017). *Revision process, standardisation and psychometric properties and a link to neurorehabilitation*. Paper presented at the European Paediatric Neurology Society Congress, Lyon, France.

Stroud, L., Foxcroft, C., Green, E., Bloomfield, S., Cronje, J., Hurter, K., …, Venter, D. (2016). *Griffiths Scales of Child Development (3rd Edition) Part I: Overview, development and psychometric properties*. Oxford, United Kingdom: Hogrefe.

Stroud, L., Foxcroft, C., Cronje, J., & Marais, R. (2014). *Promoting the relevance of developmental tests over time: The case of the Griffiths*. Paper presented at the 20th South African Psychology Congress, Durban, South Africa.

Stroud, L., Foxcroft, C., & Marais, R. (2012). *Re-standardising the Griffiths Mental Development Scales – Phase one, its challenges and triumphs.* Paper presented at the 30th International Congress of Psychology, Cape Town, South Africa.

Swanepoel, I., & Krüger, C. (2011). Revisiting validity in cross-cultural psychometric test development: A systems-informed shift towards qualitative research designs. *South African Journal of Psychology, 17*(1), 10-15.

Teasdale, T., & Owen, D. (2005). A long-term rise and recent decline in intelligence test performance: The Flynn effect in reverse. *Personality and Individual Differences, 39,* 837–843.

Tellegen, A., & Ben-Porath, Y. S. (2008/2011). *MMPI-2-RF (Minnesota Multiphasic Personality Inventory-2 Restructured Form) technical manual*. Minneapolis: University of Minnesota Press.

Ten Ham-Baloyi, W., & Jordan, P. (2016). Systematic review as a research method in postgraduate nursing education. *Journal of Interdisciplinary Health Sciences, 21*, 120-128.

Terre Blanche, M., Durrheim, K., & Painter, D. (2006). *Research in practice: Applied methods for the social sciences* (2nd ed.). Cape Town: UCT Press.

The Cochrane Collaboration. (2012). *The Cochrane Collaboration*. Retrieved from http://www.cochrane.org

Torgerson, C. (2003). *Systematic reviews.* London: Continuum Press.

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management, 14*(3), 207-22.

Vaismarodi, M., Turunen, H., & Bondas, T. (2013). Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. *Nursing and Health Sciences, 15*, 398–405.

Urbina, S. (2014). *Essentials of psychological testing* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

van der Linden, W. J. (2005). *Linear models for optimal test design.* New York, NY: Springer.

van Dulm, O. (2013). Child language assessment and intervention in multilingual and multicultural South Africa: Findings of a national survey. *Stellenbosch Papers in Linguistics, 42*, 55-76.

Van Eeden, R. (1997). *Manual for the Senior South African Individual Scale – Revised (SSAIS-R): Background and standardisation.* Pretoria: Human Sciences Research Council.

Van Eeden, R., & 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.

Venter, D. Y. (2016). *Personality traits and self-presentation on Facebook: a systematic review.* Unpublished master's treatise, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa.

Wango, G. (2017). History and systems of psychology: Timelines in the development of contemporary psychology. *Journal of Psychology and Behavioral Science, 5*(2), 29-43.

Watters, C. (2017). *A pilot child psychology developmental clinic in Nairobi Kenya using Griffiths III*. Paper presented at the 17th International Scientific Meeting of the Association for Research in Infant and Child Development, Birmingham, United Kingdom.

Wei, H., & Lin, J. (2014). Using out-of-level items in computerized adaptive testing. *International Journal of Testing, 15,* 50-70.

Wechsler, D. (1955). *Manual for the Wechsler adult intelligence scale*. New York: Psychological Corporation.

Wechsler, D. (1981) *Manual for the Wechsler adult intelligence scale - revised*. New York: Psychological Corporation.

Wechsler, D. (1997). *WAIS-III administration and scoring manual*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2008). *WAIS-IV administration and scoring manual*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2012). *Wechsler Preschool and Primary Scales of Intelligence* (4th Edition). London: Pearson.

Wechsler, D. (2014). *Wechsler Intelligence Scale for Children* (5th Edition). Bloomington, MN: Pearson.

Weisz, G., Cambrosio, A., Keating, P., Schlich, T., & Tournay, V. J. (2007). The emergence of clinical practice guidelines. *The Milbank Quarterly, 85*(4), 696-727.

Wiberg, M, & von Davier, A. A. (2017). Examining the impact of covariates on anchor tests to ascertain quality over time in a college admissions test. *International Journal of Testing, 17*, 105–126.

Wiener, C. L. (2000). *The elusive quest: Accountability in hospitals*. Hawthorne, N.Y.: De Gruyter.

Williams, N. A. (2017). The national guideline clearinghouse. *Journal of Electronic Resources in Medical Libraries, 14*(2), 82-92.

Woolf, S., Schünemann, H. J., Eccles, M. P., Grimshaw, J. M., & Shekelle, P. (2012). Developing clinical practice guidelines: Types of evidence and outcomes; values and economics, synthesis, grading, and presentation and deriving recommendations. *Implementation Science, 7*, 61.

Wyse, A. E., & Albano, A. D. (2015). Considering the use of general and modified assessment items in computerized adaptive testing. *Applied Measurement in Education, 28*(2), 156-167.

Yu, C. H., & Ohlund, B. (2010). *Threats to validity of research design*. Retrieved from http://www.creative-wisdom.com/teaching/WBI/threat.shtml

Zhu, J., & Chen, H-Y. (2011). Utility of inferential norming with smaller sample sizes. *Journal of Psychoeducational Assessment, 29*(6), 570-580.

Zuiderent-Jerak, T., Forland, F., & Macbeth, F. (2012) Guidelines should reflect all knowledge, not just clinical trials. British Medical Journal, 345: e6702.

Appendix A: Output Classification Sheets

Output Classification Sheet

Output No: **1**

Reference: **Adams, K.M. (2000). Practical and ethical issues pertaining to test revisions. *Psychological Assessment, 12,* 281-286**

Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
		X		

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
	X			

Why was above option chosen?: The article was published in an American Psychological Association journal, an organisation which publishes psychometric guidelines.

Area(s) of investigation: Practical aspects of test revisions.

Further specifics of investigation:

Nr	Findings / Recommendations
1	Multiple conditions may dictate the need for test revision, including outdated material (p.282)
2	Economic returns of a test can become a salient factor in decisions about a test (p.282)
3	The forces of time and cost can affect a test revision process (p.283)
4	Multiple role players are included in a test revision, including subject-matter experts, psychometric experts, statisticians, financial managers, marketing managers, representatives of the corporate brand. The roles and responsibilities of these members need to delineated, balanced, and coordinated (p.283)
5	Revised tests have become increasingly expensive for test users. They need to budget for this increase… Test makers become frustrated by the unwillingness of test users to pay good money for quality revised products (p.283)
6	A revised test may include items, formats, and scoring systems of the previous version (p283)
7	During field-testing psychologists reactions to a revised test can be sourced to ascertain their likely degree of acceptance (p284)
8	Test publishers would do well to request comments from test users through popular trade publications (p285)

Included in developing guidelines? (Yes/No): **Yes**

Output Classification Sheet

Output No: **2**

Reference: **American Educational Research Association. (2014).** ***Standards for educational and psychological testing.*** **Washington, DC: AERA.**

Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
X				

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
X				

Why was above option chosen?: The document represents the official guidelines for members of the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education.

Area(s) of investigation: Broad-based psychometric best practice standards.

Further specifics of investigation:

Nr	Findings / Recommendations
1	1.9. "When validation rests in part on the opinions or decision of expert judges, observers, or raters, procedures for selecting such experts and eliciting judgments or ratings should be fully described. The qualifications and experience of the judges should be presented" (p.25).
2	2.19. "Each method of quantifying the reliability/precision of scores should be described clearly and expressed in terms of statistics appropriate to the method" (p.47) (including reporting the sampling procedures and test takers).
3	4.0. "…Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population" (p.85).
4	4.9. "When item or test form tryouts are conducted, the procedures used to select the sample (s) of test takers as well as the resulting characteristics of the sample(s) should be documented. The samples(s) should be as representative as possible of the populations s) for which the test is intended" (p.88).
5	4.10. "When a test developer evaluates the psychometric properties of items, the model used for that purpose (e.g., classical test theory, item response theory, or another model) should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are screened and the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF) for major examinee groups, should also be documented. When model-based methods (e.g., IRT) are used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented (p.88).

6	4.24. “Test specifications should be amended or revised when new research data, significant changes in the domain represented, or newly recommended condition of test use may reduce the validity of test score interpretations. Although a test remains useful need not be withdrawn or revised simply because of the passage of time, test developers and test publishers are responsible for monitoring changing conditions and for amending, revising, or withdrawing the test as indicated” (p.93).
7	4.25. “When tests are revised, users should be informed of the changes to the specifications of any adjustments made to the score scale, and of the degree of comparability of scores from the original and revised tests. Tests should be labeled as “revised” only when the test specifications have been updated in significant ways” (p.93).
8	7.14. “When substantial changes are made to a test, the test’s documentation should be amended, supplemented, or revised to keep information for users current and to provide useful additional information or cautions" (p.129).
9	10.1 “Those who use psychological tests should confine their testing and related assessment activities to their areas of competence, as demonstrated through education, training, experience, and appropriate credentials” (p.164).

Included in developing guidelines? (Yes/No): **Yes**

Output Classification Sheet

Output No: **3**

Reference: **Bush, S.S. (2010). Determining whether or when to adopt new versions of psychological and neuropsychological tests: Ethical and professional considerations. *The Clinical Neuropsychologist*, 24, 7-16.**

Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
		X		

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
		X		

Why was above option chosen?: The article was in a peer-reviewed journal, which is not affiliated to a specific organisation.

Area(s) of investigation: Aspects to consider when deciding on when and if to adopt a revised test.

Further specifics of investigation:

Nr	Findings / Recommendations
1	"Some tests undergo revisions, typically to improve their psychometric properties, normative data, relevance of stimuli, and ease of administration" (p.7).
2	"… it can take years after publication of a revised test for research with special patient populations to be performed and published" (p.7).
3	"… the profession of psychology has established a community standard regarding the transition to the newest test revision that ranges from 6 months to 1 year" (p.10).
4	"Neuropsychologists should request and obtain from test publishers the psychometric properties of the new version of the test to compare to the prior version. Test publishers can facilitate this process by (1) conducting such studies prior to publishing revisions of tests, and (2) providing charts containing such comparisons with their pre-release professional publications and promotional materials" (p.13).
5	"In addition to improving dialogue with test publishers, neuropsychologists…" (p.13).

Included in developing guidelines? (Yes/No): **Yes**

Output Classification Sheet

Output No: **4**

Reference: **Butcher, J.N. (2000). Revising psychological tests: Lessons learned from the revisions of the MMPI. *Psychological Assessment, 12*, 263-271.**

Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
		X		

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
	X			

Why was above option chosen?: The article was published in an American Psychological Association journal, an organisation which publishes psychometric guidelines.

Area(s) of investigation: Insights from test revision of the MMPI, leading to suggested practice guidelines.

Further specifics of investigation:

Nr	Findings / Recommendations
1	"…the goals and scope of the revision need to be carefully staked out before a revision is undertaken" (p.263).
2	"If a test is in wide use, there are likely to be more arguments against its revision and more resistance to change even though everything else around it has changed, resulting in an instrument that becomes even more out of date" (p.263).
3	"…that it takes a great deal of time and research effort to effect a successful revision and gain broad acceptance by the professional community" (p.264).
4	"The revised version of a psychological test must be a clear improvement over the earlier standard in order for it to be accepted… An instrument that falls short or simply meets the earlier standards would likely be unsuccessful" (p.264).
5	"Alterations of a major psychological test should not be based strictly on market forces or commercial interests, but should be based on a clear empirical justification and with a clearly thought-out rationale" (p.264).
6	"It is important to obtain broad and diverse input into needed changes from researchers and test users early in the revision planning" (p.264).
7	"Because altering traditional standards can result in conflictual decision-making situations, working relationships on a revision program can become strained. The revision team needs to have expertise to conduct the work, commitment to complete 1he job, and the authority to make critical decisions" (p.264).
8	"Settle issues such as arrangement of credit, work responsibility, and royalty arrangements before the revision gets underway" (p.264).

9	“Clearly establish the ties with the original version of the instrument to circumvent criticism that the revised version is not the same test as the original” (p.264).
10	“Changes in test stimuli or items require careful implementation study and need to be evaluated in pretest before they are implemented” (p.264).
11	“Choose the most generalizable normative approach. Avoid shortcuts that might make data collection easier but paint you into a comer in terms of user acceptability” (p.264).
12	“Provide empirical evidence on the validity of the revised measure” (p.264).
13	“Although diverse input needs to be solicited and weighed, revisers cannot allow small special interests to alter the course of the program revision strategy” (p.264).
14	“Develop a reasonable phase-out period for the superseded version, publicly advertise the end point, and stick to it” (p.264).
15	“Provide nationally based training programs and practically oriented workshops in the months leading up to and following the publication of the revision in order to assure that test users can obtain quick access to the revised instrument and incorporate the most recent version into their practice” (p.264).

Included in developing guidelines? (Yes/No): **Yes**

Output Classification Sheet

Output No: **5**

Reference: **Camara, W.J. (2007).** ***Standards for educational and psychological testing: Influence in assessment development and use.*** **Retrieved from http://www.teststandards.org/files/standards%20-%20influence%202007.pdf**

Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
			X	

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
		X		

Why was above option chosen?: Although the output is self-published, the author has published referenced works in peer-reviewed journal. He also mentions the College Board (USA) underneath the heading, which suggests that the paper may have been circulated within the institution.

Area(s) of investigation: Insights based on the 1999 Standards for educational and psychological testing.

Further specifics of investigation:

Nr	Findings / Recommendations
1	"Test developers are responsible for ensuring that assessment products and services meet applicable professional and technical standards and should be familiar with the Standards and other applicable requirements" (p.7).
2	Test developers "have a responsibility for providing technical documentation on their tests, including evidence of reliability and validity that supports inferences that will be made from test scores. Technical qualities for many educational tests also include construct representation…" (p.7).
3	"Evidence that differences in performance across major subgroups are related to the construct being measured and not due to construct irrelevant variance is also a professional responsibility of developers" (p.8).
4	"Sometimes the demand of test production may outstrip the resources of a test publisher and result in errors that may have been prevented with a more reasonable schedule" (p.8).
5	"…several instances where insufficient piloting and pretesting led to spurious results, and time schedules for accountability tests didn't allow for all the quality control procedures needed to detect and correct errors prior to test administration" (p.9).

Included in developing guidelines? (Yes/No): **Yes**

Output Classification Sheet

Output No: **6**

Reference: **European Federation of Psychologists' Associations (EFPA). (2013a). *EFPA review model for the description and evaluation of psychological and educational tests: Test review form and notes for reviewers* (Version 4.2.6). Retrieved from http://www.efpa.eu/download/650d0d4ecd407a51139ca44ee704fda4**

Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
	X			

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
X				

Why was above option chosen?: The output contains the criteria that are considered for classifying a psychological test within the European Federation, and as such is endorsed and applied by all countries in the European Union.

Area(s) of investigation: Benchmarks against which all psychological tests are measured in the European Union.

Further specifics of investigation:

Nr	Findings / Recommendations
1	How old are the normative studies? Inadequate, 20 years or older; Adequate, norms between 15 and 19 years old; Good, norms between 10 and 14 years old; Excellent, norms less than 10 years old (p.39).
2	"Thoroughness of the item analyses and item analysis model" (p.25).
3	"Norms: Clear and detailed information provided about sizes and sources of norms groups, representativeness, conditions of assessment etc." (p.27).
4	Comprehensive section on reliability and validity, including required statistical levels (pp.43-61).
5	"Summary of relevant research" (p.26).
6	Comprehensive presentation of content, construct, and criterion validity, with a range of studies (pp.26-27).
7	Theoretical foundations of constructs should be presented (p.26).
8	"Adequacy of documentation available to the user (user and technical manuals, norm supplements, etc.)" (p.28).

Included in developing guidelines? (Yes/No): **Yes**

Output Classification Sheet

Output No: **7**

Reference: **European Federation of Psychologists' Associations (EFPA). (2013b).** ***Performance Requirements, Context definitions and Knowledge & Skill specifications for the three EFPA levels of qualifications in psychological assessment.*** **Retrieved from http://www.efpa.eu/download/1b272a998e297c248413fbb761134697**

Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
X				

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
X				

Why was above option chosen?: Official EFPA standards for test users in the European Union.

Area(s) of investigation: Standards for psychological test users.

Further specifics of investigation:

Nr	Findings / Recommendations
1	"Establish that the constructs being measured are relevant for the assessment need" (p.6) [Standard 2.1 B]
2	"Keep up with relevant changes and advances relating to assessment methods and procedures that you use" (p.4) [Standard 1.2 D]
3	"Only offer assessment services, modes of administration and assessment methods and procedures for which you are qualified. B. Accept responsibility for the choice of assessment methods or procedures used, and for the recommendations made" (p.4). [Standards 1.3 A & B]

Included in developing guidelines? (Yes/No): **Yes**

Output Classification Sheet

Output No: **8**

Reference: **Education Testing Service (ETS). (2014).** ***ETS standards for quality and fairness.*** **Retrieved from www.ets.org.**

Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
X				

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
X				

Why was above option chosen?: Official standards from Educational Testing Service (USA).

Area(s) of investigation: Standards for quality in testing, and fairness to test takers.

Further specifics of investigation:

Nr	Findings / Recommendations
1	"Periodically review test specifications, active items, tests, and ancillary materials to verify that they continue to be appropriate and in compliance with current applicable guidelines. Revise materials as indicated by the reviews. Notify test takers and test users of changes that affect them" (p.32). [Standard 7.7]
2	"Obtain substantive advice and reviews from diverse internal and external sources, including clients and users, as appropriate" (p.12) [Standard 3.3]
3	"For a new or significantly revised product or service, provide a plan for addressing fairness in the design, development, administration, and use of the product or service. For an ongoing program, document what has been done to address fairness in the past as well as documenting any future fairness plans" (p.19) [Standard 5.1]
4	"Document and follow procedures designed to establish and maintain the technical quality, utility, and fairness of the product or service. For new products or services or for major revisions of existing ones, provide and follow a plan for establishing quality and fairness" (p.12) {Standard 3.2]
5	"If the intended use of a test has unintended, negative consequences, review the validity evidence to determine whether or not the negative consequences arise from construct-irrelevant sources of variance. If they do, revise the test to reduce, to the extent possible, the construct-irrelevant variance" (p.18). [Standard 4.5]
6	"When feasible, pretest items with test takers that represent, to the extent possible, the intended population for the test. Document the sampling process and the characteristics of the resulting sample. Document the statistical procedures used to evaluate the items and the results of the analyses, including, as appropriate, the fit of the model to the data. If pretesting

	is not practicable, use small-scale pilot tests, and/or collateral information about the items, and/or a preliminary item analysis after an operational administration but before scores or other test results are reported" (p.32). [Standard 7.5]
7	"Document the desired attributes of the test in detailed specifications and other test documentation. Document the rationales for major decisions about the test, and document the process used to develop the test. Document the qualifications of the ETS staff and external subject-matter experts involved in developing or reviewing the test" (p.29). [Standard 7.2]

Included in developing guidelines? (Yes/No): **Yes**

Output Classification Sheet

Output No: **9**

Reference: **Education Testing Service (ETS). (2009).** ***ETS international principles for fairness review of assessments: A manual for developing locally appropriate fairness review guidelines in various countries.*** **Retrieved from www.ets.org**

Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
	X			

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
	X			

Why was above option chosen?: Official publication from Educational Testing Service (USA).

Area(s) of investigation: How to assess fairness of tests.

Further specifics of investigation:

Nr	Findings / Recommendations
1	"You should select panelists who represent the important subgroups of the country's population. For example, if there are both male and female test takers, there should be both male and female panelists. If there are several official languages, members representing each language group among the people taking the test should be included. If there are significant differences among regions of the country, then representatives from each of the regions should be included. If there are different racial, ethnic, or religious groups within the country, then members of the various groups should be included on the panel to the extent possible, and so forth" (p.5).
2	"The primary purpose of fairness review is to identify invalid aspects of test items that might unfairly hinder people in various groups from demonstrating their relevant knowledge and skills" (p.7).
3	"Translation of test items without also accounting for cultural differences is a common source of construct-irrelevant knowledge. Translation alone may be insufficient for many test items, as shown by the example of an item that required knowledge of United States coins. The content of items must be adapted for the culture of the country in which the items will be used" (p11).
4	It is important to develop a comprehensive plan for how fairness will be strived for in a test, and to train reviewers to perform fairness reviews accurately and consistently.

Included in developing guidelines? (Yes/No): **Yes**

Output Classification Sheet

Output No: **10**

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

Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
		X		

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
		X		

Why was above option chosen?: The article was in a peer-reviewed journal, which is not affiliated to a specific organisation.

Area(s) of investigation: Developing a test for a multicultural population

Further specifics of investigation:

Nr	Findings / Recommendations
1	"…test developers need to grapple with basic issues such as what methods of test administration might be appropriate or inappropriate for certain cultural groups and what language to develop the test in..." (p.8)
2	"Readers should take note that although the aspects of the plan are logically ordered, there is a dynamic interplay between the various aspects that will often result in the developer revising a decision made about one or other aspect" (p.9).
3	"From a cross-cultural perspective, the panel of experts who develop curriculum frameworks and learning outcomes or who perform job analyses should represent a mix of cultures" (p.12).
4	"Throughout this section it should be clear that test developers should be sensitive to the fact that their choice of presentation mode, item format, and response mode, represent potential sources of construct-irrelevant variance" (p.13).
5	"It is thus recommended that a multicultural test development team be assembled that demonstrates a rich mix of cultural and language groups and test development expertise" (p.14).

Included in developing guidelines? (Yes/No): **Yes**

Output Classification Sheet

Output No: **11**

Reference: **Geisinger, K.F. (Ed.). (2013).** ***APA handbook of testing and assessment in school psychology and education*** **(Volume 3). Washington, DC: APA.**

Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
	X			

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
	X			

Why was above option chosen?: The handbook is an official publication from the American Psychological Association.
Area(s) of investigation: All aspects of psychological testing.
Further specifics of investigation:

Nr	Findings / Recommendations
1	"A score-equating plan that links a new form to multiple old norms is preferable to a plan with a link to a single old form" The preference is for a common anchor block which will assist in linking test forms, considering the relative difficulty of the two tests, and the relative ability of different groups. Problems in linking test scores across forms can be overcome by using a common set of examinees, and the other is to use a common set of items (p.507).
2	"The use of multiple-language versions of tests is not only desirable but necessary for many tests that involve individuals from different languages and cultures… tests are often adapted to many different languages to provide valid measurement and minimise bias". Adaptation of tests is also necessary for use in other cultures and countries and part of cross-cultural research and international comparisons" (p.545).
3	Original version is referred to as the source version, and the adapted test as the target version (p.545).
4	With the ever-widening and deepening of diagnoses, it has become difficult to include studies on all special groups in a test manual. "Most recently published test manuals have instead provided at least a little information about test scores of children in various special groups, usually compared with matched samples of children without disabilities. The samples used in these studies are usually small, so examiners must wait for larger studies to appear in the literature, but the data do serve to demonstrate the test's validity for differentiating various groups of children…" (p.47).

Included in developing guidelines? (Yes/No): **Yes**

Output Classification Sheet

Output No: **12**

Reference: **International Test Commission. (2016).** ***ITC Guidelines for translating and adapting tests*** **(2nd Ed.). Retrieved from http://www.intestcom.org/**

Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
X				

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
X				

Why was above option chosen?: Official guidelines from the International Test Commission for their members.
Area(s) of investigation: Test adaptation and translation
Further specifics of investigation:

Nr	Findings / Recommendations
1	"TD-1 (4) Ensure that the translation and adaptation processes consider linguistic, psychological, and cultural differences in the intended populations through the choice of experts with relevant expertise" (p.11).
2	"PC-3 (3) Minimise the influence of any cultural and linguistic differences that are irrelevant to the intended uses of the test in the populations of interest" (p.36).
3	"TD-3 (6) Provide evidence that the test instructions and item content have similar meaning for all intended populations" (p.14).
4	"TD-5 (8) Collect pilot data on the adapted test to enable item analysis, reliability assessment and small-scale validity studies so that any necessary revisions to the adapted test can be made" (p.16).
5	"C-1 (9) Select sample with characteristics that are relevant for the intended use of the test and of sufficient size and relevance for the empirical analyses" (p.17).
6	"C-3 (11) Provide evidence supporting the norms, reliability and validity of the adapted version of the test in the intended populations" (p.23).
7	"Doc-1 (17) Provide technical documentation of any changes, including an account of the evidence obtained to support equivalence, when a test is adapted for use in another population" (p.28).
8	A-1 (13) Prepare administration materials and instructions to minimise any culture- and language-related problems that are caused by administration procedures and response modes that can affect the validity of the inferences drawn from the scores" (p.25).

Included in developing guidelines? (Yes/No): **Yes**

Output Classification Sheet

Output No: **13**

Reference: **International Test Commission. (2015).** ***Guidelines for practitioner use of test revisions, obsolete tests, and test disposal.*** **Retrieved from https://www.intestcom.org/files/guideline_test_disposal.pdf**

Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
X				

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
X				

Why was above option chosen?: Official guidelines from the International Test Commission for their members.
Area(s) of investigation: Dealing with test revisions, test disposal and obsolete tests.
Further specifics of investigation:

Nr	Findings / Recommendations
1	"Test Publishers Shall Describe and Justify the Need for the Revised Test" (p.9).
2	"Test Developers and Users Have a Reciprocal Relationship" (p.8).
3	"Financial Considerations Influence Adoption of Revised Tests" (p.11).
4	"Test Selection Decisions Should Be Based on Evidence Regarding the Scientific Merits of the Revised Version" (p.9).
5	"Test Selection Shall Be Based, In Part, On a Review of Changes Made In the Revision" (p.9).
6	"Practitioners Should Obtain Training in the Use of the Revised Test" (p.9).
7	"Practitioners Shall Not Justify the Use of an Older Version Due to Their Personal Attachment" (p.10).
8	"Test Developers Should Provide Evidence of the Revised Test's Validity" (p.12).
9	"Test Selection Shall Consider the Revised Test's Reliability" (p.12).
10	"Test Selection Should Consider the Correspondence between Prior and New Norms and Their Possible Impact" (p.12).
11	"Practitioners Should Consider Non-financial Considerations When Adoption a Revised Test" (p.11).

Included in developing guidelines? (Yes/No): **Yes**

Output Classification Sheet

Output No: **14**
Reference: **International Test Commission. (2013a). *ITC guidelines on test use.* Retrieved from http://www.intestcom.org/files/guideline_test_use.pdf**
Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
X				

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
X				

Why was above option chosen?: Official guidelines from the International Test Commission for their members.
Area(s) of investigation: Test use.
Further specifics of investigation:

Nr	Findings / Recommendations
1	"The tests are unbiased and appropriate for the various groups that will be tested. The constructs being assessed are meaningful in each of the groups represented. Effects of group differences not relevant to the main purpose (e.g., differences in motivation to answer, or reading ability) are minimised" (p.17).
2	"Monitor and periodically review changes over time in the populations of individuals being tested and any criterion measures being used" (p.22).
3	"The developers have been sensitive to issues of content, culture and language" (p.17).
4	"Use appropriate norm or comparison groups where available" (p.21).
5	"Consider each scale's reliability, error of measurement and other qualities which may have artificially lowered or raised results when interpreting scores" (p.21).
6	"Avoid the use of tests that have inadequate or unclear supporting technical documentation. Use tests only for those purposes where relevant and appropriate validity evidence is available" (p.17).
7	"Determine that the test's technical and user documentation provides sufficient information to enable evaluation" (p.16).
8	"Only offer testing services and only use tests for which they are qualified" (p.15).
9	"Keep up with relevant changes and advances relating to the tests they use, and to test development, including changes in legislation and policy, which may impact on tests and test use" (p.14).
10	"Be aware of the need to re-evaluate the use of a test if changes are made to its form, content, or mode of administration" (p.22).
11	"Be aware of the need to re-evaluate the evidence of validity if the purpose for which a test is being used is changed. Where possible, seek to validate tests for the use to which they are being put, or participate in formal validation studies. Where possible, assist in updating information regarding the norms, reliability and validity of the test by providing relevant test data to the test developers, publishers or researchers" (p.23).

Included in developing guidelines? (Yes/No): **Yes**

Output Classification Sheet

Output No: **15**

Reference: **International Test Commission. (2013b).** ***ITC Guidelines on quality control in scoring, test analysis, and reporting of test scores.*** **Retrieved from https://www.intestcom.org/files/guideline_quality_control.pdf**

Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
X				

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
X				

Why was above option chosen?: Official guidelines from the International Test Commission for their members.

Area(s) of investigation: Test analysis, quality control of the testing proses.

Further specifics of investigation:

Nr	Findings / Recommendations
1	"Agree upon who has final responsibility and authority to decide how to proceed when problems occur and how to resolve them" (p.13).
2	"Agree in advance which member of staff is responsible for each stage" (p.14)
3	"Monitoring should be carried out in collaboration with all stakeholders, with the aim of auditing specific processes, for example, monitoring inter-rater reliability and checking data entry error rates" (p.14).
4	"Document all activities. Use standard check sheets to show that each process has been carried out and checked off accordingly" (p.14).
5	"Conduct item analysis after the test is administered or analyze accumulated data (e.g., within 3-5 years of administration) if the test is given periodically. Consider performing item analysis on partial data, (before full data is available), so you can identify errors quickly" (p.19).
6	"In the technical manual or associated materials, give a detailed description of the procedures used to convert raw scores to standardized scores. Because this technique may be different for different test forms, the procedure should be described for each test form" (p.22)
7	"Account for changes that occur in the scale over time" (p.22)
8	"Advise other professionals of mistakes in an appropriate and timely manner, sometimes in a special meeting devoted to error prevention. Document how to prevent future mistakes or errors" (p.15).

9	"Conduct studies to determine the expected correlation between background data and scores, look for inconsistencies in the patterns of scores in the current data with respect to other information – previous data sets, research findings etc" (p.16). "Compare sample data to the range of scores that can be expected, and compare descriptive statistics to the test publishers' norms, if those are provided. Sample statistics can be expected to deviate somewhat (beyond what is expected by sampling error variance), but large effect size differences should be noted and potentially investigated" (p.17) "For high-stake tests make every effort to replicate equating results independently and involve a third party external to the equating process" (p.21).

Included in developing guidelines? (Yes/No): **Yes**

Output Classification Sheet

Output No: **16**

Reference: **King, M.C. (2006). Adopting revised versions of psychological tests. *The CAP Monitor, 23*, 6-7.**

Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
		X		

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
			X	

Why was above option chosen?: The article was in a local organisation's newsletter, and would probably not have been extensively peer-reviewed prior to publication.
Area(s) of investigation: Adopting revised tests.
Further specifics of investigation:

Nr	Findings / Recommendations
1	"Some require revision when the trait or ability they purport to measure has changed… Other tests require revising or updating as the cognitive science underlying them advances… Still other instruments require revision to address problems with the original test construction and norming, the "aging" of test content, and to reflect advances in measurement and statistical analyses of their data" (p.6)
2	"Where psychological theory, data, or methods have pushed tests past their best-before dates, psychologists should adopt the new instruments in the service of client care or show cause why they have not done so" (p.6). "Some inadequate justifications for continuing to use outdated or obsolete test instruments • I've still got hundreds of test forms from the old version in my office. • I'm just more comfortable with the old version. • The new version is too expensive. • I don't have time to learn how to use / interpret the new version. • I refuse to pay (a test publisher) all this money every time they feel like changing one of their tests and putting it in a new carrying case" (p.7).
3	"In the case of assessment services, this standard appears to require adoption of the new versions of test instruments when it is clear that those new instruments represent an advance over their predecessors" (p.6).
4	"Second, tests may be applied to populations (for example, distinct cultural, language, or age groups) for which appropriate normative data may not yet have been acquired in the test revision process. In such cases, it might be appropriate to consider using previous test versions or norms, with the appropriate cautions, until more appropriate data are available" (p.6).

Included in developing guidelines? (Yes/No): **Yes**

Output Classification Sheet

Output No: **17**

Reference: **Liu, J., & Dorans, N.J. (2013). Assessing a critical aspect of construct continuity when test specifications change or test forms deviate from specifications. *Educational Measurement: Issues and Practice, 32*, 15-22.**

Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
		X		

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
		X		

Why was above option chosen?: The article was in a peer-reviewed journal, which is not affiliated to a specific organisation.
Area(s) of investigation: Changes to tests over time. Equating scores.
Further specifics of investigation:

Nr	Findings / Recommendations
1	"planned changes…including topic coverage, item format, item type, test length, the relative emphasis given to different aspects of the measured domain, and the addition of a new measure to the existing test battery" (p.15).
2	"A testing program needs to be responsibly responsive to the environment in which it operates" (p.15).
3	"Score equity assessment (SEA) can be used as a tool to assess a critical aspect of construct continuity, the equivalence of scores, whenever planned changes are introduced to testing programs" (p.15).
4	"Possible factors included changes in test difficulty, the test admission process, public perception of testing, the test-taking population, patterns of test preparation, face validity, cost, fairness, and scaling constraints" (p.15).
5	"Test assembly processes may not strictly follow explicit blueprints. The content and difficulty composition of the item pool are not constant: Some content areas are easier to replenish than other areas. Very easy and very hard items may become scarce over time while middle difficulty items continue to abound" (p.15).
6	"Innovations in education, curriculum, and technology may lead to changes in tests that reflect contemporary school curricula, reinforce educational standards and practice, and maintain test fairness for an increasingly diverse test-taking population" (p.15).

Included in developing guidelines? (Yes/No): **Yes**

Output Classification Sheet

Output No: **18**

Reference: **Mattern, K.D., Kobrin, J.L., & Camara, W.J. (2012). Promoting rigorous validation practice: An applied perspective.** ***Measurement: Interdisciplinary Research and Perspectives, 10*****(1-2), 88-92.**

Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
		X		

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
		X		

Why was above option chosen?: The article was in a peer-reviewed journal, which is not affiliated to a specific organisation.
Area(s) of investigation: Issues around test validity.
Further specifics of investigation:

Nr	Findings / Recommendations
1	"While test sponsors, publishers, and developers may strive to adhere to the Standards, competing demands often pose challenges to strict compliance, and shortcuts may be taken" (p.88). "…bigger threats contributing to "impoverished validation practice" emerge from the absence of adequate criterion measures and data, a lack of monetary and personnel resources, and business concerns that could influence the type and quality of validity research" (p.89). "We encourage our profession to focus on methods that can inform test developers and users how to develop and produce validation evidence rather than perfecting the definition of validity. This work can be improved by publishing best practices and specific examples that illustrate the strengths and weaknesses of different lines of validity evidence, show how to prioritize those studies, and offer guidance in determining the appropriate uses of assessment results based on sufficient evidence to support those interpretations or uses" (p.90).
2	"…a committee could develop a template or checklist that would guide test developers and users in assembling the types of evidence required to support each separate and distinct intended use of an assessment" (p.91).

Included in developing guidelines? (Yes/No): **Yes**

Output Classification Sheet

Output No: **19**

Reference: **Naglieri, J.A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological Testing on the Internet: New Problems, Old Issues.** ***American Psychologist, 59*****(3), 150-162.**

Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
		X		

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
		X		

Why was above option chosen?: The article was in a peer-reviewed journal, which is not affiliated to a specific organisation.

Area(s) of investigation: Test publisher responsibilities to maintain test information post-launch.

Further specifics of investigation:

Nr	Findings / Recommendations
1	"Other issues…include access of tests by individuals who are not qualified to administer such tests and interpret the results; and tests that have been modified, changed, or translated without appropriate permission or validation" (p.3).
2	"For example, revising a paper-and-pencil test requires printing and distributing new test forms and answer keys and printing new or revised test manuals, an expensive process that may take several months or years" (p.7)
3	"It would be unethical to develop new measurement tools that cannot be held to existing psychometric standards…without providing arguments and evidence for new or revised standards" (p.56).
4	"However, publishers and authors must scan the web for whole and partial elements of tests that require professional training for administration or interpretation...publishers must also protect their copyrights on test materials" (p.58).
5	"Further, it is quite easy for web publishers to forget about published pages on the Internet that may be updated in different places, yet the old materials remain available to the public" (p.59).

Included in developing guidelines? (Yes/No): **Yes**

Output Classification Sheet

Output No: **20**

Reference: **Oliveri, M.E., Lawless, R., & Young, J.W. (2015).** ***A validity framework for the use and development of exported assessments.*** **Princeton, NJ: Educational Testing Service.**

Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
	X			

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
	X			

Why was above option chosen?: Official publication of Educational Testing Service (USA).

Area(s) of investigation:

Further specifics of investigation:

Nr	Findings / Recommendations
1	"In defining the domain, the claim is that test takers' performances provide valid evidence of the assessed construct without the introduction of sources of construct-irrelevant variance" (p.9). "The warrant in defining the domain posits that assessments administered to multiple populations are construct-relevant for the examined domain. The presence of construct-irrelevant factors undermines our ability to make valid score-based inferences and is important because tests containing high cultural loads, meaning that items on the test require specific knowledge of, or experience with, mainstream culture, also may be assessing test takers' level of acculturation or learning of the culture(s) in which the person is expected to demonstrate competence in addition to the assessed construct" (p.11).
2	"The analysis of field-test data is an important step in test construction, particularly in relation to administering tests with multiple test-taker populations. Field-testing helps identify items that may contain construct-irrelevant factors due to cultural or linguistic differences…Results from field-test analyses can inform decisions on whether particular items should be retained or discarded from the assessment… or modified… Field-test results are also useful for monitoring how equating/linking items perform (if these were included in the assessment). Using the field-test results, DIF analyses can be conducted if the items are administered to a large enough sample of examines (around 300 from each population) and can detect whether the items function similarly across the intended populations (p.17).
3	"…generalization involves examining whether the configuration of the tasks is appropriate for the intended score interpretations and whether the test has a sufficient number of tasks to demonstrate the test takers' knowledge, skills, and abilities (KSAs) in the construct(s) of interest" (p.20).

4	"…samples of test takers should be drawn from the various groups to which the test will be administered and efforts must be made to include them in the universe of generalization" (p.20).
5	"Thus, some of the issues of developing and using multilingual international assessments may be related, but not limited, to challenges in translation and adaptation" (p.5). "Expert reviews need to be carefully planned and implemented by selecting experts familiar with the original test population and targeted population. They should also be familiar with the nuances of the language of the test, able to recognize the features of the language that may be problematic, and be sensitive to text that may be specific to a particular culture to ensure that the tests are fair for the new population(s)" (p.13).

Included in developing guidelines? (Yes/No): **Yes**

Output Classification Sheet

Output No: **21**

Reference: **Strauss, E., Spreen, O., & Hunter, M. (2000). Implications of test revisions for research. *Psychological Assessment, 12,* 237-244.**

Peer-review / organisational authority of research output (mark with X):

Organisational standard / guideline	Organisational publication	Peer-reviewed journal publication	Self-published output	Unknown
		X		

Possible level of support (institutional or from industry sources) for guidelines provided:

Very High	High	Medium	Low	Very Low
	X			

Why was above option chosen?: The article was published in an American Psychological Association journal, an organisation which publishes psychometric guidelines.

Area(s) of investigation: Research on test revisions.

Further specifics of investigation:

Nr	Findings / Recommendations
1	"There are good reasons for revising instruments. These include updating norms, providing age extension, additional sampling of minorities, removing dated items, and improving item effectiveness as well as test validity. Other revisions include translations of tests into other languages or a change to a computerized format" (p.237).
2	"These gains may result from improved nutrition, cultural changes, experience with testing, changes in schooling or child-rearing practices, or other factors as yet unknown" (p.237).
3	"Understanding how revisions compare with previous measures is important because failure to do so may lead to misleading results" (p.237). "A related implication of the Flynn effect is that one cannot directly equate performances on earlier and later versions of the test" (p.238). "Another method to equate performance involves converting raw scores on different versions of an instrument to T scores corrected for age, education, and gender" (p.238).
4	"The relationships of the subtest scores to one another and of test scores to other measures of cognitive functioning are also of interest" (p.238). "It is not appropriate to rely on the apparent simi1arity of two tests without actually checking them out" (p.240). "When comparing old and new versions, avoid global scores if factor structures are not equivalent and restrict comparison to those components that appear equivalent on both versions" (p.243).
5	"A special case of test revision is the conversion of standard tests into computerized versions, which has been occurring more frequently in recent years" (p.242).

6	"Use the revised version when there is evidence of a substantial normative shift (that is, a large Flynn effect); If the revised version has succeeded in measuring new and important constructs, then the choice is clear; If it is important to maintain continuity with the previous literature, then use the older version. However, if new normative data are available then the new norms should be consulted for interpretation; Use the same version when serial testing is an issue; Avoid decision rules based on different test revisions"

Included in developing guidelines? (Yes/No): **Yes**

Appendix B: Summarising Map

Theme		Document Number																				
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	Reasons for revising a test																					
1.1	• Factors internal to the test	X		X					X													X
1.2	• Factors external to the test	X	X		X	X	X		X						X		X	X		X		X
2	Role players in test revision	X	X		X					X	X					X						
3	Revision planning																					
3.1	• Revision scope and process	X			X	X					X					X		X		X		
3.2	• Post-launch activities				X											X			X			
4	Relationship between test editions	X			X							X		X		X		X				X
5	Test item development		X		X		X		X				X								X	
6	Norm development approach		X		X		X					X	X		X						X	
7	Test validity and reliability		X	X	X	X	X						X	X	X							X
8	Fairness of test results across different groups		X			X			X	X	X	X	X		X						X	
9	Test Users																					
9.1	• Information to test users		X	X			X		X					X		X				X		
9.2	• Test user feedback	X			X				X													
9.3	• Test user responsibility		X					X						X	X							
9.4	• Adopting revised tests	X		X	X									X			X					X

Appendix C: Letter of Invitation to Participants

NELSON MANDELA
UNIVERSITY

Good day

Invitation to participate in the PhD study of Mr Johan Herman Cronje

You are being asked to participate in a research study as part of a team of national and international reviewers of a guidelines document for the revision of psychological tests. I am a PhD student registered within the Department of Psychology at the Nelson Mandela University. The title of my study is "The development of a set of guidelines for the revision of psychological tests and the use of revised psychological tests". One of the objectives of this study is to develop guidelines.
You have been identified as a professional with experience in the field of assessment and/or test revision. I am inviting you to comment on the 30 guidelines that were developed. The guideline document is 15 pages in length, and your involvement in the study would be to provide constructive written feedback on the guidelines. All feedback will be considered in consultation with the promotors of the study, Prof MB Watson and Prof L Stroud, and used to refine the guidelines.
Please note that:

- This study has been granted ethics approval by the Faculty of Health Sciences Postgraduate Studies Committee of the Nelson Mandela University (Ref: H15-HEA-PSY-012).
- Your participation in this study is voluntary, and you may withdraw your participation at any time.
- You have the right to approach me with questions or concerns regarding the study at any time.
- Although your identity will at all times remain confidential, the results of the research study may be presented at scientific conferences or in specialist publications.

This informed consent statement has been prepared in compliance with current statutory guidelines.
I would appreciate your input on the guidelines. Please email general comments to me by **Wednesday, 28 February 2018.** If you wish to add comments within the document, please do so, and attach the document in your email to me.

Sincere regards,

Johan Cronje

Lecturer: Department of Psychology
Tel: +27 (0)41 504 2334
Johan.cronje@mandela.ac.za

Appendix D: Ethics Permission Letter

Nelson Mandela Metropolitan University
for tomorrow

Copies to:
Supervisor: Prof MB Watson
Co-Supervisor: Dr DJL Venter

Summerstrand South
Faculty of Health Sciences
Tel. +27 (0)41 504 2956 Fax. +27 (0)41 504 9324
marilyn.afrikaner@nmmu.ac.za

Student number: 194112520

7 July 2015

MR J CRONJE
15 ANNA STREET
UITENHAGE
6229

RE: OUTCOME OF PROPOSAL SUBMISSION

QUALIFICATION: PHD (PSYCHOLOGY)
FINAL RESEARCH/PROJECT PROPOSAL:
THE DEVELOPMENT OF A SET OF GUIDELINES FOR THE REVISION OF PSYCHOLOGICAL TESTS

Please be advised that your final research proposal was approved by the Faculty Postgraduate Studies Committee (FPGSC) subject to the following amendments/recommendations being made to the satisfaction of your Supervisors:

COMMENTS/RECOMMENDATIONS:

1. The proposal was well prepared and of a high standard.
2. Cover page
   Change the name of the degree to the English version.
3. Page 17 under objectives
   The researcher refers to a dearth of published guidelines. A systematic review is done on an abundance of publications. This statement was confusing. Rephrase.
4. The work and time schedule was unrealistic (less than two years was indicated to complete the study).
5. Output classification sheet
   The researcher did not indicate how he will assess the quality of the documents that will be reviewed. This could lead to non-quality conclusions.

Faculty Postgraduate Studies Committee (FPGSC) reference number: **H15-HEA-PSY-012.**

Please be informed that this is a summary of deliberations that you must discuss with your Supervisors and make the necessary amendments.

Please forward a final electronic copy of your appendices, proposal and REC-H form to the Faculty Postgraduate Studies Committee (FPGSC) secretariat.

Appendix E: Guideline Document for Test Revision and the Use of Revised Tests

Guidelines for the Revision of Psychological Tests and the Use of Revised Psychological Tests

Introduction

The guidelines below were developed through a systematic review of existing standards and guidelines of authors and organisations published from 2000-2017. In all, 20 original sources were included in the systematic review. The guidelines form part of a PhD study. The author is a lecturer in psychology in the subject areas of research methodology, psychometrics, and data analysis. The guidelines are intended for practitioners embarking on a revision process of a psychological test, and for users of revised psychological tests. Thirty guidelines were developed across ten phases of test revision conceptualised by the author. Each guideline starts with a broad topic statement that is explained in greater detail in the text that follows it. The guidelines are placed according to the content area they relate to, but as they convey overarching themes, there may be some overlap or repetition of themes throughout the explanatory texts.

The following acronyms are used for the purpose of referencing in the document:

AERA	American Educational Research Association
EFPA	European Federation of Psychologists' Associations
ETS	Education Testing Service
ITC	International Test Commission

Phase One: Pre-Planning

1.1 Test revisions should endeavour to improve the quality, utility, accuracy and fairness of a test. During a test revision, revision teams have to take cognisance of preceding versions. This can be both a benefit and a challenge. The benefit is that from previous versions a body of research evidence and feedback from test users as well as the test market in general is available to provide insight into areas of the test that may be improved on. This includes expansion in the use of the test beyond the original test taker market, and the need to consider new markets in the content, materials, and standardisation sample and norms of the revised test.

There are different aspects that can be revised in a test. If a considerable period has lapsed between revisions, improvements may include refinements to the underpinning construct of the test, the relevance of stimuli, and normative information (Bush, 2010). Expanded reasons for revision may include an extension of the age range of the test population, broadening of the intended test population in terms of ethnic, cultural or language groups, improved accuracy of the test, and alternative forms of administration, scoring and reporting, including fixed computer-based or internet delivered modes (Strauss, Spreen, & Hunter, 2000).

One challenge in the existence of a previous test version is that it creates an immediate benchmark against which the revised version will be measured. Revision teams need to take cognisance of both the benefit and challenge of having previous test versions. Ultimately, whatever changes are made during a revision must be shown to be a clear improvement on

earlier versions of a test. A revised test that fails to demonstrate this advancement will likely be poorly received by the market and consequently be unsuccessful (Butcher, 2000). The primary aim should always be to deliver the highest quality test for the ultimate benefit of test takers.

1.2 Revision teams should consist of a mix of internal and external stakeholders of the test. Test revision should be steered by a dedicated, core project committee, called a revision team. As such, the revision team will reflect the existing academic and economic aspects of a test. The revision team would consist therefore of multiple role players, including experts in the subject matter, the field of psychological testing and statistics, as well as representatives from the test publisher involved in test marketing and financial management (Adams, 2000). It is also important that revision teams reflect the rich mix of subgroups within the test's intended population, including different racial, ethnic, and language groups, as well as gender (ETS, 2009; Foxcroft, 2004). These representatives should have expertise that will allow them to represent the linguistic and cultural differences of the intended test population (ITC, 2016). When aspects of the revision rest on the opinions or decisions of experts, the process of selecting those experts, their relevant areas of expertise and experience, together with their qualifications should be documented (AERA, 2014; ETS, 2014).

A revision team should create a vision and mission for the project at the outset, against which all project decisions can be measured (Butcher, 2000). To maintain forward movement on the project it is also important that the roles and responsibilities of revision team members are agreed on in advance (ITC, 2013b). The revision team should agree at the outset on who has final responsibility for completing different tasks, and who has final authority on project decisions. If decision-making power resides with the revision team, then it should be stipulated by which majority vote decisions will carry. Issues related to credit for different components of the revision, and arrangements regarding financial matters, such as salaries, stipends, or royalties should be finalised before the project commences, as these may become sources of conflict at later stages (Butcher, 2000). Above all, the main criteria for project members are their expertise and willingness to perform tasks, their commitment to the test, their decision-making ability, and a commitment towards engaging in a collaborative effort intended to promote the welfare of test takers (Butcher, 2000).

Phase Two: Initial Investigation

2.1 Test publishers are responsible for monitoring the context within which tests operate, including the use of and feedback about tests, and the industry requirements for psychological tests, as this information may inform the decisions of revision teams. Tests operate in a dynamic and changing environment. Test publishers have a responsibility to monitor changes in test conditions and the use of their test products (AERA, 2014). They should be proactive and maintain a responsive attitude (Liu & Dorans, 2013). This includes a variety of actions. Publishers should familiarise themselves with the professional and technical industry standards that apply to tests (Camara, 2007). Test specifications, items, materials, and publications should be reviewed periodically to ensure that their products meet the required standards (ETS, 2014). Changes to industry standards may require a publisher to revise a test to align it with the updated standards.

If significant test information or content has been published within the public domain, it may challenge test validity, which will require test revision earlier than anticipated. Publishers

should protect test security therefore by enforcing their copyright of test materials. As part of this process, the internet must be scanned for complete or partial test components that require professional training and registration, as test use by unauthorised persons may diminish test security and cause harm to test takers (Naglieri et al., 2004).

Publishers should be proactive in seeking feedback from test users and researchers by inviting comments through trade publications and test user forums (Adams, 2000). In the event that any changes to the use of a test are made, test users should be informed of the changes that affect them (ETS, 2014). This includes the intention to embark on a new revision, as this may affect test users at a future date.

2.2 A test should be revised or withdrawn when new research data, significant changes in the test domain, or altered conditions of test use may affect the validity of test score interpretations. It can be challenging to choose the correct moment to revise a test. Important cues can come from research findings, changes in the domain of the test, and amended conditions to the use of tests that are implemented by external bodies.

An important cue is when critical test components have become outdated (Adams, 2000). A key indicator that this has occurred is changes to the theoretical framework that underpins the test. In addition to this, advances in measurement theory, psychological testing practice, and norm development are also important considerations (King, 2006). Changes in the intended test population over time may also necessitate a change. For some tests, a shift in popular culture may date some test items. For tests that challenge, and that rely on the difficulty of individual items, changes in test performance, such as the Flynn effect, may reduce the overall difficulty of tests. Improved nutrition, health care, child-rearing practices, and education have been mentioned as possible causes for the Flynn effect (Strauss, Spreen, & Hunter, 2000). This is especially pertinent for developing countries. Concern about the effects of time on the validity of interpretations of test results is evident in industry publications. For instance, the EFPA label a test as inadequate if the normative and standardisation information is 20 years or older (2013a). To obtain a rating of ‘Excellent’, such information should be less than a decade old. Requirements such as these are intended to inform test users, as well as publishers who should remain cognisant of changes to important industry standards and benchmark their products against them.

Furthermore, changes in test taker behaviour and performance over time should be scrutinised, as these would affect the perceived value of the product by potential users, and the face validity of tests by test takers (Liu & Dorans, 2013). Sometimes, a test that has reached its expiration date due to outdated norms may still be used meaningfully in already existing longitudinal research. Such a test may also be preferable for certain clinical groups if it is supported by adequate research. However the proviso is that such a test should not be used for decision-making based on its norms, but rather for its qualitative value as one part of a comprehensive portfolio of evidence (AERA, 2014). Revision teams should define the period for which outdated tests may still be used for these purposes and inform test users.

The above aspects highlight the intricate climate of time, culture, and context within which tests are used. Revision teams should take heed of the impact of different factors to determine the correct moment to embark on a test revision.

2.3 During a test revision, feedback should be obtained from diverse internal and external sources, including test users and test takers. It is important to gather feedback from test users and researchers early in the project regarding changes that are required in the test. This

should be a broad consultation of both users of the test as well as those who do not use the test, information from experts in the test's subject matter, and experts in the theory and practice of psychological testing. The purpose of such consultations is to update the test revision team on current knowledge within the subject field, and opinions from the test user market (Butcher, 2000). Requesting input serves multiple functions. Firstly, it recognises and values the experience of test users and makes them feel included in the revision. Secondly, it allows for identification of latent experts on the test, who may be drawn during later phases of the revision project (ITC, 2013b). Thirdly, it creates a sense of collaboration between the revision committee and test users. Finally, it creates a database of interested users and researchers who may be approached later to review the revised test and to provide feedback on the likely acceptance of the product by the broader market (Adams, 2000; ETS, 2014).

Phase Three: Project Planning

3.1 Revision teams should provide a plan to address fairness in the design, development, administration, and use of a revised test. A psychological test is a controlled, standardised observation. Its ultimate goal is to measure a construct or set of constructs accurately and fairly, without any interference from sources that are not integrally linked to the construct(s). Revision teams should consider measures to promote the fairness of their tests. The intended changes of a test revision should include therefore plans to improve fairness and accuracy (ETS, 2009). A suggested starting point would be to document historical actions of previous test versions to address fairness (ETS, 2014). Creating such an overview will set a trajectory for the current test revision, by highlighting strengths and potential gaps in previous test editions. It will similarly provide insights into the specific actions that were more successful in increasing fairness, as well as those that were less helpful. It may also assist in constructing future fairness plans. For a current test revision, the measures taken to improve fairness, validity and reliability, including the analyses used and results thereof should be documented (AERA, 2014). Revision teams are ethically obliged to represent the level of fairness in a revised test accurately, including its potential shortcomings in this regard for specific populations.

3.2 The rationale, goals, scope, and process of a test revision should be planned, followed and documented. Test revision takes considerable planning and effort. Without appropriate management, it has the potential to veer off course (Butcher, 2000). It is essential to delineate the goals and scope of the project at the outset to act as a compass. Each step in the process should be documented to demonstrate how technical quality has been achieved (ITC, 2013b). The rationale for major decisions about the current test revision should also be explained in detail, as these will be important for existing users of previous versions of the test, as well as for future revisions of the test (ETS, 2014).

3.3 Revision teams should consider constraints in terms of time, cost, and resources when designing a test revision. Test revision can be an expensive process that may take years to complete (Naglieri et al., 2004). Projects are limited however by the extent of available resources. The timeframe and budget for the test revision should be considered early in the project, as these will have implications as to the extent of the project and the quality of the final product (Adams, 2000). The demands of the eventual project may outstrip the resources initially allocated, which can result in errors that could have been avoided (Camara, 2007). It is therefore advisable that revision teams have a realistic concept regarding these resources in order to plan

their allocation throughout the test revision project from the outset, and to set aside a contingency fund and additional time for unforeseen problems.

Phase Four: Academic Enquiry

4.1 The conceptualisation and operationalisation of components of revised tests should be reviewed and appropriately revised to minimise construct-irrelevant sources of score variance. Tests should strive to measure constructs accurately, without interference from factors that are outside the scope of the construct. The variance in test scores should be linked directly to variance in the assessed construct, and not because of construct-irrelevant sources (Camara, 2007). As such, performance should provide valid evidence of the test construct for test takers from all populations for whom the test was designed (Oliveri, Lawless & Young, 2015). Revision teams should conduct research to determine the extent of construct-irrelevant interference in test scores, as such interference may affect the recommendations that are based on test scores (ETS, 2014). Different sources of interference in accurate measurement can all undermine a test user's ability to make valid inferences about test takers (Oliveri, Lawless, & Young, 2015).

Culture and language are important considerations in this regard. Specific words or phrases may have different meanings for people from different countries or contexts. Cultural differences may also affect how a construct should best be assessed (ITC, 2016). Test items with high cultural loads require test takers to have specific knowledge of the mainstream culture of a specific country. This can unfairly discriminate against test takers from other countries or cultures that differ from the perceived mainstream as such items may measure acculturation as an irrelevant secondary construct, in addition to the intended primary construct of the test (Oliveri, Lawless, & Young, 2015). It is important that a revised test measures the construct accurately for all intended populations (ITC, 2013a).

Test takers from different backgrounds may be less familiar with certain item formats or modes of testing, which may negatively affect their performance. For instance, in developing countries such as South Africa, test takers from underprivileged communities may not be familiar with computers or tablet-based technologies such as keyboards and touch screens. Using these modes of testing on test takers from these communities would negatively affect their test performance (Foxcroft, 2004). Revision teams should periodically review any sources of construct-irrelevant interference, to identify invalid components that may unfairly prevent test takers of certain groups from demonstrating their abilities in the intended test construct (ETS, 2009). Such components should be revised as far as possible, or removed during a test revision (ETS, 2014).

Revision teams should take cognisance therefore of the effects of language and culture on how constructs are assessed and take proactive steps to counteract these sources of measurement interference in revised tests.

4.2 Revision teams should balance the needs of test users and the domain measured, when deciding on test items and the nature of tasks required from test takers of a revised test. As part of the academic enquiry of a test revision, revision teams should familiarise themselves with the needs of practitioners who use the test. It is important to understand the contexts in which a test is used, as well as for what purpose. The nature of tasks included in a revised test should be informed by the contexts in which the test is utilised, as well as by the test

takers. As there may have been changes in the contexts and the tasks test takers may be expected to perform since the launch of the previous version of a test, these changes should be considered in the types of items included in the revised test. These needs would be particularly important if a test is expected to interface with other forms of assessment in formal settings, including hospitals or educational systems, for diagnostic purposes or to track the effectiveness of remediation or intervention strategies (Liu & Dorans, 2013).

4.3 Utilising careful analysis, optimally functioning components of a test should be considered for inclusion in a revised test to act as anchor items, or to foster a sense of brand familiarity between different test editions. Major test revisions face heightened scrutiny from existing test users. The product of a revision that reflects a shift in underpinning constructs, test questions, target populations, as well as scoring or norming methods, can create a sense of disconnect between the revised and previous test versions. Steps to address this potential lack of connection are to include items from previous versions in a revised test. Such legacy items would need to be selected after expert and statistical vetting as being useful for the revised edition. These items would create an anchor block, which can assist in establishing the link in test difficulty between different versions (Geisinger, 2013). Another strategy is for a revised test to utilise the same system of item formats or scoring as previous versions to minimise errors in administration and scoring by users of the previous version (Adams, 2000). All decisions in this regard would be guided however by expert opinion, user feedback, and statistical analysis.

Phase Five: Item Development

5.1 The development of test items should consider multicultural contexts, and the possibility that revised tests may be used eventually in settings for which they were not initially intended. A popular test may be used eventually for applications for which it was not originally designed. This is particularly prevalent in the global environment, where tests have only been developed for a specific country but are eventually used in other countries. Revision teams need to be aware of this possibility and develop items that either are applicable for a global audience or easily adapted for other cultures (Foxcroft, 2004). The benefit for revision teams is that they may already have some insight of the global exposure of the previous test, which will inform the test items that are included in the revised test.

Test publishers should periodically review the changes in contexts in which their tests are used, as well as in the test populations, as this may suggest possible aspects to be addressed in test revision (ITC, 2013a). Another trend in psychological testing is the conversion of standard tests to computer-based or online tests. These modes of testing require special considerations and adaptations. The equivalence of traditional and technological versions of a revised test would be improved if revision teams were mindful of such future developments, and if they created test items from the outset that could be extended to other modes of testing (Strauss, Hunter & Spreen, 2000).

5.2 When authoring item content and test instructions, revision teams should anticipate translation of a revised test into other languages in the future. The issues of culture and language remain amongst the most challenging aspects in the accuracy of test results as they directly affect test content (ITC, 2013a). A popular practice in the test industry is to translate tests into other languages to extend the test user market and for cross-cultural research. Multiple-language tests are not only desirable, but also often necessary to reduce bias and

promote accurate and fair testing in international settings (Geisinger, 2013). Translation from the original source language to a new target language without accounting for cultural differences can be a significant source of construct-irrelevant interference.

In some cases, it may be impossible to create a direct translation of items, due to non-existent words in the target language. This would necessitate adaptation of items for the target language (ETS, 2009). Revision teams should generate a comprehensive vision for a revised test, utilising feedback from the review of the countries and populations using the previous test version. By knowing the initial countries where a revised test will be used, test instructions and items for a revised test must be authored in such a way as to simplify future translation and adaptation, as these practices can affect the success of multilingual international tests (Oliveri, Lawless, & Young, 2015).

If materials are developed in multiple languages as part of the test revision, these should minimise language bias as a nuisance variable. This would require knowledge of the source and target languages, and experts should be selected who are familiar with the different languages and regional nuances within a single language, as well as knowledgeable of all intended test populations. Attention must be paid to attaining equivalent difficulty in texts for different languages and populations. Revision teams should provide evidence of the similarity in meaning for all intended populations for a revised test (ITC, 2016; Oliveri, Lawless, & Young, 2015).

Phase Six: Test Piloting

5.1 The development of test items should consider multicultural contexts, and the possibility that revised tests may be used eventually in settings for which they were not initially intended. A popular test may be used eventually for applications for which it was not originally designed. This is particularly prevalent in the global environment, where tests have been developed for a specific country but are eventually used in other countries. Revision teams need to be aware of this possibility and develop items that either are applicable for a global audience or easily adapted for other cultures (Foxcroft, 2004). The benefit for revision teams is that they may already have some insight of the global exposure of the previous test, which will inform the test items that are included in the revised test.

Test publishers should periodically review the changes in contexts in which their tests are used, as well as in the test populations, as this may suggest possible aspects to be addressed in test revision (ITC, 2013a). Another trend in psychological testing is the conversion of standard tests to computer-based or online tests. These modes of testing require special consideration and adaptation. The equivalence of traditional and technological versions of a revised test would be improved if revision teams were mindful of such future developments, and if they created test items from the outset that could be extended to other modes of testing (Strauss, Hunter & Spreen, 2000).

5.2 When authoring item content and test instructions, revision teams should anticipate translation of a revised test into other languages in the future. The issues of culture and language remain amongst the most challenging aspects in the accuracy of test results as they directly affect test content (ITC, 2013a). A popular practice in the test industry is to translate tests into other languages to extend the test user market and for cross-cultural research. Multiple-language tests are not only desirable, but also often necessary to reduce bias and promote accurate and fair testing in international settings (Geisinger, 2013). Translation from the

original source language to a new target language without accounting for cultural differences can be an important source of construct-irrelevant interference.

In some cases, it may be impossible to create a direct translation of items, due to non-existent words in the target language. This would necessitate adaptation of items for the target language (ETS, 2009). Revision teams should generate a comprehensive vision for a revised test, utilising feedback from the review of the countries and populations using the previous test version. By knowing the initial countries where a revised test will be used, test instructions and items for a revised test must be authored in such a way as to simplify future translation and adaptation, as these practices can affect the success of multilingual international tests (Oliveri, Lawless, & Young, 2015).

If materials are developed in multiple languages as part of the test revision, these should minimise language bias as a nuisance variable. This would require knowledge of the source and target languages, and experts should be selected who are familiar with the different languages and regional nuances within a single language, as well as knowledgeable of all intended test populations. Attention must be paid to attaining equivalent difficulty in texts for different languages and populations. Revision teams should provide evidence of the similarity in meaning for all intended populations for a revised test (ITC, 2016; Oliveri, Lawless, & Young, 2015).

Phase Seven: Test Standardisation

7.1 Revision teams should give due consideration to the representativeness and size of standardisation samples, in order to develop normative information for a revised test that is applicable to intended test takers. Test norms are an important consideration during test development and test revision processes. Norms are the interface between client-focussed test performance and the external macro-system that surrounds the test taker and within which they function. Norms assist to transform potentially meaningless test scores to an objective peer-informed interpretation of test behaviour. It is therefore important for a revision team to design a strategy to develop norms to maximise generalisability and usability, whilst keeping costs within acceptable parameters (Butcher, 2000).

The important questions about norm samples relate to who to include in the sample, and how many test takers are required. The norm sample should consist of participants that are relevant for the intended test populations. Tests that are used for diverse populations require a more complex sampling strategy. In the event that the norm sample cannot consist of sufficient representation from all groups, research should be conducted to demonstrate the equivalence in performance of different groups on a revised test. Test norms may not be used for populations where adequate norms or research evidence of equivalence is lacking (ITC, 2013a, 2016). All information about size, composition, and source of norm groups, including their representativeness, should be provided in test manuals (EFPA, 2013a).

The size of samples used to develop norms for a revised test is an important consideration. Revision teams should familiarise themselves with the guidelines of test classification agencies when planning the test standardisation phase. The EFPA, for instance, is prescriptive in how sample size affects the score and subsequent classification of a test. The organisation distinguishes between traditional norms, which view a norm sample group as individual strata, and the increasingly popular continuous norming approaches that divide sample sets into overlapping subgroups that would allow for a more seamless norming matrix for all groups

(EFPA, 2013a). Table 1 reflects the minimum EFPA (2013a) sample requirements for a low-stakes test (which is not a primary source of evidence when making life-changing decisions) and the associated qualitative classification of different samples.

Table 1. Minimum EFPA sample size requirements for low-stakes test classification

Classification	Traditional Norming	Continuous Norming
Inadequate (1)	Below 200	Less than 8 subgroups (maximum group size of 69)
Adequate (2)	200-299	8 subgroups with 70-99 participants each
Good (3)	300-999	8 subgroups with 100-149 participants each
Excellent (4)	1000 and above	8 subgroups with at least 150 participants each

From Table 1 it appears that the EFPA would only consider a traditional norming sample as adequate if there are at least 200 participants in each normed group. Due to the seamless norming approach used in continuous norming using overlapping samples, the minimum sample size of 70 is applied for each normed group (EFPA, 2013a). Table 1 demonstrates the level of scrutiny imposed by test classification bodies. Revision teams should familiarise themselves with relevant criteria from the bodies they would approach for test classification, as a lack of compliance can be expensive and time-consuming to correct retrospectively.

7.2 Revised tests should be accompanied at launch with adequate norms and standardisation information. Revised tests should be published with the relevant documentation and information that would allow test users to determine the suitability of a test for their clients. The standard information required includes evidence to support the norms, and the validity and reliability of the revised test for the intended populations (ITC, 2016).

The main forms of validity are face, content, construct, and criterion-related validity. Face validity is the superficial appearance and presentation of a test for test takers, whilst content validity refers to coverage of the construct by test items, as judged by experts. As such, face and content validity rely on qualitative judgments that are embedded in the test development and revision process. Construct validity assesses if a test measures what it is intended for, or whether there are unintended underlying constructs embedded in the test that impact on accurate measurement. Factor analysis is a popular method for investigating construct validity. It would also be important to investigate differential item functioning for test takers from different language or cultural groups to determine the cohesive functioning of the test and its ability to measure accurately across different test populations (EFPA, 2013a).

Construct validity can also be investigated by comparing performance on different tests that measure a similar construct. Such investigations into concurrent or convergent validity aim to underscore the validity of a revised test by comparing it to an established test with proven validity. A correlation coefficient of 0.6 between test performances would provide adequate evidence of concurrent or convergent validity (EFPA, 2013a). Criterion-related validity can be investigated using postdictive (ability to predict former behaviour), concurrent (ability to predict current behaviour), and predictive (ability to predict future behaviour) studies. In this context, concurrent studies could refer to the ability of a revised test to predict test performance on other similar tests or on present real-world behaviour. Findings from a range of studies with samples exceeding 200 would be considered excellent (EFPA, 2013a).

Reliability is the evidence of consistency of measurement of a revised test over time, test versions, internal consistency and test administrator. A test-retest correlation coefficient of below

0.6 would be inadequate for a reliable measure, whilst a coefficient exceeding 0.8 would be an excellent result (EFPA, 2013a). A range of coefficients can be used to measure the internal consistency of a test or subtests, including Cronbach's alpha, Kuder-Richardson 20 or 21, Lambda-2.factor analysis (omega or theta), and greatest lower bound estimate. According to EFPA (2013a), coefficients of below 0.7 would be inadequate, and higher than 0.9 would be excellent. Correlation studies between different forms of a revised test or multiple scorer ratings can also offer valuable insights into reliability.

Item-response theory (IRT) may be used to offer insight into item discrimination, item difficulty, and guessing of answers. The use of IRT can extend to developing a theoretical model of the test and estimates of a revised test's ability to measure underlying trait factors. Such studies are largely dependent on adequate sample size, with a suggested minimum guide of 200 participants for a one-factor (discrimination), 500 for a two-factor (discrimination and difficulty), and 700 for a three-factor (discrimination, difficulty, and guessing) studies (EFPA, 2013a). Revision teams should investigate different forms of reliability in a range of studies, using adequate and applicable samples, to form a clear picture of test reliability. An indication should be supplied of error of measurement in a revised test as well as measures implemented by the revision team in the norms and interpretations to overcome such artificial lowering or raising of scores (ITC, 2013a).

Some tests are used to assist in the diagnosis of certain disorders or illnesses, and to monitor the effectiveness of treatment for clients. With the fragmentation of traditional diagnoses into ever-widening and deepening layers, producing norms or research relevant to each category has become unfeasible. Revision teams should therefore provide at least some information in the manuals of revised tests about the scores of test takers from certain clinical groups, compared with matched samples from non-clinical samples. Such information could qualitatively guide test users about potential uses of a revised test for clinical populations until additional research is published (Geisinger, 2013).

Phase Eight: Conduct Supporting Research

8.1 Revision teams should prioritise research into all target populations of a revised test, including clinical and non-clinical samples. It may take years after publication for research to be conducted with a revised test on clinical populations. This being said, revision teams should identify key populations and conduct research to guide the use of the test for such populations, for inclusion in the test manuals and training materials, together with a communication that such research serves as a starting point for ongoing research on different populations.

Research should draw on samples from various clinical and non-clinical populations, and effort should be made to produce research that will maximise the usability and generalisability of findings (Oliveri, Lawless, & Young, 2015). Revision teams should prioritise such research and inform test users of research results on revised tests as soon as possible. This may include releasing pertinent information early through presentations and bulletins to test users, prior to dissemination of the full results in professional publications (EFPA, 2013a). Research should be based on sufficient data, as it would not be in the best interests of test users to rely on the perceived similarity of certain populations or levels of test performance (Strauss, Spreen, & Hunter, 2000). Users of revised tests should request research information on clinical populations

from test publishers, and consider contributing to such projects if it is within their field of interest or expertise (Bush, 2010).

8.2 Multiple methods should be employed to investigate the relationship between previous and revised editions of a test. It is important for test users to understand how a revised test compares to its predecessors. Failure to do so would lead to misleading results, and result in unintended and inappropriate use of a revised test (Strauss, Spreen, & Hunter, 2000). This information includes a comparison of the validity and reliability of the previous and revised editions, differences in the intended populations, conditions for test use, administration and scoring guidelines, and how norm tables should be used and results interpreted.

As the performance of similar test populations may change over time due to anomalies, such as the Flynn Effect, test users should be informed that test performance on different test editions may not be used interchangeably or be directly equated (Strauss, Spreen, & Hunter, 2000). Revision teams should investigate test performance on different test editions, including score equity assessments aimed at analysing construct continuity and equivalence of scores (Liu & Dorans, 2013). If, during test revision, changes have been made to the underlying constructs of the test, it will restrict comparison of global scores, as the factor structures of the different editions will be dissimilar. One method of providing some insight into performance includes converting scores from different versions to a common comparable scale, such as T-scores, that have been corrected for biographical variables such as age, gender, and culture (Strauss, Spreen, & Hunter, 2000).

8.3 Research should be conducted into the validity and reliability of a revised test. Revision teams have a responsibility to provide comprehensive evidence of the test validity and reliability of a revised test (Butcher, 2000). This information should include technical documentation that highlights different types of validity and reliability (Camara, 2007). The evaluation of tests by classification organisations requires comprehensive presentation of validity at statistically significant levels, as evidenced by a range of studies (EFPA, 2013a). Professional bodies similarly require test users to base test selection on reviews of validity and reliability, as a minimum best practice requirement (ITC, 2015). Revision teams should not rely exclusively on the validity and reliability evidence of previous editions of a test but should fully investigate these areas on the revised test. This evidence should be supplied in the test manuals of the revised test and be expressed clearly with statistical information appropriate to the methods used (AERA, 2014).

It is of concern that some test users employ tests without adequate training or tests that have been adapted, translated, or revised without adequate supporting research or validation (Naglieri et al., 2004). Test users are advised by registration bodies to avoid using tests that have inadequate or unclear technical documentation, an oversight that can be directly linked to failings on the part of revision teams and publishers (ITC, 2013a). Although revision teams may strive to meet these professional standards, they often have to balance competing demands and tight deadlines. This may result in taking shortcuts in the phases just prior to the launch of a revised test, most notably research on test validity and reliability (Mattern, Kobrin, & Camara, 2012). Common sources of an impoverished validation practice are a lack of staff resources or capacity, monetary or business concerns, and a lack of research data on a revised test.

Various aspects may be of interest to test users. This includes technical information regarding the construct representation of the test (i.e., content validity) (Camara, 2007). Within the competitive test publishing market test users would also take note of how test takers perform

on a revised measure as compared to other similar products (i.e., concurrent validity) (Strauss, Spreen, & Hunter, 2000). Test revision teams should note the different forms of validity (including content, construct, concurrent, and criterion validity) and reliability (such as test-retest, split-half, inter-rater, and intra/inter-scale reliability). Research is ever expanding in these fields, but revision teams should focus on tried-and-tested methods that communicate the strengths and weaknesses of a revised test in a clear and unbiased fashion (Mattern, Kobrin, & Camara, 2012).

Phase Nine: Test Product Assembly and Launch

9.1 The extent of a revision should be communicated in the product description of a test. Butcher (2000) identifies 'light', 'medium' and 'extensive' as three types of test revision. A 'light' revision entails changes made mostly to the test manual. Aspects that could fall within this type are minor updates to item wording or editorial changes. A 'medium' revision is more intensive and includes changes to or replacing non-performing items, and updating the norms of a test. An 'extensive' revision involves a complete reanalysis and reconstruction of the test. This could include re-examining the theoretical foundation of the test and major changes to items or subscales, together with a new set of test instructions. An extensive revision would also include new norm data, as well as validity and reliability studies (Butcher, 2000).The term 'revised' should only be attached to tests that have been updated in significant ways, such as in 'medium' and 'extensive' revisions.

If the test has not been changed significantly after a 'light' revision, the test should rather be marketed as containing minor changes or updates. The extent of these changes should be clearly communicated to existing and new test users.

9.2 When tests are revised, users should be informed of the changes to the specifications, underlying constructs, and changes to the scoring method. Revision teams should present any changes to a revised test in comprehensive technical documentation. Documentation should also focus on how the revised test differs from its predecessor (ITC, 2016).

The theoretical foundations for updates to constructs should be supplied (EFPA, 2013a). Any differences in target populations, methods of norm development, and the correspondence between norms from previous and revised test editions and their potential impact should be unpacked (ITC, 2015). Differences and similarities in the techniques used to convert raw scores to standardised scores must be explained to avoid confusion amongst test users (ITC, 2013b). Documents should do more than reflect on different editions of a test and should go further in justifying the need for a revised version (ITC, 2015). Emphasis should be placed on evidence regarding how the revised test builds or improves on its predecessor, as it would be unethical to develop a test that cannot at least be held to the standards of its predecessor (Naglieri et al., 2004).

9.3 Test users should be clearly informed of the comparability and relationship between the previous and revised editions of a test. There are many reasons why the ties between the previous and revised editions of a measure should be clearly established. The first is that a revision team may face change resistance from established test users, who may make unfounded claims that the revised test is too different from its predecessor or, more likely, so similar that the expense to purchase the revised test is unnecessary (Butcher, 2000). The second

reason is that test users conduct an assessment based on the construct in question. It would be important for test users to be aware of the comparability of the constructs being measured by previous and revised test versions to assess the relevance of the revised test for the assessment need (EFPA, 2013b). A third motivation is that, despite following explicit blueprints in test revision, changes may occur over time (ITC, 2013b). A revised test may draw on content from its predecessor, but there may also be new test questions. With regard to the latter point, items in some content areas are more difficult to replace than others, which may result in marked differences between previous and revised test editions.

For challenge tests that include items with a range of difficulty, items with difficulty levels at the extreme low and high ends are more difficult to develop, clone or replicate. This may affect the overall difficulty of the revised test, which will affect how its test scores compare to a previous version (Liu & Dorans, 2013). Any changes to the difficulty level of a test between different versions should be clearly explained, as well as the comparability of test scores from different editions (AERA, 2014).

9.4 Documentation for revised tests should be amended and dated to keep information for test users current. Any substantial changes to a test should be reflected in its updated documentation, and with supplementary information to existing test users. This includes general information as well as cautions regarding test use (AERA, 2014). The focus should be on the adequacy of information for test users, including administration guidelines, technical information, and norm supplements (EFPA, 2013a). The main purpose of this information is to enable evaluation of the revised measure for its use on individual test takers, as well as certain populations (ITC, 2013a).

9.5 Test publishers should consider the economic circumstances of test users when determining the cost of a revised test. The cost of revised tests has continued to escalate. Test users have their practices in a variety of settings, which affects the availability of funds to purchase new tests. Those in private practice may need to budget for a revised test, while those in institutional settings may have to apply for funds from their employers. Some test users may be from developing countries, where the availability of finance is lower (Adams, 2000). These financial considerations will influence the rate at which a revised test is adopted (ITC, 2015). Test publishers should consider their test user market and price revised tests accordingly. Test users should also be informed as early as possible what the price range for a revised test would be, to enable them to plan for this expense, and to engage with test publishers.

Phase Ten: Post-Launch Activities

10.1 Test publishers and users share a joint responsibility to engage with each other regarding revised tests. The quality of the psychological testing services is informed by the relationship between test publishers and test users (ITC, 2015). This requires concerted effort from both stakeholder groups to improve the dialogue concerning psychological tests, including revised test editions (Bush, 2010). One area in which test publishers can improve this relationship is by communicating openly and accurately with test users regarding revised tests. This would include their online presence and the information provided on publisher websites. Test publishers should remember that existing and potential test users consult these online resources, and the relationship can be supported by providing accurate and updated information

regarding tests (Naglieri et al., 2004). Test users should connect with the publishers of the tests they utilise and engage with the information provided by test publishers.

10.2 Test publishers should develop a reasonable strategy to assist test users to switch to a revised test edition. Financial considerations play an important role in the speed with which test users will adopt a revised test (ITC, 2015). Test publishers become frustrated by the perceived unwillingness of test users to invest in revised tests (Adams, 2000). Many test users will also adopt a waiting strategy to evaluate new research after the launch of a revised test. This may not necessarily be due to change resistance, but the need to be convinced that a revised test represents a tangible improvement over the previous version (King, 2006). The result is that it can take time for a revised test to gain acceptance in the professional community. Test publishers should assist test users by developing a reasonable strategy to transition to the revised test edition. This may include financial assistance, such as a reduced pre-launch order price. Test publishers need to decide on and advertise an end date of use for the previous version, and remain steadfast in their resolve, whilst assisting test users to adopt the revised test (Butcher, 2000).

10.3 Test publishers should offer comprehensive training to promote the level of competence with which test users employ revised tests. Test users are required to remain current with changes and advances in tests, and to only offer services for which they are qualified (EFPA, 2013b; ITC, 2013a). Test publishers can assist test users to achieve this practice standard by offering training programmes and practical workshops in the months leading up to and subsequent to the publication of a revised test. This will enable test users to adopt a revised instrument faster (Butcher, 2000). In reality, some common mistakes in the use of a revised test will surface. Test publishers should advise users of these common errors in a timely manner through a variety of ways, such as in writing or through special error prevention meetings. Publishers must document and disseminate information on common errors in the use of a revised test, as well as how to prevent such mistakes (ITC, 2013b).

10.4 Test users should guard against resistance to change, keep current with changes to tests, and strive to adopt a revised test as soon as possible, with due consideration for the best interests of their clients. Test practitioners rely on the psychological tests they employ. Over time, this reliance can become ingrained, which can result in attachment to a specific test edition that is outdated (Butcher, 2000). Attachment to a previous test version is not an acceptable justification for not adopting a revised test (ITC, 2015; King, 2006). Test users must accept responsibility for the tests they use, and the accuracy of the recommendations they make (EFPA, 2013b).

The industry standard is for test users to transition to a revised test within six months to a year post-launch (Bush, 2010). This decision should be informed by the relevance of the test for each test taker and the purposes of the test user. Users should have an unbiased approach to revised tests and review the scientific merits of revised tests before reaching a decision. Despite the cost implications of adopting a revised test, economic considerations should not be the primary basis for decisions about test selection. The merits of the test in facilitating an accurate assessment of a test taker should be the most important criteria (ITC, 2015).

A revised test should be adopted as soon as possible if evidence exists that there has been a shift in norms from the previous test edition (such as a large Flynn effect), or if there have been updates to the conceptualisation or measurement of the test constructs (Strauss, Spreen, & Hunter, 2000). Previous test versions may still be used for research purposes, and for test takers

assessed from groups (such as language, culture, age, or specific disability) for whom there is an absence of appropriate test norms or validity studies on the revised test, but which is available on the previous edition (King, 2006).

Test users should only offer assessment services they are qualified to render (AERA, 2014). This requires that they update their knowledge when switching over to a revised test, by studying the test materials and by undergoing training (EFPA, 2013b). Test users should not assume that the method of test administration, scoring, and interpretation used for a previous version would still apply to a revised edition. Time should be taken to learn how to competently use and interpret a revised test (King, 2006; Strauss, Spreen, & Hunter, 2000). In addition, test users should be committed to lifelong learning by refreshing their knowledge about the tests they employ, through follow-up seminars and experiential training (AERA, 2014). They should also remain current in their knowledge of legislation, policy and psychological testing practice. This includes advice, warnings, and guidelines from their professional bodies and their employers (ITC, 2013a).

10.5 Revision teams should develop a comprehensive post-launch research strategy and encourage the dissemination of independent research studies. At launch, a revised test is accompanied by the research performed during its development. As it is adopted by test users, a revised test is used in many contexts with test takers from different backgrounds. Each test session is unique and provides an opportunity for research and learning. Revision teams should spearhead ongoing research into a revised test, and engage with researchers and test users internationally. They should develop a list for test users and researchers that highlights the evidence required to validate a revised test for use on different populations (Mattern, Kobrin, & Camara, 2012). In addition, revision teams should encourage independent research aimed at replicating the validity and reliability claimed in test materials (ITC, 2013b). Minor deviation from the claimed statistics is acceptable, but significant differences beyond expected patterns of performance should be noted, published, and researched further. Test users should be open to participating in research studies, and lend their expertise to assist in data collection, providing relevant anonymised test data, and sharing interesting test experiences on a revised test (ITC, 2013a).

24 May 2018

Developed by: Mr Johan Cronje
Nelson Mandela University
Johan.cronje@mandela.ac.za

References

Adams, K. M. (2000). Practical and ethical issues pertaining to test revisions. *Psychological Assessment, 12,* 281-286.

American Educational Research Association. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Bush, S. S. (2010). Determining whether or when to adopt new versions of psychological and neuropsychological tests: Ethical and professional considerations. *The Clinical Neuropsychologist, 24,* 7-16.

Butcher, J. N. (2000). Revising psychological tests: Lessons learned from the revisions of the MMPI. *Psychological Assessment, 12,* 263-271.

Camara, W. J. (2007). *Standards for educational and psychological testing: Influence in assessment development and use.* Retrieved from http://www.teststandards.org/files/standards%20-%20influence%202007.pdf

Education Testing Service (ETS). (2014). *ETS standards for quality and fairness.* Retrieved from www.ets.org.

Education Testing Service (ETS). (2009). *ETS international principles for fairness review of assessments: A manual for developing locally appropriate fairness review guidelines in various countries.* Retrieved from www.ets.org.

European Federation of Psychologists' Associations (EFPA). (2013a). *EFPA review model for the description and evaluation of psychological and educational tests: Test review form and notes for reviewers (Version 4.2.6).* Retrieved from http://www.efpa.eu/download/650d0d4ecd407a51139ca44ee704fda4

European Federation of Psychologists' Associations (EFPA). (2013b). *Performance requirements, context definitions and knowledge & skill specifications for the three EFPA levels of qualifications in psychological assessment.* Retrieved from http://www.efpa.eu/download/1b272a998e297c248413fbb761134697

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

Geisinger, K. F. (Ed.). (2013). *APA handbook of testing and assessment in school psychology and education* (Volume 3). Washington, DC: APA.

International Test Commission. (2016). *ITC Guidelines for translating and adapting tests* (2nd ed.). Retrieved from http://www.intestcom.org/

International Test Commission. (2015). *Guidelines for practitioner use of test revisions, obsolete tests, and test disposal.* Retrieved from https://www.intestcom.org/files/guideline_test_disposal.pdf

International Test Commission. (2013a). *ITC guidelines on test use*. Retrieved from http://www.intestcom.org/files/guideline_test_use.pdf

International Test Commission. (2013b). *ITC Guidelines on quality control in scoring, test analysis, and reporting of test scores*. Retrieved from https://www.intestcom.org/files/guideline_quality_control.pdf

King, M. C. (2006). Adopting revised versions of psychological tests. *The CAP Monitor, 23,* 6-7.

Liu, J., & Dorans, N. J. (2013). Assessing a critical aspect of construct continuity when test specifications change or test forms deviate from specifications. *Educational Measurement: Issues and Practice, 32,* 15-22.

Mattern, K. D., Kobrin, J. L., & Camara, W. J. (2012). Promoting rigorous validation practice: An applied perspective. *Measurement: Interdisciplinary Research and Perspectives, 10*(1-2), 88-92.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the Internet: New problems, old issues. *American Psychologist, 59*(3), 150-162.

Oliveri, M. E., Lawless, R., & Young, J. W. (2015). *A validity framework for the use and development of exported assessments.* Princeton, NJ: Educational Testing Service.

Strauss, E., Spreen, O., & Hunter, M. (2000). Implications of test revisions for research. *Psychological Assessment, 12,* 237-244.