The Validation of a Rating Scale for the Assessment of Compositions in ESL

K. Hattingh

Thesis submitted for the degree Doctor of Philosophy at the Potchefstroom Campus of the North-West University

Promoter: Prof. J.L. van der Walt

2009

ABSTRACT

Keywords: assessing second language writing; rating scale development; empirical scale development; writing rating scales; validity; validation; validation argument; validation framework; scoring validity

This study aimed to develop and validate a rating scale for assessing English First Additional Language essays at Grade 12 level for the final National Senior Certificate examination.

The importance of writing as a communicative skill is emphasised with the re-introduction of writing as Paper 3 of the English First Additional Language examination at the end of Grade 12 in South Africa. No empirical evidence, however, is available to support claims of validity for the current rating scale.

The literature on the concept of validity and the process of validation was surveyed. Theoretical models and validation frameworks were evaluated to establish a theoretical base for the development and validation of a rating scale for assessing writing. The adopted framework was used to evaluate the adequacy of the current rating scale used for assessing Grade 12 writing in South Africa. The current scale was evaluated in terms of the degree to which it offers an appropriate means of assessing Grade 12 Level essay writing while adhering to requirements of the National Curriculum Statement. It was found lacking and the need for a new, validated rating scale was established. Various approaches to scale development were considered in consideration of factors that impact scores directly, viz. the type of rating scale, rater characteristics, scoring procedures and rater training.

A new scale was developed and validated following an empirical procedure comprising four phases. The empirical process was based on an analysis of actual performances of Grade 12 English learner writing. A combination of quantitative and qualitative methods was used in each of the four phases to ensure the validity of the instrument. The outcome of this project was an empirically developed and validated multiple trait rating scale to assess Grade 12 essay writing. The proposed scale distinguishes five criteria assessed by means of a seven-point scale.

OPSOMMING

Sleutelwoorde: assessering van tweedetaal skryfwerk; meetingskaalontwikkeling; empiriese skaalontwikkeling; geldigheid; validering; geldigheidsargument; bepuntingsgeldigheid

Die doelwit van hierdie studie was om 'n geldige meetingskaal daar te stel om Graad 12 skryfwerk te assesseer vir die Nasional Senior Sertifikaat eksamen.

Die belangrikheid van skryfvaardighede word benadruk deur die herinstelling van skryf as Vraestel 3 van die Engelse Eerste Addisionele Taal eksamen aan die einde van Graad 12 in Suid Afrika. Geen empiriese bewyse is egter beskikbaar om 'n geldigheidsargument vir die huidige skaal te staaf nie.

'n Literatuurstudie van die konsep geldigheid en die proses van geldigmaking (validering) is onderneem. Teoretiese modelle en valideringsraamwerke is geëvalueer om 'n teoretiese grondslag vir die ontwikkeling en validering van die skaal daar te stel. Die raamwerk wat gevolg is is gebruik om die huidige skaal wat gebruik word om Graad 12 skryfwerk te assesseer the evalueer. Die huidige skaal is beoordeel in terme van die mate waartoe dit 'n geskikte instrument bied vir die assessering van Graad 12-vlak opstelle en voldoen aan die voorskrifte van die Nasionale Kurrikulumverklaring. Tekortkominge is geidentifiseer en die noodsaaklikheid vir 'n nuwe, geldige metingskaal is aangetoon.

Verkseie benaderings tot skaalontwikkeling is oorweeg met in agneming van faktore wat uitslae affekteer, soos die tipe skaal wat gebruik word, eienskappe van nasieners, nasien prosedures en opleiding van merkers.

'n Nuwe skaal is ontwikkel en gevalideer deur 'n empiriese proses wat vier fases beslaan het. Die empiries proses is gebasseer op 'n analise van werklike voorbeelde van die skryfwerk van Graad 12 leerders. 'n Kombinasie van kwantitatiwe en kwalitatiewe metodes is gebruik in elk van die vier fases om die geldigheid van die instrument te verseker.

Die uitkoms van hierdie projek is 'n empiries ontwikkelde en gevalideerde multi-eienskap meetingskaal om Graad 12 opstelle te beoordeel. Die voorgestelde skaal onderskei vyf kriteria wat aan die hand van 'n sewe-punt skaal beoordeel word.

ACKNOWLEDGEMENTS

This study is dedicated to my parents.

I would sincerely like to thank everyone who contribution to the successful completion of this study. Thank you to the Almighty for granting me the opportunity and ability to pursue and realise my ambition, and for each of the following parties who contributed to the success of this study.

Those who provided professional guidance and support:

- Prof. J.L. van der Walt for his clear, prompt, realistic, continuous and patient guidance as my promoter. Thank you for your enthusiasm and understanding throughout the duration of this project.
- My colleagues at the School of Languages (NWU); in particular the English staff for their advice, encouragement and accommodating me so generously, enabling me to complete this study.

For financial support:

- National Research Foundation

My personal support system:

- My parents for always believing in me and helping me to believe in myself. Your love and support have been invaluable throughout the period of this study.
- My brother for always being proud of me, being available when I needed a break and also understanding when I was unavailable.
- My extended family for their interest and motivation.
- All my friends – old and new from across the globe –for their amazing support and helping me to maintain a balance in life.
  In particular, thank you Sorita, Susan and Marni, who were closely involved and endured the ups and downs with me. Thanks also to Hanta, Toinette, Thea and Marali for their sincere interest, excitement and continuous support.

Participants

All those who were involved in the study, for sharing their expertise, time and resources, and for their enthusiastic participation. In particular:

- Prof. H.S. Steyn for his assistance with the analyses of the data;
- Those involved in the empirical process: Prof. Brenda Spencer, Dr. Mirriam Lephalala, Prof. Adelia Carstens, Prof. Carisma Nel and Dr. Ian Butler; and all the expert raters who participated willingly.
- Those who provided additional time, input and resources: Dr. A.J. Weideman; Dr. P.J. Mafisa; Dr. Stuart Shaw and Dr. Barry O'Sullivan; and Elsa van Tonder and Teresa Smit at the Language and Literature in the South African Context Research Unit, NWU.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Problem statement and motivation

Language assessment is a complex and multi-faceted process. Hudson (2005) describes language as possibly the most complex of human abilities and states that assessing language ability can be expected to be as complex. What we are in fact assessing is "an individual's performance interacting within a very social context" (Hudson, 2005:208).

The communicative approach to teaching language renders writing an increasingly valuable skill. Weigle (2002:1) notes: "[T]he ability to write effectively is becoming increasingly important in our global community, and instruction in writing is assuming an increasing role in both second- and foreign-language education". Writing is thus regarded as one of the most important skills imparted by educational systems, and is often a major part of the assessment process. It is a form of communication that supplies teachers with a record of learners' attempts to use a language communicatively. Great value is placed on learners' ability to organise and express ideas. Research, for example by Oller (1979), shows that acquisition of the writing skill carries over into other skills of language use.

Rating scales are a popular means of assessing writing (cf. North, 2000; Lynch, 2003). Lynch (2003:57) notes that scales provide "a consistent reporting format for results from various levels of testing and assessment". As such, rating scales potentially provide a common framework for different stakeholders (viz. learners, teachers, parents and administrators) (North, 2000:11-12; Lynch, 2003:57).

Scales have traditionally been constructed by a committee of experts, using their intuition and expert judgement alone. This approach has increasingly been criticised (cf. North & Schneider, 1998) and an empirical approach to the validation of scales has been advocated more recently (cf. Upshur & Turner, 1995; Fulcher, 1996; Taylor, 2000; Weigle, 2002; Weir, 2005).

Empirical scale development entails developing scales based on analyses of actual samples of learner writing. Such analyses may reveal typical traits of how the construct is manifested in practice. These traits can then be described in the rating scale. An empirically-based approach also involves investigating how criteria and descriptors are likely to be interpreted and applied by raters. This is especially necessary if a centralised and standardised scale is to be used. According to Weir (2005:15), validation entails evaluating an instrument based on a variety of quantitative and qualitative forms of evidence, indicating whether inferences from test scores are verifiable. A combination of quantitative and qualitative methods should therefore be used to collect evidence justifying claims of validity.

This study is concerned with the valid assessment of written compositions produced by Grade 12 level English First Additional Language (FAL) learners. The problem addressed is the development of an empirically validated rating scale. An argument is presented for the validity of the proposed scale, which can be used for its intended purpose and context.

The assessment of writing in English FAL (Paper 3) was reintroduced as a national examination in 2008, as stipulated by the Subject Assessment Guidelines (SAG) in the National Curriculum Statement (NCS) (2005). Learners who write the Grade 12 National Senior Certificate examination (NSC) are expected to produce cohesive and coherent writing, using appropriate content, style and register within a specific context, while fulfilling a function such as arguing or describing.

The Writing paper is notorious for being the most difficult one in which to achieve valid and reliable assessment (Schoonen, 2005). The reliability of scores is influenced by various factors including the rating scale used to assess performances for a particular purpose in a particular context (cf. Bachman, 1990; Bachman & Palmer, 1996; Shaw & Weir, 2007). The current rating scale used for assessing essays written for Paper 3 of the English First Additional examination was originally designed by a committee of examiners and moderators and was based on their experience and expectations of Grade 12 learners. It was not derived from actual examples of learner writing, and it has not been empirically validated. In addition, the scale comprises only two main criteria – language and content – which are assessed on a two-dimensional grid. Issues such as these raise the question of whether the current rating scale of two criteria is sufficient for providing accurate information on learners' writing abilities.

The problem addressed in this study thus boils down to the development and validation of the assessment instrument used to score the essays written in the Senior Certificate Examination in South Africa. It is clear that there is a need for an empirically validated rating scale for assessing essay performances. Such a new scale is likely to increase the reliability of scores and provide a common standard and interpretation of writing ability in Grade 12.

1.2 Research aim and objectives

1.2.1 Aim of the study

The aim of this study is to develop an empirically validated rating scale for assessing the English FAL essays in the final matriculation examination in South Africa.

1.2.2 Objectives

The above aim can be operationalised in terms of a number of objectives. They are to:

- investigate the concepts of *validity* and *validation*;
- identify and describe an appropriate framework for the validation of writing assessment;
- evaluate the current rating scale;
- examine examples of Grade 12 learner writing performances to guide the development of a new scale;
- draw up and validate a new scale by means of quantitative and qualitative procedures;
- propose an empirically validated rating scale for assessing writing in the Grade 12 FAL examination.

1.3 Central theoretical statement

An empirically validated rating scale will produce accurate, fair and reliable results for assessing essays in the Grade 12 examination and will enable the operationalisation of a common standard of writing in Grade 12.

1.4 Method of research

1.4.1 Survey of the literature

Relevant literature on the concepts of *validity* and the *validation process* was reviewed. Frameworks for validating rating scales – and in particular for writing assessment scales – were examined in order to establish a suitable framework for the purpose of this study. Literature on the aspects addressed in the adopted framework (Shaw & Weir, 2007) was reviewed, as well as relevant official documentation published by the Department of Education. The current rating scale was evaluated in terms of its adherence to the requirements in these documents. Literature on quantitative measures such as Rasch analysis and generalisability estimates as well as qualitative procedures such as verbal protocol reports were also considered.

1.4.2 Empirical research

1.4.2.1 Research design

A combination of quantitative and qualitative analyses and procedures was used to develop and validate the proposed rating scale.

1.4.2.2 Data collection and analysis

A new rating scale was developed by following an empirical process that consisted of four phases. In the first phase, sixty-four essays written by Grade 12 learners were benchmarked to illustrate typical examples of writing at seven performance levels. These compositions were analysed in the second phase to identify salient features of writing at each level and incorporate them in a draft scale. The draft scale was then revised and refined in the third phase and piloted in the fourth phase. Relevant quantitative and qualitative methods were use in each phase to achieve the outcomes. Quantitative procedures included Rasch analyses, correlation coefficients and generalisability procedures. Qualitative methods used included expert judgements, written feedback reports and questionnaires.

1.5 Programme of study

Chapter 2 traces the evolution of the concept *validity*. It emphasises the change from the fragmented traditional interpretation to modern views of validity as a unified concept. As validity can only be accessed through validation, Chapter 3 discusses the concept of validity as an argument, with validation as the process of constructing the argument. Scale development must be guided by a validation framework grounded in a model of language competence. Chapter 4 considers models of communicative competence and frameworks of language assessment validation. Shaw and Weir's (2007) framework for validating writing assessment is adopted for the purpose of the present study. Chapter 5 discusses the concept of *scoring validity* as a component of Shaw and Weir's framework. Scoring validity is a central concept in the present study, as it relates to as all aspects that influence scores directly, including the rating scale. Chapter 6 provides an overview of the method of research followed. Chapter 7 discusses the empirical process followed to develop and validate the new rating scale. Chapter 8 concludes the study and makes recommendations for further research.

CHAPTER 2

Perspectives on Validity

2.1 Introduction

Language tests are used to collect information about learners' language abilities. This information is used to make decisions about learners' progress and ability to perform in academic, professional and social situations. Measures must provide accurate information about learners' abilities so that fair inferences can be made about an individual's abilities based on the outcome of a test. Kane (2005:136) highlights the importance of evaluating the extent to which measurement procedures reach this goal. Test developers should ask: Do the scores provide information relevant to the context? Do the scores help to make good and fair decisions? Does the test measure the construct it claims to measure? Do scores mean what they are understood to mean? These questions concern the validity of a test.

This chapter considers the traditional and modern interpretations of the complex, multi-faceted concept of validity. First, it explores the evolution of the meaning of validity. In the discussion of the traditional, segmented view of validity, the constituent parts – or types – of validity are discussed. Then, the modern interpretation of validity as proposed by Messick (1989) is considered. In conclusion, the concept is defined for the purposes of the present study.

2.2 Defining validity

Test-users trust measurement instruments to provide accurate information about test takers' abilities on which to base decisions regarding individual test takers. For Weir (2005:1), the main concern in language testing is the degree to which an instrument can be shown to produce scores that accurately reflect learners' abilities in a specific area, such as reading for main ideas in texts, writing argumentative essays, breadth of vocabulary knowledge, spoken interaction with peers, and so on. Evaluating an assessment procedure thus entails determining whether it generates scores that provide the required type of information to aid such decisions. This is a question of validity (Kane, 2004:136).

The term "validity" is often used to refer to the quality or acceptability of a test (e.g. Henning, 1987:89), but the scope and the meaning of validity have changed significantly over the years (Chapelle, 1999:254). Although the general understanding of validity seems clear and the term seems to be used with a stable meaning in most language assessment papers, the precise meaning of the concept has proven difficult to pin down. Kane (2004:136) perceives the generally accepted definition of validity as too broad. As long ago as 1961, Ebel (1961:640) expressed frustration in this regard, while describing validity as one of the major divinities of psychometrics. Different interpretations and uses of the term are difficult to analyse – unlike, for example, mathematical models – making it problematic to formulate an exact definition of validity.

Weir's (2005:1) concern about accurate scores reflects the importance of identifying the construct to be measured and the way in which the construct should be measured. A clear definition of validity in a language assessment context is fundamental, since it affects all test users and stakeholders and is necessary. It is also necessary for ensuring accurate measurement and fair decisions about issues that affect test users, such as placement, progress made, admission into international or special courses, and admission to educational facilities such as universities. The value of a particular test depends on accepted assumptions about validity in the particular context. These assumptions must therefore be clearly defined (Chapelle, 1999:254).

Both the trait and the method must be appropriate for the purpose of the test. *Trait* refers to the 'what' of a language test, i.e. the underlying construct. *Method* refers to the 'how' of a language test: how the construct is being measured, and what instrument (such as a rating scale) is used to gather information about the trait (Weir, 2005:1; cf. also 3.5.1). If either of these is inappropriate for the purpose of the test, the results are likely to be misleading.

A common problem that affects the validity of testing instruments is that they are often misused to measure abilities which they were not intended to measure. Misusing a testing instrument in this manner renders invalid scores, which makes fair inferences and decisions impossible. The construct, purpose for which and contexts in which a test is valid must therefore be stated explicitly (Alderson, Clapham & Wall, 1995:170).

2.2.1 The traditional conception of validity

Validity is traditionally understood to be concerned with the question of whether a measurement instrument, such as a test or scale, measures what it claims to measure. A test is valid to the degree to which it measures what it is supposed to (Lado, 1961:132; Cronbach, 1971:463; Henning, 1987:89). Traditional definitions of validity assume that assessment is meant to measure something real and that questioning the validity of the assessment means questioning whether it really does measure that specific 'something' (Fulcher & Davidson, 2007:4). Thus, all tests can be valid for some purposes, but not for others.

When the American Psychological Association (APA) first codified validity standards in 1954, four types of validity were identified, corresponding to different test aims. Validity was seen as an indication of the degree to which a test can be used for a certain type of judgment (APA, 1954:13; Shepard, 1993:408). It was seen as mainly comprising three individual entities, namely criterion, content and construct validity, also labelled *the holy trinity* (Guion, 1980).

Traditionally, validity was regarded as a static characteristic of the test instrument. Validity was defined as the extent to which an assessment instrument produced useful information relevant for a specific purpose (Goodwin & Leech, 2003:182). It also incorporated the *trinity* view of validity presented by Cronbach and Meehl (1955).

The different entities of the tripartite concept were used like separate tools in a toolbox, each with a specific function in validating test score interpretations (Kane, 2004:138). Criterion validity was used to validate placement tests. As an aspect of criterion validity, predictive validity was used when making predictions about learners' future performances based on test scores. Concurrent validity, as a second aspect of criterion validity, involved comparing a new test with an external criterion to determine if it could serve as a substitute for an existing, but less convenient test. Content validity was used to validate achievement tests and to describe performances on a universe of tasks. Finally, construct validity was used to examine unseen abilities such as intelligence or anxiety, and was calculated when validating theory-based, explanatory interpretations (Shepard, 1993:408-409; Kane, 2004:138). The three entities are discussed below.

2.2.1.1 Criterion validity

Criterion validity concerns the extent to which the instrument correlates with an external independent criterion, viz. another test or scale designed for the same purpose and context that has been established as valid. Learners' performances on the test in question are compared to their performances on a criterion that is believed to measure the same construct as the test in question accurately. Similar scores for learners' performances on the test and the criterion would indicate a valid test instrument (Hughes, 1989; Bachman, 1990; Weir, 2005; Fulcher & Davidson, 2007).

Criterion validity may not be concerned so much with whether an instrument measures the construct that it is meant to measure or not (Hughes, 1989:22), but correlations with an external variable will only be useful if the external criterion aims to assess the same construct for the same purpose as the instrument in question. Shepard (1993:410-411) points out that empirical evidence is necessary to show that there is a relation between the instrument and the criterion in order for the correlation to be meaningful and indicate criterion validity. The concern is that even if empirical evidence indicates a relation, test and criterion may share the same bias, resulting in a false correlation.

Criterion validity is mainly an *a posteriori* and quantitative concept (Weir, 2005:35), consisting of concurrent validity and predictive validity (Hughes, 1989; Alderson et al., 1996; Weir, 2005; Davidson & Fulcher, 2007).

Concurrent validity refers to cases in which the assessment and the chosen criterion are completed at the same time (Alderson et al., 1996:178; Fulcher & Davidson, 2007:5).

Predictive validity concerns how well an assessment such as a proficiency test can predict the success of learners' behaviour (i.e. a future criterion such as academic success) in future situations, based on their current behaviour (Alderson et al., 1996:181). Placement tests that are used to decide whether learners can enrol in a particular course or function effectively in a foreign learning or business context would typically be investigated for predictive validity.

Fulcher and Davidson (2007:4) note the importance of defining successful behaviour in future real-life settings in order to assess current performances accordingly. For example,

report writing requires the ability to draw conclusions and make clear recommendations based on certain facts regarding a situation. In order to determine the extent of learners' abilities to perform such a task, they have to be assessed according to criteria for successful report writing in the actual working environment. Assessing their present abilities to draw conclusions and make recommendations based on given facts only provides an indication of how well they would be able to cope with this task in real life situations. If they are assessed according to the criteria for success in the future setting, this definition of successful behaviour in a real-life situation is central to establishing validity, since validity is context-bound (Fulcher & Davidson, 2007:5).

Criterion validity was regarded as the most important type of validity or the *golden standard* (Kane, 2004:137) for most of the 20th century. Until the 1950's, correlations alone were the standard measure used to judge the accuracy of a test. Early interpretations of validity assumed that tests are valid if the scale according to which the construct is measured could be correlated positively with a dependent variable (Guilford, 1946:429; Cureton, 1951:623; Shepard, 1993:409). Guilford (1946:429) states that "a test is valid for anything with which it correlates".

Kane (2004:137) criticises views such as Guilford's (1946) that argue for validity of a testing instrument based only on a positive correlation with a dependant external variable. Such a wide interpretation allows for a measurement instrument to be a valid measure of anything with which it correlates.

A major problem with the criterion model is that it is only successful if a valid criterion is readily available. In such cases, the model works simply, elegantly and effectively (Kane, 2004:137). It is often difficult to find or develop an acceptable and valid criterion measure (Alderson et al., 1996:178; Kane, 2004:137). Alderson et al. (1996:178) point out that the criterion measure has to be expressed numerically and must not be directly related to the test itself. It must be external and proven to be a reliable and valid measure of the construct for the exercise to be meaningful. Such criteria may not be available, which then makes establishing concurrent validity impossible (Moller, 1982; Bachman, 1990; Weir, 2005).

A more fundamental problem facing the criterion model is therefore that of validating the criterion (Ebel, 1961:640; Kane 2004:137-138). Even if a suitable criterion is available, the

criterion must also be validated against its own external criterion. If an assessment under question correlates strongly with another measure of the same construct, the external criterion must also be proven valid; i.e. it must correlate with another valid external criterion measuring the same construct. However, another testing instrument that measures the same ability with the same purpose in mind is difficult to find. The problem is pushed back to a point where at least one external criterion must be validated in a way different from correlation. The problem of establishing a valid external criterion thus remains.

Sampling models is one possible alternative method of independent validation. The relationship between the measurement procedures used to generate the scores and the proposed interpretation of the scores is investigated statistically. However, Kane (2004:138) notes that sampling is questionable on various grounds, especially regarding the size and representativeness of the sample and the motivation of the examiners.

2.2.1.2 Content validity

Content validity concerns the extent to which the items in the instrument measure the full construct domain and whether the tasks learners are asked to complete are relevant for the purpose and context of the assessment (Fulcher, 1999; McNamara, 2000; Brown & Hudson, 2002). McNamara (2000:50) explains: "The issue here is the extent to which the test content forms a satisfactory basis for the inferences to be made from the test performance". The substantive aspect of relevance here, according to Fulcher (1999:226), is the extent to which the items included in an instrument represent the trait to be measured in the particular context and for the relevant purpose.

Learners must implement a variety of skills and linguistic structures to perform one communicative construct. The assessment tasks must elicit those skills and structures that provide the most accurate and complete picture of learners' abilities to perform the construct being measured in that context. Assessment tasks should elicit a representative sample of all aspects of the construct. The content of the test must be selected rationally to ensure that the content represents the domain of the construct being tested. Furthermore, the instrument must be constructed according to specifications that consider aspects related to the ability (construct) being tested. These include aspects such as the performance context, characteristics of the text, format of items, the assessment rubric, as well as linguistic and

cognitive abilities related to the ability being tested (McNamara, 1996:96; Fulcher, 1999:492; Brualdi, 1999:3).

If the content of an assessment instrument over- or under-represents a certain aspect of the construct domain, the assessment may lead to invalid scores, unfair inferences and negative washback effects. Therefore, the content should reflect the detailed test specifications according to which an assessment is constructed (Alderson et al., 1996:176; Brualdi, 1999:5). The purpose of the assessment determines which skills and structures related to the ability are most relevant. Test developers must, for example, specify those linguistic features that will provide the most comprehensive picture of learners' abilities to perform the construct in the early stages of test construction. The choice of test content must furthermore be based on a theory of language ability measurement (Hughes, 1978; Wall, Clapham & Alderson, 1991; Fulcher, 1996; Brown & Hudson, 2002; McNamara, 2000).

Fulcher (1999:227) argues that the level of task item difficulty, the quality of rubrics and the accuracy of the scoring key should also be considered under content validity. Items that are too difficult or easy, poor rubrics and inaccurate scoring keys cause construct irrelevant variance.

Face validity is a traditional concept closely related to content validity, but the two must not be confused. Face validity refers to the surface credibility of a test. In other words, it concerns whether an assessment seems appropriate for its particular purpose. Face validity is not so much concerned with technical validity, in other words what the test actually measures, but rather with what the test appears to measure (Anastasi, 1976:139). A test that looks authentic has face validity (Jones, 1979:51, Bachman, 1990:307).

There is a risk, however, that teachers may choose a measuring instrument simply because it looks valid without investigating the original construct and context for which the assessment was intended. An instrument that looks valid on the surface is not necessarily representative of the construct domain to be tested. Developers and teachers must be careful not to rely on a quick overview to determine whether an assessment is content valid or not.

The content approach to validity is useful when interpretations are made on the basis of a well-defined construct domain, but not so for interpretations outside the specified domain (Kane, 2004:138).

It can be seen as contributing to content validity, but face validity alone is not sufficient to establish content validity. Many researchers have spoken out against the use of face validity as a means for justifying test interpretations (cf. for example Lado, 1975; Bachman, 1990). However, poor face validity may influence performance, making the assessment seem less serious and less credible than what it is. Performances influenced by poor face validity will result in an inaccurate picture of learners' performances and in turn produce invalid scores.

The major limitation of content validity identified by Bachman (1990:247) is that it focuses on the test, as opposed to actual learner performances, test scores and how these are interpreted. Bachman (1990:247) notes that showing the content of an instrument to be representative of the construct domain does not entail considering how learners actually perform on the assessment. Content validity is a characteristic of the test itself, and since the content of the test does not change, neither will its content validity. However, the individuals taking the test do change, as well as the context and the way that the test results are interpreted and used (Hambleton, Swaminathan, Algina & Coulson, 1978:38-39; Bachman, 1990:247).

Underhill (1987:106) sees little difference between content and construct validity. He equates content validity with determining the extent to which test content reflects the course syllabus and programme outcomes. According to Underhill (1987:106), the test developers' knowledge and judgment of the implicit objectives of the course largely determines validity.

Test developers tend to use content and face validity as touchstones of test validity (Fulcher, 1999:223-224). Stevenson (1985:111) objects to such "naïve, face-valid judgments" about what language tests measure. Firstly, it reinforces the misguided notion that validity lies only within a test (a traditional view that is opposed by modern interpretations of validity, as discussed below). Secondly, defining a target domain is very difficult.

Bachman (1990:245) notes that language test developers rarely have a clearly and specifically defined domain that unambiguously identifies the relevant language tasks from which a test

can and should be sampled. Fulcher (1999:223) also criticises such a restricted focus on content and face validity to ensure validity of language testing. This narrow focus leads to a simplistic view of validity, namely that an authentic test that looks valid is valid. Both Bachman (1990:245) and Fulcher (1999:224) mention the almost endless list of additional factors, such as physical conditions, that form part of the testing domain and need to be specified to present a sufficiently specific definition of the domain.

2.2.1.3 Construct validity

Formally introduced in 1955 by Cronbach and Meehl, construct validity was simply described as applying scientific theory to either prove or disprove the interpretation of scores, drawing together requirements for a rational statement and empirical verification of the statement (Shepard, 1993:416; Ryan, 2002:282-292). Cronbach and Meehl (1955:28) define a construct in terms of a theory that shows a relation between a particular construct and other constructs, and to observable performances. A construct is "a postulated attribute of people, assumed to be reflected in test performances" (Cronbach & Meehl, 1955:283). In other words, a construct is the specific ability, linguistic structure or aspect of a skill that test developers aim to measure with a specific instrument.

However, the concept of *construct validity* is more complicated than this seemingly simple description. According to Fulcher and Davidson (2007:7), what makes defining construct validity difficult is firstly defining what constitutes a "construct". The term *construct* does not refer to a physical ability, but rather to an underlying ability that can only be investigated by observing behaviour (Fulcher & Davidson, 2007:7) and "is hypothesised in a theory of language ability" (Hughes, 1989:26). Ebel and Frisbie (1991:108) provide the following description of a construct:

> The term construct refers to a psychological construct, a theoretical conceptualization about an aspect of human behaviour that cannot be measured or observed directly. Examples of constructs are intelligence, achievement motivation, anxiety, achievement, attitude, dominance, and reading comprehension. Construct validation is the process of gathering evidence to support the contention that a given test indeed measures the psychological construct the markers intend it to measure. The goal is to determine the meaning of scores from the test, to assure that the scores mean what we expect them to mean.

Construct validity refers to the extent to which the relevant psychological structure that underlies a performance – such as language ability – is being measured (Brualdi, 1999:2). It also concerns the extent to which a measurement is based in theory of language ability and measurement. Garson (2006:2) explains:

> A good construct has a theoretical basis which is translated through clear operational definitions involving measurable indicators. A poor construct may be characterized by lack of theoretical agreement on its content, or by flawed operationalisation such that its indicators may be construed as measuring one thing by one researcher and another thing by another researcher. A construct is a way of defining something, and to the extent that a researcher's proposed construct is at odds with the existing literature on related hypothesized relationships using other measures, its construct validity is suspect. For this reason, the more a construct is used by researchers in more settings with outcomes consistent with theory, the more its construct validity.

A construct is a way of classifying behaviour, providing a definition of an ability that allows us to theorise about how that ability relates – or does not relate – to other abilities and to observed behaviour (Cronbach, 1971; Bachman 1990). Fulcher and Davidson (2007:7) state that concepts become constructs in the following manner:

> Concepts become constructs when they are so defined that they become 'operational' – we can measure them in a test of some kind by linking the term to something observable ... , and we can establish the place of the construct in a theory that relates one construct to another.

Some constructs are easy to relate to in everyday life – such as "human being" – while others are "embedded in well-articulated, well-substantiated theories" (Cronbach, 1971:462; referenced by Bachman 1990:255).

Language constructs such as writing ability are latent traits that cannot be observed directly. These must therefore be measured indirectly through observing the behaviour of writing, elicited, for example, by an appropriate test (Henning, 1991:183). Such abilities are theoretical because we theorise that they affect the way language constructs are performed. Bachman (1990:256) describes the extent to which we can make these inferences about hypothesised abilities from language performances, such as those produced in a test, as the essential issue of construct validity. In investigating construct validity, the aim is to test hypothesized relationships between scores and abilities empirically.

Performing a construct is influenced by psychological processes as well as the context in which the construct is performed. Construct validity is a function of the interaction between the linguistic and cognitive processes involved when performing the construct and the performance context. In order for an instrument to be construct valid, the construct that is being tested must be theoretically and psychologically real (Hughes, 1989:27).

The construct model was traditionally only considered useful when evidence for criterion- and content validity was not available. Construct validity was regarded as a "last way out"; the last tool if all others failed. Shepard (1993:416) describes the early version of construct validity as "too demure and too ambitious" in comparison with later interpretations of the concept. Cronbach and Meehl (1955) describe construct validity as the weak sister of the previously dominant view of validity, presenting it as an option only to be used as alternative when criterion and content validity models failed (Shepard, 1993:416; Kane, 2004:138). "[C]onstruct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not operationally defined ... and for attributes for which there is no adequate criterion" (Cronbach & Meehl, 1955:282-299).

During the 1970's, researchers such as Messick (1975; 1981) and Cronbach (1980b) noted the problematic application of the toolbox-approach to providing validity evidence. Developers chose to provide the most readily available validity evidence in a piecemeal fashion, rather than providing the most appropriate and relevant evidence of a test's validity.

2.2.2 A critique of the traditional conception of validity

Towards the end of the 1970's and during the 1980's language testers approached test qualities in a more sophisticated manner and used a wider range of analytical research tools than before. In educational measurement, the traditional definition and scope of validity came into question (Chapelle, 1999:255-256).

Brualdi (1999:2-3) notes that the traditional view of validity has been criticised for being fragmented and incomplete (cf. Messick, 1989; 1996), as it ignores evidence of the social implications of score meanings and the consequences of decisions based on the scores.

The 1985 American Educational Research Association, American Psychological Association and National Council on Measurement in Education Standards (AERA, APA & NCME, 1985:9) describe validity according to the three categories, but cautions that although different category labels are used, the categories should not be interpreted as separate types of validities. Loevinger (1957) criticises the individual parts of the validity scheme – content, predictive, concurrent (or criterion-related validity) and construct validity – for not being clearly and logically distinct or carrying equal weight. She argues that the parts represent options of validity rather than components of validity. She suggests that content and criterion-related validity rather serve as supporting evidence for construct validity. Only construct validity supplies a scientific basis for establishing the validity of an assessment instrument (Loevinger, 1957; cf. also Moss, 1992).

Performance assessment presents problems related to validity that cannot be handled sufficiently according to the traditional validity models. For example, students are allowed more freedom in interpreting and responding to tasks. Moss (1992:231) points out that learners' responses to these tasks become more complex as learners become more proficient and integrate different skills and knowledge. Issues surrounding reliability, generalisability and comparability as defined according to the traditional validity model become difficult to handle. Concerns about the social consequences of how test scores are used provide different criteria for validity than the traditional validity criteria, which results in tension (Moss, 1992:231).

During the 1970's the interrelatedness of the three types of validity in theory was recognised in the AERA, APA and NCME Standards, which stated that it seldom happens that only one of the three types of validity is important in a particular situation. The 1974 Standards document indicates that the different aspects are discussed independently only for the sake of convenience, but they are logically and operationally interrelated (AERA, APA & NCME, 1974:26).

Researchers such as Guion (1980) opposed the *holy trinity* approach to validity, which oversimplifies the principles of validity. Landy (1986) equated traditional validity practices with stamp collecting, where a test is pasted into the content, criterion or construct space. Guion (1980), Landy (1986) and later Cronbach (1988) and Messick (1989) suggest a unified view of validity as solution to the problem of fragmentation. Today, a unified view of validity

based on the construct validity model, is the generally accepted approach to validity (Shepard, 1993:415; Kane, 2004:138).

2.3 The modern conception of validity

Cronbach (1980, 1988) and Messick (1980, 1988, 1989) were primary influences in the movement towards expanding the concept of *validity* to include socially related issues. They led the shift in the conceptualisation of validity by emphasising the inferences and decisions made from test scores. Messick (1989) suggests that the evidential and consequential basis of interpretations and uses of test scores be examined. Cronbach (1988) proposes that validity be investigated from political, functional, economic and explanatory viewpoints.

When investigating consequential validity, one is interested in finding out whether the scores, interpretations of scores and impact of scores are valid. Fulcher and Davidson (2007:35) point out that "[t]he usefulness of assessment, the validity of interpretation of evidence, is meaningful only if it results in improved learning".

The traditional interpretation of validity in terms of three different entities was abandoned in favour of a view of validity as unitary concept which poses construct validity as central and content and criterion validity as components of construct validity. Van der Walt and Steyn (2007:139) describe this view as "a more naturalistic and interpretative one", considered as currently the most influential theory of validity.

Chapelle (1999:256) highlights three major developments in the 1980's that steered validity research into this new frontier. The AERA, APA and NCME Standards for educational and psychological testing revised their definition of validity as a single unified concept with construct validity as central, as opposed to the traditionally accepted definition of three validities. The 1985 Standards define validity as "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (AERA, APA & NCME, 1985:9).

Validity was no longer equated with correlation, but content and correlation analyses were suggested as ways of investigating construct validity. Researchers such as Cherryholmes (1988) started questioning the philosophical underpinnings of ways to establishing validity.

Messick's (1989) seminal paper emphasised both these points and articulated a definition of validity that incorporated research related to construct validity as well as test consequences. As a result, the issue of test consequences was taken seriously enough to cause widespread debate for the first time, although the notion was not new (Chapelle, 1999:256-257).

The idea of construct validity as an encompassing or umbrella category for test validity was only strongly promoted after the publishing of Messick's paper in 1989. It was not strongly supported earlier in the twentieth century, but as early as 1957, researchers such as Loevinger interpreted construct validity as an overarching term for validity as a whole: "... since predictive, concurrent, and content validities are all essentially ad hoc, construct validity is the whole of validity from a scientific point of view" (Loevinger, 1957:636). Today the unified view of validity is generally accepted and advocated (e.g. Bachman, 1990, 2004; McNamara, 1996, 2000; Kane, 1990, 2004; Shaw & Weir, 2007).

In essence, Messick's paper serves two main purposes. Firstly, it establishes validity as a unified concept. Secondly, it broadens the meaning of the concept of *validity* beyond the meaning of scores to include relevance, utility, value implications and social consequences. Messick (1989), supporting Cronbach (1971), demands that validity supports inferences as well as actions based on test scores. He further explicitly states the need for considering implicit assumptions about what an assessment instrument will accomplish (cf. also Shepard, 1993:423).

Traditionally, the main concern of validity was with fitness for purpose, in other words, the appropriateness of a test for the particular purpose of assessment. Originally, questions of validity asked: "Does the test measure what it is supposed to and claims to measure?" With the adoption of Messick's (1989) interpretation of validity, the main emphasis shifted to include an awareness of the social impact of assessment, and therefore onto validity as property of score interpretations rather than of the instrument.

The traditional validity question was reformulated as follows: "What is the evidence that supports particular interpretations and uses of scores on this test?" In modern terms, the social impact of how scores are interpreted and used is seen as a validity concern in addition to the assessment's suitability for the assessment purpose. This view of validity leads to the consideration of the test's consequences (Alderson & Banerjee, 2002:79). Brualdi (1999:1)

therefore describes validity as the degree to which inferences based on scores are useful, meaningful and appropriate for the particular population, purpose and context of the administration.

Messick (1989) argues that it is not the test properties that show whether an assessment is adequate, but the results of the test (responses, scores and how scores are interpreted and used): "Tests do not have reliabilities and validities, only test responses do" (Messick, 1989:14). In other words, validity is considered as a property of the test scores interpretations rather than residing in the test *per se* (cf. Weir, 2005:12). It is considered to be inherent in the interpretation and uses of an assessment instrument, rather than a property of the assessment instrument itself (Bachman, 1990; Fulcher, 1999; Weigle, 2002; Weir, 2005).

Messick (1989, 1996) uses the term "construct validity" as an over-arching term to refer to all different aspects of validity. He defines validity as "an integrative evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other methods of assessment" (Messick, 1983:13). Messick (1989) uses a progressive matrix illustrated in Table 2.1 to explain validity and the process of establishing validity (cf. Chapter 3).

	Inferences	Uses
Evidence	Construct validity	Construct validity + Relevance/ utility
Consequences	Construct validity + Value Implications	Construct validity + Relevance/ utility + Value Implications + Social consequences

Table 2.1 Progressive matrix for defining the facets of validity (adapted from Messick, 1989:20)

The matrix is progressive in the sense that each cell moves forward, while incorporating or containing the previous cell. Construct validity is the central feature of each cell and forms the overarching concept in the evaluation of language tests (Fulcher, 1999:225). In each cell, construct validity appears with an additional aspect added to it. In terms of Messick's approach, a convincing argument for the validity of test score interpretations and uses should address construct validity, relevance/utility, value implications and social consequences (Lee, 2005:2).

The first cell illustrates Messick's suggestion that construct validity forms the evidential basis for test interpretations or inferences (Messick, 1989:20). The consequential basis of using an instrument (bottom left cell) refers to evaluating the potential and actual social consequences of the particular setting. It requires considering the value assumptions implied by the concept labels and theoretical framework that guides the validity investigation explicitly (Shepard, 1993:424).

The consequential basis of test inferences (top right cell) refers to the consideration of the value implications of the construct label, the theory underlying the score interpretations, and the ideologies in which the theory is embedded. Here Messick addresses questions about the meaning of the construct and relationships that support the application of the test. The validity of outcome criteria must be investigated, considering their relevance, representativeness and multidimensionality. Statistical correlations alone are not enough to show that variance shared with a criterion is relevant to the construct and not due to a shared bias (Shepard, 1993:425).

Finally, Messick addresses the consequences of test use in the last cell. Both intended and unintended consequences must be investigated for validity. According to Lee (2005:2), construct validity, supported by evidence that the assessment is relevant for the particular purpose and setting, also forms the evidential basis of tests (bottom right cell).

Construct validity should be understood as a super-ordinate category of description referring to the various aspects of validity. It no longer only refers to the cognitive trait or theoretical construct on which the assessment is based, but also concerns aspects such as relevance and utility of an assessment, value implications, and social consequences of an assessment. Construct validity is a function of the interaction between the context, linguistic features and cognitive procedures used when a task is performed. Scoring validity, criterion validity and consequential validity are considered additionally, as Weir (2005:19) points out:

> Context validity is concerned with the extent to which the choice of tasks in a test is representative of the larger universe of tasks of which the test is assumed to be a sample. This coverage relates to linguistic and interlocutor demands made by the task(s) as well as the conditions under which the task is performed arising from both the task itself and its administrative setting.

This interpretation by Weir (2005:14-15) supports the idea proposed by Messick (1989) that validity is not only inherent in the test itself, or in the scores alone, but also in the inferences and decisions made, based on the results. Weir (2005) furthermore suggests substituting the traditional term "content validity" with "context validity", because it implies the social dimension of language testing more strongly. Context validity then consists of content validity, face validity and response validity (Alderson et al., 1995:172-177).

As Henning (1991:284) points out, the construct being assessed must be valid for the purpose of the assessment and the scores of the assessment must give an accurate reflection of the construct. In addition, the responses produced by learners must be relevant to the elicited construct.

Response validity concerns the appropriateness of learner responses to test prompts as a result of the cognitive procedures they apply to produce the response. Response validity can be tied to one of the guidelines for validity suggested by Anastasi (1976), namely to determine whether a response is appropriate in terms of the construct being measured, rather than to the content of the task. Various factors may influence responses, such as levels of motivation, how well instructions are understood and how willing learners are to obey all the restrictions and exam conditions (Wall, 1991:220).

Messick (1989) argues that content is a central validity issue. He refers to Ebel (1983:8), who summarises the argument for the centrality of content in validity. Addressing content validity means providing the rationale for an assessment. It involves providing a written document that (a) defines the construct to be measured, (b) describes the task to be included in the test, and (c) explains why such tasks are used to measure the specific ability. According to Fulcher (1999:222), content as a central issue of validity supports the real-life approach to validity and means establishing the degree to which an assessment accurately samples the relevant construct domain in a setting that represents the content and format of real life tasks as closely as possible. All tasks should be representative of tasks from a well-defined domain (Bachman, 1990:310).

Lee (2005:2) points out that test use, as addressed in the second column in Table 2.1, was not considered in relation to validity before Messick emphasised the relevance of issues such as the misuse of tests, social consequences and test fairness. These issues are crucial under

Messick's framework. If negative consequences result from a test, the validity of the use of the test becomes questionable (Messick, 1989:20; Lee, 2005:2).

Messick also highlights the multi-faceted nature of validity as it is understood in the modern sense. Messick (1996:9-15) distinguishes six aspects of validity, viz. content and substantive validity (content validity), structural validity (theory-based validity), generalisability (scoring validity), external validity (criterion validity), and a consequential aspect (consequential validity). Douglas (2000:257-258) illustrates the complex nature of validity by comparing it to a mosaic:

> [E]ach piece of ceramic or glass is different, sometimes only slightly, sometimes dramatically, from each other piece, but when they are assembled carefully, indeed artfully, they make a coherent picture which viewers can interpret. The process of validation is much like this, presenting many different types of evidence which, taken together, tell a story about the meaning of a performance on our test. It is for this reason that I employ the term validity mosaic to characterize the process.

Each aspect that Messick identifies addresses a central aspect of validity. *Content validity* concerns specifying the boundaries of the construct domain (Messick, 1996:9-15). All tasks in an assessment instrument must be relevant and representative of the construct domain. An important aspect is determining the knowledge and skills that will be revealed by the task in order to guard against over- or under-representing the construct (Brualdi, 1999:3).

Substantive validity concerns the domain processes involved in performing the task. Substantive theories and process models can be used to identify the relevant processes. The assessment task must elicit an appropriate sampling of the involved domain processes and provide an appropriate sample of the domain content. Furthermore, empirical evidence must show that the elicited processes are indeed the relevant ones related to the domain and that learners do engage in these processes when performing the task (Embretson, 1983; Messick, 1989; Brualdi, 1999).

The *structural aspect* of validity refers to the notion that the theory of the construct domain should guide the selection of tasks, scoring criteria and rubrics. Thus, the structure of the assessment instrument must be consistent with what is known about the in internal structure of the construct domain (Messick, 1996:11). Brualdi (1999:4) explains that the way in which

the performances are scored should be based on how the implicit process of the learners' actions "combines dynamically to produce effects".

Generalisability means that the assessment must provide representative coverage of both the content and the processes of the construct domain so that the score interpretations are not limited to the sample of tasks in the assessment, but can be related to the broader construct domain.

The *external aspect* of validity relates the assessment to other external measures of the same construct. The construct should account for any correlation pattern between the assessment and the external criterion. Score interpretations must be supported externally by evaluating the degree to which empirical evidence supports the meaning of the scores. Construct theory indicates how relevant the relationship between the scores and the criterion measure are (Messick, 1992:12; Brualdi, 1999:4).

Finally, *consequential validity* is the aspect of construct validity that concerns evaluating both intended and unintended consequences of the score interpretation and use. The potential and actual consequences of test interpretations and uses should support the purpose of the assessment and be consistent with social values of the time and context (Messick, 1989:18). Messick (1989:18; 1996:12-13) argues that the relevant social values considered in the outcomes, inferences and uses of scores derive from and contribute to the meaning of test scores. Thus, consequential validity is an aspect of construct validity, as Messick (1989:18) explains:

> For a fully unified view of validity, it must also be recognised that the appropriateness, meaningfulness, and usefulness of score based inferences depend as well on the social consequences of testing. Therefore social values and social consequences cannot be ignored in consideration of validity.

Any negative consequence of test score interpretations should not be relatable to construct underrepresentation or construct irrelevance. In order to enhance positive effects, test developers should investigate the construct representativeness of the instrument (Messick, 1996:13). Messick (1989:160) suggests that the best way to combat adverse social consequences is to minimise potential sources of invalidity during the measurement process, especially construct underrepresentation and construct-irrelevant variance. Furthermore,

decisions made based on test scores must be in line with social values and ideas of the time, because social values influence interpretations of test scores. As these change through time, so will the meaning of test scores for certain situations.

More recent publications, such as Weir (2005) and Fulcher and Davidson (2007), echo Messick's concern about the social impact of the interpretation and application of test results and stress that test administrations reflect social views and goals. The impact of test scores refers to "the effect that tests have on individuals (particularly test takers and teachers) and on larger systems, from a particular educational system to the society at large" (Weigle, 2005:53). Therefore, test results have certain consequences outside the testing situation for various parties involved. According to Weir (2005:214), the impact of tests on society and on people's lives is possibly the most difficult aspect of consequential validity to investigate (Messick, 1988, 1989; Bachman & Palmer, 1996; Alderson & Banerjee, 2002; Weigle, 2002; Weir, 2005; Fulcher & Davidson, 2007).

The influence of test results on the society at large is often overlooked, because it entails investigating the consequences of the test on stakeholders who are not directly related to the test itself. Tests can be, and often are, used as controlling tools. According to the critical language testing view (cf. Shohamy, 2001), tests are instruments of power that tend to be biased, unethical and unfair. They could be used for imposing constraints, restricting curricula, as disciplinary tools or tools for promoting a political agenda, and to encourage mechanical teaching. The fact that such extreme views persist about assessment and testing further emphasises the importance and necessity of carefully validating assessment.

Validity is a unified concept with the unifying force being the meaningfulness or trustworthiness of the interpretability of test scores and acting implications; in other words construct validity. Messick's six aspects of construct validity discussed above can be used to investigate validity to ensure that issues implicit in the central notion of unified validity are addressed. Contrary to the toolbox approach, the unified approach to validity does not allow for selectively providing validity evidence. The different aspects are not separate aspects that can be substituted from one another; rather they are interdependent, complementary forms of validity evidence that exist interdependently (Brualdi, 1999:3). Messick (1996:15-16) emphasises that none of these aspects is insufficient by itself, nor required to ensure validity. What is required is a convincing argument that whatever evidence is available is enough to

satisfy the intended score interpretations and use. The six types of validity offer a base to check that all necessary aspects are considered to establish the validity of a testing instrument.

Messick (1995:742) suggests that the different sources of validity be interpreted as "an evaluative summary of both the evidence for – and the actual and potential consequences of – score interpretation and use (i.e. construct validity conceived comprehensively)". He considers content, criteria and consequences, and integrates these into a unified model for empirically testing hypotheses about the meaning and uses of scores.

A test is not either completely valid or not. Rather, a test can be more valid for one purpose than for another, and more in one context than another. Douglas (2000:257) describes validity as to do with interpreting test scores in the light of the test purpose and, consequently, instead of questioning the validity of the assessment as such, the question should rather be whether the assessment is valid for that *specific purpose or situation.* Validity is a relative rather than a concrete concept. It is therefore important to prove the validity of a test for every different purpose it is used for and context it is used in (Van der Walt & Steyn, 2007). Bachman (1990:26) notes that validity depends on various factors outside the test itself and can therefore not be proven indefinitely. Fulcher and Davidson (2007:26) also suggest that there is no absolute answer to the question of validity. Interpretation and uses of scores must be proven valid for a particular administration. They point out that this truth may change with time and as new evidence is provided. Validity remains a complex, multi-faceted concept.

According to Messick (1989:33), the strength of one type of validity may differ from one administration of a test to another. The amount of evidence available to support each of the different types of validities may also differ. Alderson et al. (1996:188) conclude that an assessment must produce fair and accurate results and therefore both validity and reliability should be ensured in as many ways as possible.

Messick (1989, 1996) highlights authenticity and directness as standards of validity. Authentic and direct testing improves the chances of positive test consequences for teaching and learning. According to Fulcher (1999:222), the accurate sampling of real-life tasks and situations relates to the authenticity of an assessment. Authenticity in language testing refers to the degree to which the real world is brought into the testing situation. An authentic test

task looks like real-life tasks and relevant test content represents real-life situations (Bachman, 1990:307; Fulcher, 1999:222). Tasks are presented in a realistic setting that closely simulates the tasks, processes, available time and resources of the real world (Messick, 1992:3). The major concern here is that nothing important and relevant to the construct domain is left out of the assessment of the construct. Nothing irrelevant must be added to the assessment. In order to ensure positive consequences, sources of invalidity must be minimised (Messick, 1992:3-4).

Authenticity and directness relate to the two main threats of validity, namely construct underrepresentation and construct irrelevant variance (Messick, 1989, 1994, 1996; cf. also Fulcher, 1999; Brualdi, 1999; Weir, 2005). Construct underrepresentation means that a test is too narrow and does not include all the relevant dimensions of the construct (Messick, 1989:34). Construct irrelevant variance means that a test is too broad and includes variance that is irrelevant to the interpreted construct. It implies that a test measures too many and some irrelevant aspects of a construct. Some factors that are irrelevant to the construct may cause test scores to differ systematically, causing markers to attribute invalid meanings to test scores (Fulcher, 1999:226).

Within this second type of invalidity one can distinguish between "construct-irrelevant easiness" and "construct-irrelevant difficulty" (Messick, 1989:34; Brualdi, 1999:5) According to Brualdi (1999:5), these two forms entail that certain information or aspects of a task allow either appropriate but irrelevant answers, or make a task irrelevantly difficult. Construct-irrelevant easiness causes learners to score higher than they normally would when performing the ability, whereas construct-irrelevant difficulty causes learners to score lower than they normally would.

Questions of validity are questions for evidence and arguments that discount these two major threats of construct validity (Messick, 1989:34). Construct validity is undoubtedly crucial for assessments to be useful and fair, specifically with regard to writing assessment. Hamp-Lyons (1991:11-12) calls for more research on the construct of writing and its underlying constructs, especially in the second language context. She argues that collecting and scoring written samples alone is insufficient to establish whether an assessment measures what students know or can do in writing. Rather, it is necessary to look into learners' reasons for

writing what they write, how they write it, and which factors influence their writing performances.

2.4.1 Validity and reliability

The relationship between validity and reliability has long been problematic in language testing. Both are essential aspects, yet according to the traditional approach a trade-off between validity and reliability is essential. The adoption of the unified interpretation of validity brought on a shift in roles of reliability and validity.

Reliability concerns the believability of an assessment tool. It is a feature of test scores and concerns the degree to which individuals are measured consistently by a particular instrument (Bachman, 1990:26; McNamara, 1996:136). This definition of reliability has remained stable over time. Anastasi (1976:103) describes reliability as referring to "the consistency of scores obtained by the same persons when re-examined with the same test on different occasions, or with different sets of equivalent items, or under other variable examining conditions". Henning (1991:285) regards reliability as "the capacity of the assessment procedures to rank-order the same samples of writing performance consistently in the same way". According to Jones (2001:1), a reliable test is dependable since it can be expected to produce similar results on different occasions. For Weir (2005:23), reliability concerns how unbiased, stable and consistent test results are produced in a situation, and from one situation to the next.

Questioning reliability means asking whether the assigned scores are accurate and can be used to make inferences about learners' abilities. Bachman (1990:20) suggests that, if test tasks are regarded as test characteristics, then reliability can also be regarded as "a function of consistencies across different sets of test task characteristics". At least two sets of scores are required to make comparisons.

Reliability indicates how accurate an instrument measures the construct. Furthermore, it estimates how accurately learners' abilities are distinguished and classified. "Thus, the most reliable test will be those that best distinguish among students in the ability measured" (Henning, 1991:288). 'Error variance' refers to any condition irrelevant to the purpose of the test that may influence the score produced by a learner. These factors must be identified for each unique situation (Alderson et al., 1996; Anastasi, 1976).

A reliable measuring instrument is free from errors of measurement (AERA, APA & NCME, 1985; Bachman, 1990). Learners' test performances are influenced by various factors other than their own ability to perform the task. Factors that may interfere with performances include: the time of day, learners' background, motivation, fatigue, and raters responsible for scoring performances. Rater characteristics, rating scales and scoring procedures may also influence scores. These factors are potential sources of error. If the interference of such factors is minimised, the reliability of the test is maximised, because the test then gives a more accurate reflection of the learners' abilities (Bachman, 1990:161).

A completely reliable instrument only measures systematic changes and provides a "true score", as opposed to unsystematic changes. "Real" or systematic differences refer to actual changes in learners' abilities (Alderson et al., 1996:86). The true score on a test is an "idealised version" (Huot, 1990:202) of the score a learner should receive on a test. The true score perfectly reflects the learner's ability and would be exactly consistent with any future performance on the same test. It purely reflects the learners' ability free from any interference from outside that causes error. "Error" or unsystematic differences refer to influences such as learners' physical or mental fatigue, test circumstances and inconsistent rating (Alderson et al., 1996:86). In other words, a scale's reliability is an indication of the degree to which it measures differences in performance which are actually a result of change or development of the trait being measured, instead of due to chance or error. However, error is inherent in any assessment instrument (Huot, 1990:202) and all test scores are functions of the true score and some error component.

In the traditional view, reliability and consistency were regarded as distinct from validity. Tests are traditionally seen as valid if they consistently produce accurate measurements and therefore reliable. Thus, validity depends on reliability. Inconsistent measurement cannot be accurate. Bachman (1990:161) notes that increasing the reliability of a test satisfies a "necessary condition for validity: in order for a test score to be valid, it must be reliable".

Reliability is a fundamental criterion against which all language tests have to be judged (Weir, 1990:31). However, proving that a test is reliable is not enough evidence for the overall validity of a test, but a test must be proved reliable in order to establish other empirical types of validities, viz. concurrent validity, predictive validity and construct

validity (Bachman, 1990; Henning, 1991; Brown & Hudson, 2002). A reliable test is not necessarily valid, as it may give consistent results, but may not be measuring what it claims to, or may not be appropriate and meaningful within the context in which it is used. According to traditional views of reliability, the reliability of a test would have to be reduced in order to increase validity (Alderson et al., 1996:187; Hughes, 1989:42; Chapelle, 1999:255).

This competitive relationship between validity and reliability has raised many questions. Alderson et al. (1996:188), for example, point out that the differences between the two concepts are not clearly specified. Furthermore, test users must understand what a reliability index reveals (technical information) in order to decide whether a test is suitable for a specific purpose. Huot (1990:202) points out that assumptions and confusion about validity and reliability promoted the neglect of validity: "The emphasis on reliability fuelled some assumptions and confusion about the concept *validity*" (own italics). For example, the assumption was made that in establishing a common set of criteria for rating writing performances, both reliability and validity will be achieved. The confusion linked both validity and reliability to a common set of criteria (reliability), while neglecting to consider the characteristics which pertain to a true assessment of writing quality (validity). McColly (1970:149) also noted:

> It is often said that reliability is the more important of the two ... But really the inverse relationship is true. Be that as it may. The scholarly literature that deals with writing tests shows more apparent interest with reliability than with validity.

These assumptions prospered despite attempts to establish validity as the most important attribute (e.g. Lyman, 1978); and to establish reliability in itself as an insufficient condition for test validity (Popham, 1981).

Messick's unified view of validity provides a solution to the problematic relationship between validity and reliability. Messick (1989) suggests that reliability be regarded as an aspect of validity, instead of a separate (and to some extent contrasting) aspect to be investigated. An increase in reliability should be interpreted as one type of evidence for the overall validity of an assessment, as opposed to competing against other types of validity evidence. Weir (2005:43) supports this notion, explaining that reliability and validity are no longer polarised in the modern understanding of validity. Rather, reliability and validity are

seen as parts of a unified approach to establishing the overall validity of a test (Chapelle, 1999:258; Weir, 2005:43).

Scoring validity (referred to as "external validity" by Alderson et al., 1995) concerns issues regarding the procedures around producing and interpreting scores. Validity and reliability are regarded as two complementary aspects that are both concerned with identifying and minimising the effects of external factors on test performances (Bachman, 1990). A test must provide a measure that can be interpreted as a trustworthy indicator of a person's language ability, in other words it must produce fair and accurate scores (Alderson et al., 1996:188; Bachman, 1990:23).

Weir (2005:43) also uses the term *scoring validity* as superordinate to "locate reliability more centrally in the validation process". Scoring validity typically relates to aspects such as the reliability of scoring procedures and raters' ability to score performances consistently, along with other statistical elements (Weir, 2005:47). It entails investigating whether test scores give an accurate reflection of the learners' ability to perform the construct and whether the scores are consistent measures of learners' abilities (i.e. reliability). It is directly related to construct validity: "[T]he validity of score interpretations is dependent on the fidelity of the construct that is measured by the test and the resulting test scores" (Lane, 1999:3; also cf. Messick, 1989).

The reliability of test scores can be damaged in numerous ways. First, it is impossible to gain reliable test scores from an unreliable test (Hughes, 1989:36). Second, rating scales can influence the meaning of scores because they may either contain items that are irrelevant, or not contain certain items that are important. In such cases, investigating rater bias and consistency serves little purpose because the content of the assessment does not reflect the construct specified in the test specifications (cf. 2.2.1.2 and 2.2.1.3) (Lane, 1999:6). Third, scoring methods influence the meaning of scores. Different scoring methods, such as primary trait, holistic or analytic, focus on different characteristic of a written performance, so each method of scoring assigns different meanings to the scores (Lane, 1999:7).

2.4.2 A critique of Messick's unified model

The modern view of validity, strongly influenced by Messick's views, has not gone uncontested. First, Messick's unified model is considered to be very complex. Second, the necessity for a unitary interpretation of validity is questioned. Third, questions have been raised about whether the consequences of tests really are a matter of validity: Can, and should, test consequences rightfully be incorporated into the concept of *validity*? Fourth, Messick's focus on the test score as determining factor for a test's validity has caused much controversy and the debate is still unresolved.

The main concern with Messick's concept of validity is that it is too complex and impractical. Messick's unified model is sometimes opaque and his matrix has been described as "incomprehensible" and "demanding" (Shepard, 1993:429).

Shepard (1993) does not inherently disagree with Messick's argument for construct validity, but shows concern about the faceted presentation of the fourfold matrix (Table 2.1). According to Shepard (1993:426), this presentation invites a new segmented view of validity requirements. Shepard raises three concerns: First, the table may create the impression that values are regarded as separate from scientific evaluation of test scores. The two rows in the table create the impression that one would first resolve the scientific validity issues and then consider the value implications. However, this is not what Messick suggests. According to Messick (1989:62), "scientific observations are theory-laden and theories are value-laden". Scientific and value issues of validity are dealt with simultaneously.

Second, with construct validity being named in the first cell and repeated in the following cells, it is not clear whether the term refers to the part or the whole. According to Shepard (1993:427), Messick seems to imply a narrow definition of construct validity as score meaning (the first cell only). However, Shepard (1993:428) suggests that construct validity be equated with all the demands implied by the four cells, and not just a narrow definition of score meaning.

Third, the complex analysis does not help to identify the most important or relevant questions to address in order to support a specific test use. Shepard (1993:429) raises a pragmatic concern that the sequential approach may mislead the conceptualisation of theoretical

frameworks used to guide validity investigations. Shepard argues for a more straightforward means of identifying the most pressing validity questions.

Although the unified view works well in theory, it has contributed little towards ensuring the validity of assessment instruments in practice. Test validators need specific instructions for the procedure of collecting validity evidence. This calls for a reformulation of Messick's framework, providing a simpler model for prioritising questions of validity (Shepard, 1993:429; Kane, 2004:136; Lee, 2005:2).

Messick's model does not provide clear guidelines or steps for validating a test. Borsboom, Mellenbergh and Van Heerden (2004:1061) and Lissitz and Samuelsen (2007:437) comment that validity theorists and practical researchers seem to have different concepts in mind when investigating validity. Current validity theory "fails to serve either the theoretically oriented psychologist or the practically inclined tester" (Borsboom et al., 2004:1061).

Even if one accepts construct validity as a framework for a unified model, there are still some drawbacks. Messick does not regard validity as merely a characteristic of the test, but rather as an argument or series of arguments for the effectiveness of an assessment for a specific purpose (McKay, 2000:193; cf. also Hudson & Brown, 2002; Kane 2004). Kane (2004) also advocates validity as an 'argument' and warns that the original unified model is very open-ended. The lack of a well-articulated theory for the construct makes it very difficult to know where to begin and where to end with providing validity evidence. "Construct validity has been useful as a unifying framework on a theoretical level, but has not, in itself, been an effective unifying influence on an operational level" (Kane, 2004:140).

The difficulty of applying Messick's theory may be attributed to the many different kinds of validity evidence available and the lack of explicit criteria for prioritising the different kinds of evidence. Construct validation has therefore been described as an unreachable goal (Kane, 1992, 2004; Shepard, 1993; Brennan, 1998; Fremer, 2000; Lee, 2005; Lissitz & Samuelsen, 2007; Embretson, 2007). Shepard (1993:429) comments that the complex model emphasises that construct validation is a never-ending process. Although true, reaching construct validity is presented as an unattainable task. Practitioners therefore tend to resort to only providing partial validity evidence, not necessarily addressing all types of validity. Shepard (1993:429) points out that current standards are not organised coherently in a conceptual framework that

provide clear guidance for prioritising validity evidence. Validity evaluations should be organised to answer the question of what the testing practice claims to do.

Some critics of the unified model, such as Crocker (2003) and Lissitz and Samuelsen (2007), suggest returning to the notion of content validity. Lissitzs and Samuelsen (2007) also propose a change in vocabulary to remove the concept of *construct validity* from the meaning of tests.

Gorin (2007) is more optimistic about the current view of validity. She is concerned that the controversial shift in validity theory proposed by Lissitz and Samuelsen does not "move the domain forward", but can rather be seen as a step backwards to historically problematic measurement theories and practices.

> [A] return to the use of content validity as the whole of validity theory threatens to stifle many of the recent advances in test design, resulting from construct-centric models of validity. Historically, the use of content validity tools such as operational definitions as indicators of score meaning has been tried and discarded (Gorin, 2007:457).

Others, such as Borsboom, Van Heerden and Mellenbergh (2003), Gorin (2007), Moss (2007) and Weideman (2009), disagree with the notion of content validity as the most important aspect of validity.

Weideman (2009:2) calls for serious critical evaluation of the assumption that language testing needs a unitary concept of validation. He questions the perception that there is a need for a unifying concept of validity, as well as the choice of construct validity as unifying concept and the basis on which primacy is attributed to it. Messick (1988:35, 40f.; 1981:9; 1989:19) promotes the need for a unifying concept of test validity to bring all other important validity considerations into focus. Weideman (2009:2) notes that the reason for attributing this status to construct validity in particular is not very clear. He further argues that the assumption that a unifying term is necessary, and that construct validity is this concept, is not justified by the arguments that all other validity types can be subsumed under construct validity, or are less important than construct validity (Weideman, 2009:3).

The notion that validity resides in test score interpretations is echoed in Weir's (2005:12) definition of validity:

> Validity is perhaps better defined as the extent to which a test can be shown to produce data, i.e., test scores, which are an accurate representation of a candidate's level of language knowledge or skills. In this revision, validity resides in the scores on a particular administration of a test rather than in the test per se.

Yet, in order to produce the data that provide an accurate representation of learners' abilities, testers would need an instrument that is appropriate for a particular purpose, i.e. one believed to measure the ability in question. As pointed out (cf. 2.1), Alderson et al. (1995:170) comment that test instruments are often used for measuring an ability other than the one for which they were designed. Fulcher and Davidson (2007:16) note that validity in the modern sense of the word also refers to the fairness with which scores are assigned and can be regarded as an accurate measure of the trait in question.

Alderson and Banerjee (2002) question the implication that test developers can, and should, be held responsible for the use or the misuse of a test or test results. They debate whether the new term *consequential validity* is a legitimate concern or whether it is a "political posture" (Alderson & Banerjee, 2002:79). Lissitz and Samuelsen (2007) note that it is sometimes important to consider the impact of a test on stakeholders. However, they argue that any unwanted or unintended impact that the test may have "should not be considered relevant to the question of whether the test is valid" (Lissitz & Samuelsen, 2007:445).

Borsboom, Van Heerden and Mellenbergh (2003:3) regard Messick's argument as erroneous. "We do not see the need for a unified validity concept ... because we think there is nothing to unify" (Borsboom et al., 2004:1069). They offer an opposing view with the focus on the semantics of the concept *validity*. Borsboom et al. (2004) are concerned that the broadened view of validity has resulted in little attention being paid to the semantics of validity, in other words, little effort has been made to answer the basic question of what it means for a test to be valid. They argue that, although an overarching term may be necessary for all the different types of validities (different types of validity evidence), using (construct) validity as an 'umbrella term' causes the meaning of the concept to become unclear. In their study of truth and validity, issues of consequence and evidence were purposely ignored and the focus

placed on investigating the pure meaning of validity instead. Borsboom et al. (2003:3) propose the following definition of validity:

> A test X is valid for the measurement of attribute Y, if and only if the proposition 'Scores on test X measure attribute Y' is true.

This definition suggests that validity is indeed an inherent characteristic of the test in question, and that it is not applicable only to the interpretation of test scores. However, Borsboom et al. (2003) claim that their definition of validity is not inconsistent with the idea that score interpretations are validated as opposed to the tests themselves. They conclude that it is merely "not necessary to characterize validity as a property of a test score interpretation, rather than of tests ... the semantics of validity can be clarified by looking at the conditions that would make the proposition true" (Borsboom et al., 2003:3-4).

Lissitz and Samuelsen (2007:442) also oppose Messick's claim that validity cannot be inherent in the test itself. They see Messick's claim as "essentially incorrect and confusing". Instead, they argue that test validity resides "in the definition of the test, the development phase, and any psychometric theory that gave rise to that test and its associated constructs, whether latent or manifest" (Lissitz & Samuelsen, 2007:442). They argue that the question of whether a test measures what it is supposed to measure is concerned with the meaning of test scores and should be established by examining the internal properties of an assessment instrument. The assessment construct is external, whereas content is internal. Different development procedures are used to investigate the internal and external aspects of a test.

Lissitz and Samuelsen (2007) argue that the only evidence that can support the validity of a test is internal validity (which includes reliability) and that external evidence is irrelevant to the meaning of the test. The content of a test should be the main focus of validity and investigating it should be the first and most basic step towards valid testing. Studying the test construction process should be at the heart of the validation process and be the primary concern of any one investigating the validity of a test. This includes specifying the psychometric theory associated with the assessment instrument.

Embretson (2007:450) supports Lissitz and Samuelsen (2007) in the notion that the role of external validity evidence should be minimised while internal evidence should emphasized more strongly. However, Embretson does not regard external evidence irrelevant to test meaning (as Lissitz and Samuelsen do) and argues that the term "construct validity" is still the most appropriate label to use for the validity system sought by the likes of Lissitz and Samuelsen, considering the need for multiple sources of validity evidence.

Alderson and Banerjee (2002:100-101) also question Messick's view of test validity being inherent in test scores. They argue that test traits and characteristics vary from task to task and person to person, even if the same score is achieved on different tests and therefore test scores cannot be regarded as the ultimate consideration of validity. "The same score may represent different abilities or different combinations of abilities, or different interactions between traits and contexts and it is currently impossible to say exactly what a score might mean" (Alderson & Banerjee, 2002:101).

Test score interpretations can only be valid if they are based on scores produced by an instrument that is trusted to measure the construct for the purpose of assessment. Van der Walt and Steyn (2007:141) note that the interpretative view of validation is widely accepted, but point out that it is "dependent on the test results being used for the purpose for which the test is designed". They acknowledge that test score interpretations must be valid and because various factors may influence interpretations, sufficient evidence is necessary in order to draw conclusions about the quality of a test, i.e. its validity. "It starts as a local affair, with repeated use of a test for one purpose only, and ultimately one can argue that validity becomes a property of the test, i.e. that it tests what it purports to test; that it tests a property that exists and can be measured" (Van der Walt & Steyn, 2007:141).

Weideman (2009:10) notes that attempts to avoid reference to the validity of a test often leads to roundabout references, such as Messick's (1980:1025) reference to a "test ... accomplishing its intended purpose". Weideman thus poses the question of why a test that does what it is supposed to do (i.e. accomplishes its purpose) would not be valid, since it produces the desired effect, viz. yielding the intended measurement. He equates a test's validity with its power or force, i.e. "its results could become the evidence or cause for certain desired (intended or purported) effects" (Weideman, 2009:10).

Despite much controversy and many scholars' criticism of various aspects of Messick's (1989) unified view of validity, the positive effects that the revised definition of validity has had should not be ignored. Firstly, Messick's views have inspired new research on test validity and test validation. Building on Messick's views, researchers such as Bachman (1990), Weir (2005) and Shaw and Weir (2007) have provided us with frameworks for validity, making the unified view more approachable, clearer and simpler.

Secondly, Moss (2007) advocates the value of a unitary concept of validity, arguing that it does in fact provide valuable guidance to investigating validity. New research inspired by Messick's unified view has laid some stepping stones for investigating validity:

> Various kinds of guidance provided include (a) a list of categories of types of evidence, inferences, or aspects of validity, often illustrated with the kinds of studies that might be undertaken ...; (b) principles to guide choices among the myriad kinds of evidence that are arguably relevant to a given interpretation or use ...; (c) standards or guidelines about the nature of evidence that should be made available to enable professional judgment ...; (d) outlines of "interpretative arguments" ... or comprehensive plans for validity research for particular types of interpretations and uses of tests ... accompanied by examples of the types of evidence that might be or have been developed under the plan ...; (e) descriptions and (critical) analyses of actual programs of validity research, associated with a particular test or construct ...; and (f) frameworks illustrated with extended examples ... that take us from conceptualization through test development and implementation (Moss, 2007:474).

Thirdly, Messick's theory provides a solution for the tense relationship between reliability and validity. The main emphasis of assessment has moved away from establishing reliability at the cost of validity. Instead, the construct under investigation is now the main emphasis of assessment and interpretation, along with the fair application of scores.

Fourthly, it is no longer enough to provide evidence for only one type of validity. Thus the opportunistic trend to provide only the most readily available evidence is eliminated. Test developers are also encouraged continually to investigate the validity of assessment instruments as they are used in different contexts and for different purposes. Validity is an issue of test design and development, but also of scoring procedures, grading, impact of the assessment, and of revision of the instrument after every administration. In response to the question: "How much evidence is enough?", Douglas (electronic interview with Alderson & Banerjee, 2002) comments that it is not possible to fully prove validity, but only to provide

ample and relevant evidence that will convince those using the test that it is valid for the assessment purpose and context (Alderson & Banerjee, 2002:103).

Fifthly, the unified model presents a flexible vocabulary for discussing validity issues. Gorin (2007) emphasizes the value of the unified model in providing such a flexible vocabulary, as it promotes cross-disciplinary discussions of validity. Although Gorin (2007) agrees with Lissitz and Samuelson that more emphasis should be placed on internal validity evidence, she opposes the suggestion that test development procedures and internal evaluation of content are enough to prove a test's validity. She argues that, unlike content, assessment constructs are not limited to one specific context. "Constructs exist across all assessment contexts" (Gorin, 2007:457). Gorin hereby also opposes arguments such as Lissitz and Samuelsen's (2007) for choosing "content" rather than "construct" to describe all processes underlying test scores, stating that "content" better serves the broader measurement community. "Thus, what Lissitz and Samuelsen present as a radical change in validity terminology is more appropriately characterized as an issue of semantics or perhaps terminological preference" (Gorin, 2007:457).

Finally, Messick (1989) regards the main issue in modern validity as the inconsistency between theory and practice. However, despite modern theoretical views that validity is a unitary concept, social and behavioural scientists continue to practise the traditional view of validity. Shepard (1993) as well as Hubley and Zumbo (1996) points out that many researchers do not investigate all aspects of validity and continue to discuss validity as a test characteristic and not in terms of inferences made based on test scores.

For the purpose of the present study, the importance of recognising social consequences of tests as advocated by Messick is accepted. Test developers do need to keep the probable use of an instrument (i.e. its purpose) in mind when designing it. Based on the construct being assessed and the reason why it is assessed, developers can select the content of a test to be representative of the construct. The uses of test scores are particularly relevant in the present study, as the focus is on producing an objective rating instrument for assessing ESL writing. The scores on an examination are used to make decisions that seriously impact on the futures of learners, be it academic or in the working environment. Therefore, this study supports concerns for the fairness and accuracy of test scores, the meaning attributed to the scores and

the decisions based on the interpretations. It aims to address these concerns by developing a valid measurement instrument for achieving accurate scores.

Furthermore, validity is not accepted as a property of test score interpretations alone. In order to meet requirements for valid test score interpretations, an instrument must provide an accurate measurement of the ability in question. Sound test score interpretations are based on the assumption that the instrument used provides accurate and reliable data (i.e. scores) and that the scores are used for the purpose for which they were intended. In this sense, the argument that the question of whether a test measures what it is supposed to measure is concerned with the meaning of test scores (Lissitz & Samuelsen, 2007) is supported. Sufficient evidence for validity needs to support the choice of a particular instrument for assessing a construct for a particular purpose. If this evidence is not available, test score interpretations cannot be assumed to be valid. As Weideman (2009:13) points out, test scores are technically qualified and theoretically grounded objects. They are meaningless without interpretation, but the subjective interpretations are made on the basis of an objective measurement. Thus, valid score interpretations can only be achieved if the instrument being used is believed to measure the ability for the particular purpose – i.e. is valid. If a test is used over a period of time for a particular purpose and is later believed to provide accurate measurement of a particular ability for that purpose, the test itself can be seen as valid (cf. Van der Walt & Steyn, 2007).

Various types of evidence are necessary in order to establish the validity of an instrument such as a rating scale. Scoring validity is a significant contributing factor to overall validity. In this regard, a validation framework is required to provide the basis for the development of a rating scale (cf. Chapter 4).

2.5 Conclusion

Validity concerns the accuracy with which a test measures a construct, portrays learners' abilities to perform a construct, and the fairness with which decisions are made based on the scores produced by an assessment instrument. This chapter has discussed the multi-faceted concept of *validity*, reflecting its complex nature. Traditional definitions of validity emphasise the test's ability to elicit the relevant construct, while modern definitions include an emphasis on fair test scores and inferences, as well as the consequences that these

inferences may have in a social environment. Validity is no longer regarded as only a characteristic of the test, but also of the test scores and the use of scores.

As this study is concerned with providing fair and objective scores for assessing second language writing in high stakes-situations, a unified model of validity is accepted in the sense that various types of evidence is necessary in order to present sufficient evidence for validity claims. Overall validity is regarded as a function of the interaction between various types of validities. Both context and purpose of assessment are recognised as factors that impact the validity of scores. Validity is a matter of degree. Reliability is no longer regarded as a separate test characteristic that exists in a tense relationship with validity. Rather, it is seen as one contributory aspect of validity, investigated under *scoring validity*.

Validity research involves a continuous process of collecting and revising evidence for or against the argument that a particular instrument is valid within a particular context and for a specific purpose. This argument is discussed in the following chapter, which focuses on the validation process.

CHAPTER 3

Validation Procedures

3.1 Introduction

The previous chapter defined validity and discussed the various different types of validity distinguished in traditional and modern terms. One measurement instrument can be used for different purposes, but the instrument's validity must be demonstrated for each assessment purpose and context it is used for. This can be done through a process of validation.

Validation remains an abstract concept (Kane, 1992:4). Historical overviews of the conception of validity – as presented in the previous chapter – show various difficulties regarding validity that arose through history. At present, practitioners' understanding of the concept is "disjointed and inadequate" and they are left without consistent guidelines for practical implementation (Schilling, 2004:178-179).

Validation generally entails collecting different types of validity evidence to support a claim for the validity of an instrument as measure for a particular assessment situation. Few published studies clearly explain the relationship between the development and validation procedures of assessment instruments such as rating scales. These studies do not always employ the same criteria or the same types of evidence for validity. Furthermore, evidence is often provided opportunistically with only readily available support presented (Turner, 2000; Alderson & Banerjee; 2002; Schilling, 2004).

This chapter discusses the process of validation. It defines validation, considers the need for validation of assessment instruments and examines a number of validation frameworks to guide the development of a validation argument.

3.2 Validation

Validity can only be accessed through the process of validation, which reveals what validity "looks like" for the particular instrument, purpose and context (Bachman, 1990; Alderson et al, 1996, Weir, 2005). Validation is the process of proving that an assessment instrument measures what it claims to measure and what developers intend it to measure, that the

instrument is relevant to the purpose and in the context of assessment, that the instrument produces scores that accurately reflect learner abilities, that the scores are interpreted accurately and inferences made fairly. According to Bachman (1990:243; 2004:265), validation is a study designed to produce a variety of empirical evidence about score interpretations and use that supports the validation argument.:

> [W]e must collect evidence supporting the construct validity of interpreting this score as an indicator of the individual's ability and consider the value implications of various labels we could attach to this interpretation of the particular theories of language upon which these labels are based, and of the educational or social ideologies in which these theories are embedded.

Investigating validity means questioning how logical a test is; that is, questioning the logic of the instrument's design, purpose and the empirical evidence it provides (Douglas, 2000:257-258). For each assessment instance the instrument must be validated for its particular purpose and context. Validity has to do with interpreting test scores in the light of the test purpose. Consequently, the question of validity should rather be whether the assessment is valid for that *specific purpose or situation,* instead of questioning the validity of the assessment as such (Douglas, 2000:257). Weir (2005:15) provides the following definition of test validation:

> Test validation is the process of generating evidence to support the well-foundedness of inferences concerning traits from test scores, i.e., essentially, testing should be concerned with evidence-based validity ... this necessarily involves providing data related to *context-based, theory based* and *criterion-based validities*, together with various reliabilities, or *'scoring validity'* (2005:1) ... [Validation is] "a form of evaluation where a variety of quantitative and qualitative methodologies ... are used to generate evidence to support inferences from test scores.

Validation thus entails collecting empirical data on an instrument and using them to make logical arguments to prove that inferences based on the test scores are appropriate for the purpose of the assessment and the particular population (Brualdi, 1999:1).

The validation process is continuous and re-iterative, beginning with "a construct in search of appropriate assessment instruments and procedures" (Lane, 1999:1). Validity evidence is collected from the onset of the design phase of an instrument and continues after the test has been administered. Considering the evolution of the concept of validity (cf. Chapter 2), any new issues of validity that arise must be investigated and considered in the revision of the

instrument (Lane, 1999; Douglas, 2000; Bachman, 2004). The need for revision and validation of assessment instruments does not become satisfied once an instrument has been proven valid for one administration.

Fulcher and Davidson (2007:26) ascribe this responsibility to the tester who needs to collect evidence supporting score interpretations and uses that is acceptable to stakeholders. Testers need to ensure that the tests they use have been validated for the particular purpose and context of assessment. However, validation is often noted as the least satisfactory aspect of language assessment development (Schilling, 2004:178). Thus, there is a need for validation in language assessment.

3.2.1 Need for validation

According to Weir (2005:15), "[m]ost examinations lay claim to the numerous aspects of validity. However, what are often lacking are *validation* studies of actual tests that demonstrate this". Despite the fact that many researchers (e.g. Hughes, Porter & Weir, 1988; Weigle, 2002; Weir, 2005; Bachman, 2005; Fulcher & Davidson, 2007) emphasise the importance of validation, many instruments that are currently in use lack sufficient evidence to prove the fairness of language assessment instruments. Studies that address the issue of validating existing scales include those of Messick (1992), Alderson and Buck (1993), Fulcher (1996), Lumley (2002), Weir (2005), Bachman (2005) and Fulcher and Davidson (2007).

Huot (1996:561) notes that the first important step to establishing a new theoretical umbrella for assessing writing is to develop assessment procedures on an epistemological basis that considers local standards, specifies both composing and reading contexts, and can be mutually interpreted. Validation procedures that consider the context of the assessment are necessary. Furthermore, qualitative procedures such as interviews, observations and thick descriptions are necessary in addition to quantitative validation procedures (such as correlations, cf. Chapter 2) to investigate the role of an assessment in a specific context (Huot, 1996:161-162).

South African assessment developers are confronted with the lack of validity evidence and face a demanding task of validating instruments such as the current Grade 12 Writing rating.

No empirical data are available to prove that the intuitively constructed writing scale is valid for the purpose of the final FET matriculation examination. Such an instrument may provide a skewed indication of learners' abilities in writing and may misinform stakeholders who use results to make decisions concerning the test takers. Thus, the current scale may have unfair social consequences. As pointed out in Chapter 1, the current writing scale needs to be evaluated and validated for the purpose and context in which it is currently employed. Validation is a demanding task and, as pointed out in Chapter 1, the current rating scale used for assessing writing in the matriculation examination has not been validated empirically. A framework is necessary for establishing a validation argument in order to achieve this task.

3.2.2 The need for guidance in validation procedures

Traditional approaches to validity are very open-ended and do not provide clear guidelines as to how much evidence is enough, which steps should be taken to collect the evidence, or where to begin and where to stop (Turner, 2000; Kane, 2004; Moss, 2007). Turner (2000:556) recognises this lack of guidance:

> [O]ne often wonders how scales are developed. With the important role that rating scales play in performance evaluation, one would think that the literature would abound with descriptions and procedures for scale construction. But, as we quickly learn, this is not the case.

Bachman (2005:1) notes that a set of principles and procedures has not yet been developed in the fields of language testing and educational and psychological measurement for linking test scores and inferences and score-based inferences to the uses of tests and their consequences. Messick (1989) discusses test use and consequences, but provides almost no guidance as to how these issues can be investigated practically during test development.

Research such as that of McNamara (1996), Saville (2000) and Taylor (2002) provide a general blueprint procedure for developing and revising assessment instruments. In order to ensure a valid instrument, its development must be based in language theory and be conducted according to an explicit development procedure. The basic process of developing or reviewing assessment instruments includes a design stage, a construction stage and a pilot stage (McNamara, 1996; Saville 2002; Taylor, 2002).

Figure 3.1 shows Taylor's (2002:2) graphic presentation of the basic outline of a development and validation procedure. It illustrates a cyclical and iterative process. Each of the steps is made up of a series of validation activities. Instead of collecting evidence in a "piecemeal fashion", evidence is evaluated continuously "as an integrated set to determine to which degree the validity argument is supported" (Lane, 1999:1-2; Saville, 2001:6). Developers must gain "adequate data" through transparent development and administering procedures to prove that sufficient standards are achieved.

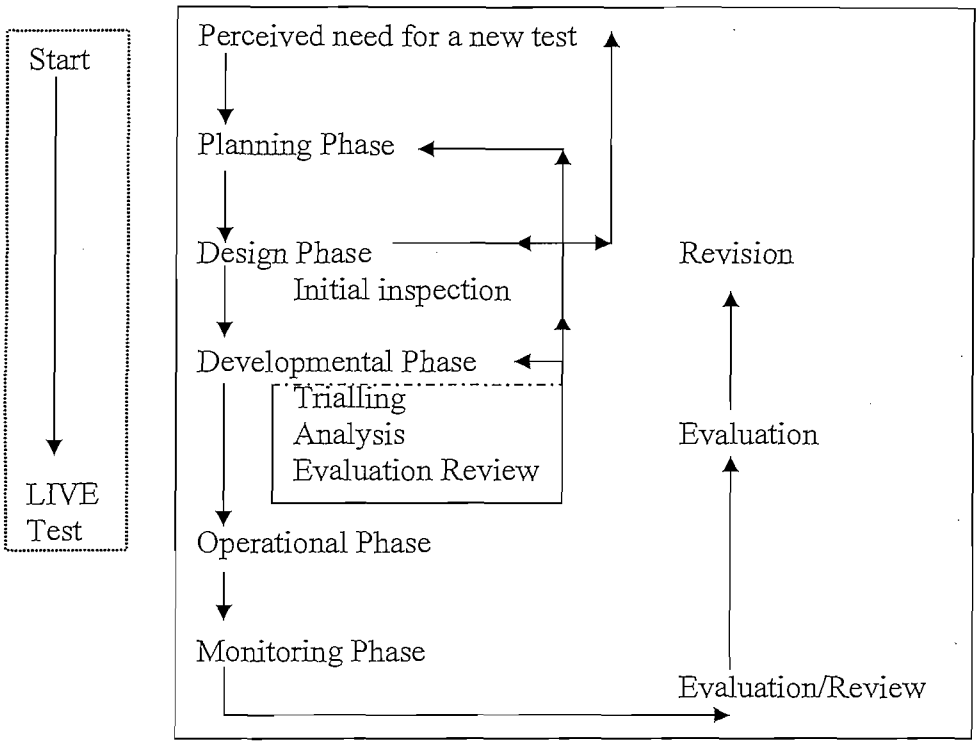

Figure 3.1 Model of the assessment instrument development process (Taylor, 2002:2)

Evidence from each stage offers new knowledge about the instrument, which is continuously worked back into its revision (Saville, 2001:5). However, evidence about the real qualities and workings (e.g. its appropriateness, accuracy and usefulness) of the instrument can only be collected once the instrument is implemented (McNamara, 2000; Saville, 2001).

The basic process illustrated in this figure does not specify the different validation activities to be conducted during each stage in the process. Kane (2004:1) notes that, although current validity theories are sophisticated, the methodology to achieve validity is generally ineffective. Stronger evidence is available for technical characteristics such as reliability of major testing programmes, than for their validity. Furthermore, attempts at validation often have a strong confirmationalist bias.

Moss (2007) points out that some guidelines are available through validation studies conducted by researchers building on Messick's unified model. A number of studies (Bachman, 1990; Hughes, 1996; Brown & Hudson, 2002; Scharton, 1996; Messick, 1996; Kane, 1992, 2004; Fulcher & Davidson, 2007) provide general guidelines for validation. Other validation studies, such as those conducted by North and Schneider (1998) and Shaw and Weir (2007), provide valuable guidelines for practical validation procedures.

However, more studies that provide validation evidence for existing scales are required (e.g. Alderson & Banerjee, 2002). In other words, there is a need for studies that specifically describe how tests and scales were developed and how the constructs were identified, operationalised, tested and revised. "Such accounts would represent a valuable contribution ... by helping researchers, not only test developers, to understand the constructs and the issues involved in their operationalisation" (Alderson & Banerjee, 2002:82). Bachman (2005:4) calls for a set of theoretically-grounded guidelines for investigating test use and consequences:

> We do not, at present, have a set of manageable procedures for investigating test use and consequences that is grounded on a coherent theoretical model of test use. Rather, we have simply several different checklists of questions and considerations with which test developers and test users should concern themselves.

Bachman (1990, 2005) as well as Toulmin (2003), Kane (1992, 2002, 2004) and Fulcher and Davidson (2007) advocates an argument-based approach to validation as it provides a systematic approach applicable to a range of score interpretations and uses. Messick (1996) and Kane (1992, 2004) distinguish explicit steps constituting an argument for validity. The following section examines the structure of a validation argument.

3.3 An argument-based approach to validation

The aim of validation is twofold: to reduce doubts around claims of assessment use and score interpretations, as well as to increase assurance that the intended use and interpretation of test scores are justified. This entails also investigating possible alternative uses and interpretations of scores, which implies coherent arguments for and against the proposed use and

interpretations. Evidence must be collected that investigate potential threats to the validity of test score interpretations.

Cronbach (1988:4) encourages researchers to think of a validation argument rather than validation research. The validity argument involves providing an overall analysis of all evidence in support of or against the suggested interpretation of test results (Cronbach 1980a, 1980b, 1988). Cronbach's suggestion of a validity argument is generally accepted today (Mislevy, 1996; Kane, 1992, 2002, 2004; Kane, Crooks & Cohen, 1999; Bachman, 2004; Fulcher & Davidson, 2007). Kane (1992, 2004) explains that the argument-based approach to validation serves to provide a methodology or technology for validation (Kane, 2004:2).

Validation as an argument entails stating a claim of validity and collecting evidence to support the claim that test scores have a particular meaning. Collecting evidence for validity, however, is more complex than merely checking items on a list. It involves hypothesising about the relationship between test scores and the meaning of the test scores and then collecting evidence to test the hypothesis (Lane, 1999; Haertel, 1999). The facts and data presented as evidence must be related to the claim and prove the hypothesis true or false, in other words that the suggested meaning of scores is acceptable. All suggested inferences in the validity argument are based on an assumption that must be supported (Lane, 1999:1; Fulcher & Davidson, 2007:160).

According to Bachman (1990:256), construct validity can be seen as a case of verifying or falsifying a scientific theory.

> [Validation is] the process of building a case – articulating an argument and collecting evidence – in support of a particular interpretation of test scores. Such an evidence-centred interpretive argument has a dual function. First, it provides guidelines for designing and developing an assessment instrument. The test construct and specifications provide the basis for developing test prompts, tasks and a scoring rubric. Second, it provides a framework for collecting the types of evidence necessary to support the intended interpretations and uses of scores (Bachman, 1990:263-264).

As pointed out in Chapter 2, traditional interpretations of validity lead to developers presenting only readily available evidence, as opposed to relevant information. Regarding construct validity as the basic overarching concept for the unified model of validity provides a solution to this problem. Kane (2004) stresses the need to operationalise construct validity

as unifying framework: "The basic principle of construct validity calling for the consideration of alternative interpretations offers some protection against opportunism, but like many validation guidelines, this principle has been honoured more in the breach than in the observance ... Most validation research is performed by developers of the test, creating a natural confirmationist bias" (Kane, 2004:140). According to Kane (2004:139), accepting construct validity as the unifying principle for validity establishes validation in the scientific tradition of clearly suggesting a theory or interpretation and challenging this interpretation empirically and conceptually.

Fulcher and Davidson (2007:18; 159) suggest that developers must continuously reconsider whether the validation claims being made and the supporting evidence are relevant and acceptable for the community of researchers and stakeholders. They view validation as a continuous, empirical process aiming to provide strong enough assurance in the form of evidence supporting claims that an instrument can be used for a particular purpose and scores be interpreted in a particular way. Fulcher and Davidson (2007:159) explain:

> The evidence may provide more confidence in a theory or interpretation, but new evidence may render the theory obsolete. We can therefore talk about validity as a process of argument for the meaning and use of test scores. This process ... requires us not only to seek evidence that supports intended or concurrent test use but also to investigate alternative interpretations of score meaning and test use ....

They propose the following features as key to a pragmatic view of validity as an argument (Fulcher & Davidson, 2007:21):

- An adequate argument to support the use of a test for a given purpose and the interpretation of scores is 'true' if it is acceptable to the community of language testers and stakeholders in open discussion, through a process of dialogue and disagreement;
- Disagreement is an essential part of the process of investigating alternative hypotheses and arguments that would count against an adequate argument;
- There are criteria for deciding which of many alternative arguments is likely to be the most adequate;
- The most convincing argument should start at the end point of considering the consequences of testing, and working backwards to test design.

The continuous process of collecting evidence for or against the validity argument extends from test design and development to scoring procedures, grading, interpretation of scores, decisions made based on the scores, and future revision of the testing instrument. As new issues of validity arise, these should be included in the investigation and the resetting of the assessment.

Collecting various types of validity evidence, with a focus on the most relevant types, provides the most secure support for a validity argument (Bachman, 1990; Hughes, 1996; Brown & Hudson, 2002). Scharton (1996:68) calls for triangulation designs, because it is easier to determine the degree to which the assessment is suitable for the intended purpose if validity evidence is demonstrated in more than one way. Evidence for different validities does not have to be equal in amount or strength, since an increase in one type of validity necessarily results in increased overall validity. Kane, Crooks and Cohen (1999) as well as Lane (1999) suggest focusing most on collecting evidence to support the weakest part of the argument. If the weakest part of the argument can be validated, it should not be too difficult to establish overall validity.

The notions of validity as an argument and validation as a process of compiling this argument offers advantages that other approaches to validity do not offer. Fulcher and Davidson (2007:163) highlight three such advantages. Test users are forced to recognise that every test is not suitable for use in every context and for any purpose. It focuses researchers' and developers' efforts on providing the most relevant types of validity evidence. It also guides them to decide which types of validity evidence are the most relevant for the purpose of testing. Finally, it serves to prevent test users from taking existing tests and simply "refitting" them for a similar purpose. In other words, test users are forced to develop new tests or at least provide evidence to prove that a test is valid for the purpose they mean to use it for.

As stated in Chapter 1, the present study aims to present a validation argument for the proposed rating scale.

3.4 The structure of a validation argument

Toulmin (2003:8) regards a sound argument and a well-grounded or firmly backed claim as "one which will stand up to criticism, one for which a case can be presented coming up to the standard required if it is to deserve a favourable verdict". In order to present such a sound

argument, developers have to be explicit about the claims being made and the evidence used to support it.

The main questions regarding the validation procedure concern how many procedures or steps are necessary, when these steps should be conducted and the nature of the evidence (i.e. qualitative or quantitative) to be collected at each stage. Kane (1992, 2002, 2004) and Toulmin (2003) describe the structure of a validation argument in terms of six stages. In the first three stages, a preliminary argument is phrased, which is subjected to criticism and rephrased and/or revised as necessary in the last three stages.

Kane (1992, 2002, 2004) refers to these two phases as the formative and summative phases. Both phases are anchored in the definition of the construct. In a seminal article, Kane (1992) introduces the notion of an interpretative argument to provide a framework for collecting evidence to evaluate the intended score interpretation. He suggests that an argument for validity comprises two sub-arguments or phases: an interpretative and a validation argument. The formative phase entails specifying the intended interpretation and use of scores in terms of an interpretative (proposed) argument that leads from observed performances to conclusions and decisions. Kane (1992:18-19) explains:

> Test-score interpretation always involves an argument, a chain or network of inferences, with the test score as the premise ... and the statements, predictions, decisions etc. involved in the interpretation as the conclusion. This argument is being referred to as the interpretive argument ... In interpreting test scores, the conclusions, including proposed actions, are typically stated explicitly.

The summative stage entails a critical evaluation of the plausibility of the proposed argument within a validity argument. Evidence for and against the proposed use and interpretation of scores are examined and the argument is revised based on the evidence collected.

The six steps that Kane (1992, 2004) identifies as comprising an argument of validity are: 1) Specify the intended score interpretation by stating a clear interpretive argument; 2) Evaluate the plausibility of the interpretative argument by examining the reasonableness of its assumptions and inferences; 3) Adjust the interpretative argument-based on the evidence; 4) Identify potential weaknesses in the argument; 5) Empirically examine the most problematic assumptions identified in step 4; 6) Evaluate the new argument that results from steps 4 and

5. These steps are not meant as a checklist, but serve to outline the argument-based approach in detail without being specific to the point of being restrictive. Furthermore, if any of these stages are omitted, the argument should justify such an omission (Messick, 1996:15-16). Kane (1992, 2004) and Messick (1996) provide details about what the main phases entail in terms of the six steps.

Fulcher and Davidson (2007:163) consider Kane's arguments as practical for investigating validity, showing how evidence is used to formulate the "most plausible and convincing argument that we can".

Toulmin (2003:36) describes a basic, field-invariant format of a validity argument, comprising of three basic components and three modifying components. Thus, Toulmin (2003:36) also distinguishes two main aspects, referring to the force of an argument and the criteria for evaluating the argument. The force refers to the probability related to the claim described by field invariant modal terms, while the criteria for evaluating the argument are field specific standards. The six basic components that the force and the criteria consist of are: claim, grounds, warrant, backing, modal terms and rebuttal. The claim is the conclusion of the argument we want to justify. Grounds refer to facts, evidence or data available and warrant refers to the justification of the claims based on the available grounds. Backing refers to any support required for the warrant. The modal term indicates how strong the warrant is while the rebuttal refers to the counter-claim, stating that the warrant does not justify the grounds as support for the claim. These roughly coincide with Kane's (1992, 2002, 2004) classification.

Haertel (2004:175) recognises the value of separating the two main components of a validation argument (Kane's interpretive and validation arguments; Toulmin's force and criteria), describing it as an "important conceptual advance". An explicitly stated interpretive argument serves three critical functions. Firstly, it guides developers because it points out the assumptions that scores need to meet.

Secondly, the conceptual framework of what scores are supposed to mean guides developers as to the types of evidence that might be collected to evaluate the suggested interpretations of the scores in consideration of the assessment purpose and context (AERA, APA & NCME, 1999:9; Kane, 2004:140).

Thirdly, the interpretative argument provides a base for evaluating the validity argument. Therefore clearly specifying the intended use and interpretation of scores prevents developers presenting only the most readily available evidence to support claims of validity, or only that evidence that is most likely to support the intended score meaning. Without an indication of what the suggested use and meaning of scores are, it is easy to adopt ambitious interpretations in discussion, but a less ambitious interpretation for practical validation purposes (Shepard, 1993; Kane, 1992, 2004). Bachman (1990:261) points out that the way in which the scores are interpreted, used and the consequences of the results are integral parts of the assessment that influence the interpretive argument.

Validity is explicitly associated with the credibility of the assumptions and inferences involved in the interpretation. The aim in the second phase (formative stage or application of criteria) is to present the rationale for the suggested interpretation and use of scores, along with the preliminary evidence that supports this suggestion. A preliminary case supporting the credibility, reasonableness and relevance of the intended interpretations of scores has to be made for the assessment purpose and context of language use.

This preliminary case is based on available evidence, as well as the relationship between the assessment procedures followed to generate scores and the intended interpretation. It must then be scrutinised and tested against specific criteria. The validity argument presents a case for believing the preliminary or interpretive argument regardless of how strong or weak it may be.

According to Bachman (1990:256-257), validation requires both a logical analysis and an empirical investigation to test the plausibility of interpretations and uses of scores. The construct must be outlined theoretically and operationally. The theoretical definition provides a base for counterhypotheses about the proposed relationship between constructs. Counterhypotheses play a fundamental role in the validation argument (Kane, 1992:1-2).

Thus, an investigation into validation aims to do two things: 1) reduce uncertainty or doubt around a claim of validity; and 2) provide assurance that a claim is justified (Fulcher & Davidson, 2007:160). As long as the reasoning for particular interpretations and uses of the

scores are reasonable and credible, the intended interpretation and meaning of scores are valid (Crooks, Kane & Cohen, 1996; Kane, 1992, 1994, 2004).

According to Kane (2004:151), the role of validators is to be constant critics of claims of validity. The interpretive argument, just like any other theory, is *never proven* and is always subject to new challenges and falsification (Bachman, 1990; Kane, 2004). Kane (1992:37) suggests that once the most problematic assumptions have been proven acceptable, the instrument can be classified as valid for the particular assessment situation. He explains that a clear and coherent interpretive argument, along with a validity argument that supports inferences and assumptions in the interpretive argument, make for adequately validated intended interpretations of scores (Kane, 2004:143).

Kane (1992, 2001, 2004) makes a valuable contribution to assessment instrument development and validation. He provides a commendable frame for addressing basic difficulties of validation in a step-by-step fashion (Schilling, 2004:178) and makes the "daunting enormity to the test validation enterprise" very clear (Briggs, 2004:174). Haertel (2004:175) states that Kane "advances the state of the art in using argument-based reasoning to guide evaluation of test interpretations". Briggs (2004:171) also notes the contribution that Kane's (1992, 2001, 2004) argument-based approach makes in terms of scientifically grounding validation. Assessment instruments should be developed in consideration of scientifically based research and, according to Briggs (2004:171), the suggestion that a preliminary argument is phrased and then subjected to criticism falls firmly within the principles that should underlie "all scientific inquiry in educational research", as identified by the National Research Council (NRC, 2002).

Bachman (2005:1) points out that although an argument-based formulation of validity such as Kane's provides a logical procedure for investigating and supporting claims, it does not consider the consequences of test use. Bachman (1990, 2005) emphasises the need to consider both validity and test use. Test developers are responsible for providing "as complete evidence as possible" that tests are valid indicators of the abilities in question, that these abilities are appropriate for the intended use, and to insist that the evidence be used to determine the use of the test (Bachman, 1990:285). Bachman (2005:7) notes that research on validity and validation tends to ignore issues of test use, while discussions of test use and consequences ignore validity.

Bachman (2005) builds his "assessment utilisation/use argument" on Toulmin's argument model. He extends the argument to include consideration of consequences of test use by combining Toulmin's argument structure with Messick's view that score meanings have to be both relevant and useful for the particular assessment context. Bachman (2005), following Messick (1989), argues that both the relevance and usefulness of scores and score interpretations must be sufficiently presented. Bachman (2005) explores the link between validity and test use. He suggests an assessment use argument structured as an overall logical framework for linking test performance to interpretations and interpretations to test use. The assessment use argument includes an assessment utilisation argument and an assessment validity argument. The first links an interpretation to a decision and contains claims, warrants, backing to support the intended use, and rebuttals of potential unintended consequences of test use. The second part consists of the validity argument which links a performance to an interpretation (Bachman, 2005:16).

According to Bachman (2005:14), formulating an argument for the validity of interpretations (as put forward by Kane, 1992, 2002, 2004 and Toulmin, 2003) and collecting evidence to support these interpretations are necessary but insufficient components for justifying the way we use language tests. It is not guaranteed that scores will be useful, relevant and sufficient for the intended purpose, even if the interpretations are valid. Furthermore, there is no guarantee that the scores and interpretations will not be used for purposes other than the intended purpose. Finally, the validity argument alone does not provide a basis for investigating potential negative consequences that may result from the way the scores are used. In other words, even if the score interpretations are valid, the results can still be used inappropriately. Thus, Bachman argues that sources of negative consequences that are "beyond invalidity" must also be described in the argument for the use of an assessment (Bachman, 2005:15-16).

Fulcher and Davidson (2007) describe how an argument of validation can be constructed on various levels, such as to argue that a particular item be included in an instrument, or that an instrument is valid for a particular purpose. They apply the generic validation argument structures of Kane's (1992, 2002, 2004) and Toulmin's (2003), following Bachman's (2005) extension of the argument to consequences of test use.

As an example of a test-level argument, Fulcher and Davidson (2007:169-170) consider the claim (a widely held view) that cloze tests are a valid measure of general language proficiency. They also consider the counter-arguments presented by Alderson (1983), Bachman and Palmer (1992) and Vollmer and Sang (1983). The argument is illustrated in Figure 3.2.

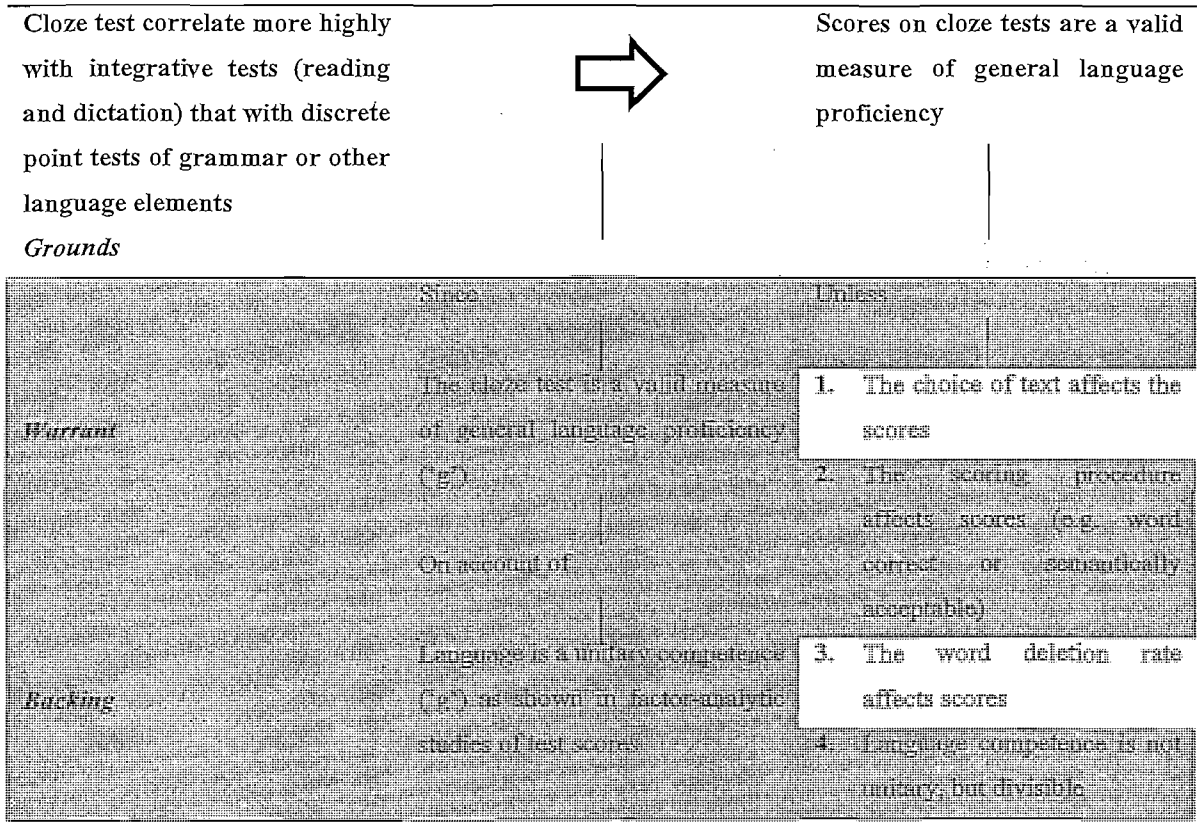

Figure 3.2 The structure of a validation argument as illustrated by Fulcher and Davidson (2007:169-170): The Cloze Argument

The claim that cloze test scores give a valid indication of learners' general language proficiency is supported by data (i.e. the data provide grounds for the claim). The data show that cloze tests correlate highly with integrative measures of language proficiency, but not with discrete-point tests of language elements such as grammar. These grounds invite the inference that a cloze test is a valid measure of general language proficiency. The warrant in this argument is that cloze tests are a valid integrative measure of general language proficiency. The claim is hereby linked to the grounds in a way that is empirically testable.

The rebuttal (questioning the warrant) in this case is presented in terms of four challenges. Three of these are directed at the warrant, namely that cloze scores can be influenced by

parameters other than language proficiency, such as the chosen text, the scoring method, and the word deletion rate. The fourth rebuttal, aimed at the backing, contends that the unitary competence is componential, rather than unified. The factor-analytic studies supporting the unitary competence hypothesis were flawed. In this example, the evidence in support of the rebuttal was strong enough to render the claim that cloze tests are a valid measure of general language proficiency invalid.

As established above (cf. 3.2), one aim of validation is to develop a scientifically sound argument in support of the intended interpretations of the scores, but also to investigate alternative meanings of scores (AERA et. al, 1999:9; Fulcher & Davidson, 2007:159). Based within a model of language ability, a validation framework is valuable in governing the empirical process of arguing for the meaning and use of a test. A number of validation models and frameworks are available. Bachman's (1990) model (adjusted by Bachman and Palmer (1996)) is discussed and the frameworks suggested by Cambridge ESOL, Weir (2005) and Shaw and Weir (2007) are reviewed below.

3.5 Validation frameworks

Valid language assessment is based in the theory of measurement and communicative competence. It is therefore essential that an assessment be based on a model of communicative language ability that can be supported empirically (O'Sullivan, 2006:95).

The terms "model" and "framework" are used in many different ways in discussions on language testing. Fulcher and Davidson (2007:36) distinguish between the two terms, defining a "model" as "over-arching and relatively abstract theoretical descriptions of what it means to be able to communicate in a second language", whereas a framework refers to "a selection of skills and abilities from a model that are relevant to a specific context". The purpose of a validation framework is to guide the establishment of a validation argument. A validation framework suggests the type of evidence that should be collected, as well as when and how it can be collected.

Bachman's (1990) model of communicative language ability – later adapted by Bachman and Palmer (1996) – provides the general mould for various validation frameworks, viz. the VRIP

framework used for developing and validating Cambridge ESOL instruments, Weir's (2005) socio-cognitive framework and Shaw and Weir's (2007) interactionist framework.

Bachman's (1990) views on validity have been highly influential. His chapter on validity in his 1990 publication, *Fundamental Considerations in Language Testing*, has been described as "the most influential mark of the 1990s" (Chapelle, 1999:257) and provides the "conceptual foundation" (Bachman, 1990:1) or intellectual basis for much current work in language testing. McNamara (2003:466) notes that "the publication of Bachman (1990) was a major event" and considers it a crucial text for various reasons, of which the most memorable was the introduction of Bachman's model of communicative language ability.

Bachman applies Messick's view of validity (discussed in Chapter 2) to language testing and describes validation of language assessment in the light of the unified concept of validity. Following Messick (1989), Bachman (1990) broadens the scope of validity to include consequences of assessment. The object of validation is not the test itself or the score itself, but the way the results are interpreted and the way the information gathered during the assessment procedure is used. He suggests that the value systems that justify suggested interpretations and uses of scores must be investigated. Validity concerns more than the reliability of scores; it also concerns "the relationship between test performance and other types of performance in other contexts" (Bachman, 1990:237).

Reliability is treated as separate from validity, but Bachman (1990:160) points out that the distinction between the two is not always clear-cut for language tests. Validity and reliability are seen as "complementary aspects of a common concern in measurement – identifying, estimating, and controlling the effects of factors that affect test scores".

In his model, Bachman (1990) takes the following aspects into account: the context which determines the use of test and scores, the nature of language abilities, and the nature of measurement. He attempts to describe the process through which the various components interact with each other and with the context in which communication takes place (Bachman, 1990:81).

Bachman's model works on a double level of abstraction (McNamara & Roever, 2006:31). First, the target domain being assessed remains an abstraction in the form of a construct (cf.

Messick, 1989). Secondly, a general model of communicative ability is advocated with developers drawing on the specified domain to outline a range of skills and knowledge. Conceptualising the demands of the target domain is critical in developing a validation framework.

According to Bachman (1990, 2005), validation of language assessment must be based on a detailed description of the abilities to be measured (what) and of the facets of test method (how), in consideration of the purpose and intended use of the assessment instrument. He characterises the domain of communicative language testing (which is the performance of communicative tasks in different contexts of use) in two stages. First, all contexts make demands on test takers' competence. He then lists the various aspects of competence in a model of communicative language ability, which serves as a menu to draw from to model the second stage, namely context of use. Bachman (1990, 2005) describes these two stages as the "what" and the "how" of language testing. He regards these as the main components that influence language performance in language assessment.

The "what" of language testing concerns the various attributes that test takers bring to an assessment, including communicative language ability. This is the construct that we intend to test. Language abilities are observed indirectly, and thus we have to hypothesise about how an ability influences language use and test performance. The extent to which we can make inferences about a hypothesised ability based on test performance is the fundamental issue in construct validity (Bachman, 1990:256).

Bachman (1990) describes the various aspects of test taker competence in a model of communicative language ability. He "reworks and clarifies" earlier models of communicative language ability (CLA), such as Hymes (1972) and Canale and Swain (1980) (McNamara & Roever, 2006:31). He recognises that the ability to use a language communicatively involves knowledge of a language, as well as the ability to implement or use this competence appropriately and contextually, in other words what Hymes (1972) referred to as 'communicative competence'.

In Canale and Swain's (1980) influential model of communicative competence, grammatical competence, sociolinguistic competence, discourse competence and strategic competence are distinguished as aspects. Bachman's interactional model, later extended by Bachman and

Palmer (1996), comprises three components of communicative language ability (CLA): language competence, strategic competence and psychological processes.

Language competence refers to a "set of specific knowledge components" used in communication through language. Bachman illustrates the main components of language competence (or "language knowledge", cf. Purpura, 2004:54) as in Figure 3.3.

This multi-componential model of communicative competence proposed by Bachman (1990) provides the "most comprehensive conceptualisation of language ability to date" (Purpura, 2004:54). Figure 3.3 illustrates the hierarchical relationship between the components of language competence.

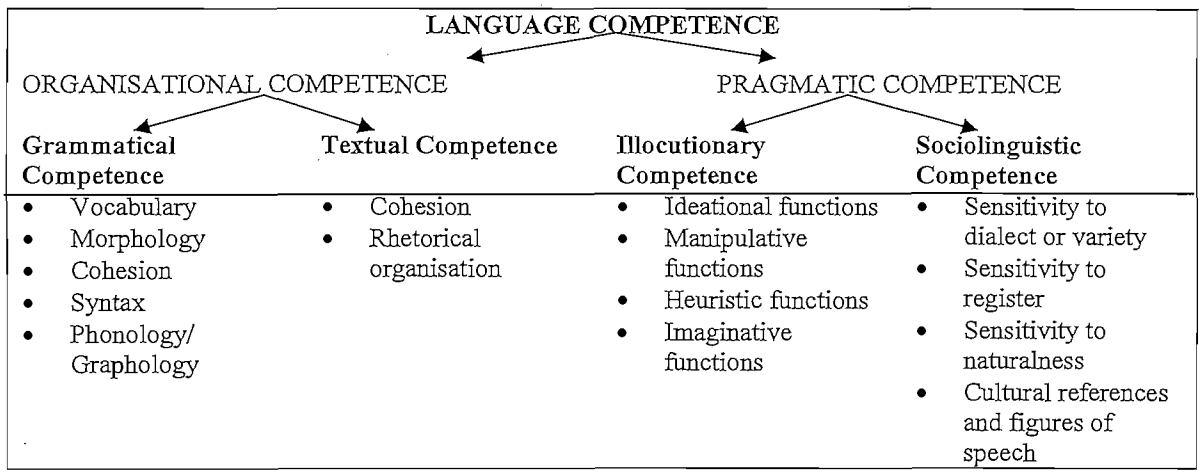

Figure 3.3 Components of language competence (Bachman, 1990:87)

Bachman (1990, 2005) clearly distinguishes between "knowledge" and "skills". Organisational competence refers to the way individuals control language structures to produce grammatical texts. Pragmatic competence refers to the way individuals communicate meaning and produce language that is appropriate in the context. The components on the third level (grammatical, textual, illocutionary; sociolinguistic competences) are divided even further. Grammatical competence concerns morphology, syntax, vocabulary and phonology. Textual competence relates to cohesion and organisation.

Illocutionary or functional competence refers to the ability to express and understand language with a particular illocutionary force. It includes various functions viz. ideational function, manipulative function, heuristic function and the imaginative function.

Sociolinguistic competence is connected with sensitivity towards different dialects or varieties of a language, to different registers, sensitivity to naturalness, and the ability to interpret cultural references and figures of speech (Bachman, 1990:84-94).

Strategic competence concerns the mental ability to use the components of language competence in appropriate and contextualised communication. It relates the language competencies to the performance context and the capacity to perform the ability appropriately. Strategic competence involves the assessment, planning and execution functions to achieve a communicative goal.

Assessment involves: 1) identifying information that is necessary to complete a particular communicative task in a certain context; 2) identifying the language competencies (e.g. first/ second/ foreign language) we have available to perform the given task; 3) determining which abilities and information we share with the interlocutor; and 4) evaluating the extent to which we achieved the communicative goal. The planning component retrieves the necessary components from language competence and constructs a plan to achieve the communicative goal. Finally, the execution component uses relevant psychological mechanisms to implement the plan appropriately for the context of the communicative goal.

Psychological processes refer to those processes – neurological and psychological – that are involved in the performance or execution of language as a physical entity (Bachman, 1990:84).

Purpura (2004:54-55) explains that language knowledge interacts with the language users' topical knowledge, other internal characteristics such as affect (e.g. anxiety; motivation), and with the communicative context to produce language. He defines language use as the internal interaction between the learner's attributes, as well as the external interactions between characteristics of the language use context. Both linguistic (language competence) and non-linguistic (strategic competence and psychological processes) are involved in communicative language ability. Thus, by recognising the three components of CLA – language competence, strategic competence and psychological processes – Bachman's model characterises the language abilities that are typically measured by language tests, the processes of interaction between the components, and the interaction between the abilities and the context in which abilities are assessed (Bachman, 1990:81).

The second main component, the "how" of language testing, refers to the methods used to elicit the necessary evidence as described by the tasks (Fulcher & Davidson, 2007:91), or the facets (characteristics) of test methods that influence performance on language tests. Thus, the "how" refers to the context and describes the different aspects of the assessment task and situation. Test method features are characteristics of the way in which language performance is elicited in an assessment. They are contextual features that determine the nature of the performance expected in a particular test situation (Bachman & Palmer, 1981; Shohamy, 1984; Turner, 2006).

Earlier models presented by Carroll (1968) and Clark (1972), for example, suggest that test tasks be distinguished according to the difference in stimulus and response characteristics. Based on these models, Bachman (1990) identifies five major categories of test method features that can potentially influence learners' performance on an assessment, viz. testing environment, testing rubric, nature of input available to test taker, nature of the expected response, relationship between the input and the response.

These five characteristics are not meant to be exhaustive and are only meant to guide empirical research. Turner (2000:557) also considers variables involved in the development stages of an instrument as method characteristics. The teachers and a sample of student performances involved during the development stages of, for example, a rating scale inherently affects the development of the instrument. The extent to which these variables influence the final instrument and the final score also needs to be considered.

Bachman (1990:155) notes the importance of describing the various aspects of test methods in detail, as this is an essential part of the validation process. Different assessment techniques vary along different dimensions. It is these dimensions that are of interest when investigating differences in test method facets. The method used to measure language ability has a major impact on performance, but some facets are potentially controllable and test developers should aim to do so (Bachman & Palmer, 1981; Shohamy, 1984; Turner, 2006).

Bachman's model has been criticised for its complexity as it has proven difficult to work with. Although the model provides valuable theoretical information, it has proven difficult to implement and operationalise in practice due to its complexity. After a "decade and more of

grappling with the operationalisation" of such a complex framework, the feasibility of validating inferences from scores has become debatable (McNamara & Roever, 2006:33).

Another point of criticism is that the concepts addressed in the model are not always clear. Davies and Elder (2005:804) describe Bachman's conception of validity as "intricate". In addition, Phakiti (2008:238) points out that the specific nature of strategic competence is unclear, despite a decade's worth of research into the powerful model. Bachman and Palmer (1996) only provide a preliminary discussion of strategic competence that does not clarify the nature of the concept (Purpura, 1999; Phakiti, 2008). A fuller discussion of the model with cross-reference to psychology, social psychology and pragmatics may aid a clearer understanding of the model and its components (McNamara, 1996; Phakiti, 2008). According to Purpura (2004:56), the Bachman and Palmer (1996) model discusses meaning to a degree, but needs to describe how grammar is used to encode meaning at different levels:

> It addresses meaning to some degree under the rubric of organisational knowledge (vocabulary), textual knowledge (cohesion), functional knowledge and sociolinguistic knowledge; however, given the central role of meaning in language instruction and communicative language use, a more explicit depiction of this aspect of language knowledge would be helpful.

Bachman (1990) suggests that the target language use context, and thus the content of the testing instrument, must be characterised in terms of the general model of language ability in order to make inferences about the ability based on the interaction of the ability with the test content. McNamara and Roever (2006:32) criticise this *a priori* approach to characterising the social context of language use. They note that this characterisation in terms of a model of individual ability "severely constrains the conceptualisation of the social dimension of language testing context". Although the model includes terminology referring to the social dimension of language competence, they consider the model to be mainly cognitive and psychological, without definite consideration of the social context; that is it is not "a theory of social context in its own right" (McNamara & Roever, 2006:32).

Bachman and Palmer (1996) adapted Bachman's (1990) model and introduced the concept of test usefulness to replace construct validity as superordinate term (Fulcher & Davidson, 2007:15). They place the focus on usefulness as "the essential basis for quality control"

(Bachman & Palmer, 1996:17). The model includes six test qualities: reliability, construct validity, authenticity, interactiveness, impact and practicality.

These qualities all contribute to the usefulness of an instrument and should be balanced and evaluated in combination with each other for each assessment situation and context. An instrument is only useful in some situations and its usefulness must be determined for those specific situations. Additionally, a balance must be established between the elements of test usefulness, the characteristics of target language use domain, the test-takers, and the construct definition for each individual administration (Bachman & Palmer, 1996).

Both the Bachman (1990) and Bachman and Palmer (1996) models distinguish grammatical, textual, functional and sociolinguistic components of language competence (language knowledge, Purpura, 2004:55), but neither provides a clear explanation of how these components interact during language use, or how grammatical knowledge may act as means for the interaction to occur. In addition, Purpura (2004:55) does not find Bachman and Palmer's (1996) description of grammatical knowledge very helpful in assessing form and meaning. Bachman and Palmer (1996) only describe grammatical knowledge as consisting of various components of linguistic form, which is strictly related to sentence based phonology, graphology, vocabulary and syntax. Bachman (1990) and Bachman and Palmer (1996) do not distinguish between the different types of meaning that grammatical forms encode.

Bachman (2005:6) notes that Bachman and Palmer (1996) provide a list of questions about potential construct validity and the impact of score interpretations. These questions are useful as means of quality control during instrument development and to investigate overall usefulness of instruments in use. However, Bachman and Palmer (1996) do not indicate how, if at all, the qualities of test usefulness are related to each other, or how construct validity and test use are directly related. Although Bachman and Palmer's (1996) notion of test usefulness suggests a different way of conceptualising validity and validation, it downgrades construct validity to an aspect of usefulness. According to Fulcher and Davidson (2007:15), this view has not challenged mainstream thinking after Messick, which may be the reason that the notion of usefulness has not been followed extensively in language testing literature.

Purpura (1999) and Phakiti (2008) propose that the Bachman (1990) and Bachman and Palmer (1996) models be validated, calling for empirical longitudinal data from high-stakes

tests to establish the validity of the frameworks. Phakiti (2008) attempts such a validation study.

McNamara and Roever (2006:33) suggest that Bachman's approach may be more valuable as "a powerful intellectual framework for considering validity in language test", that acts as a "conceptual mould" for other useful works on language assessment (Alderson, 2000; Douglas, 2000; Luoma, 2003; Purpura, 2004; Weigle, 2002; Bachman 2004). A validation framework may be more viable to guide validation in practice as it provides more explicit guidance regarding various aspects that need to be addressed for validation, and how they can be addressed, as well as the interaction between the aspects. The models proposed by Bachman (1990) and Bachman and Palmer (1996) have served as such a conceptual mould for various frameworks of language test development and validation, viz. those proposed by Cambridge ESOL (VRIP), Weir (2005) and Shaw and Weir (2007).

3.5.1 The Cambridge ESOL framework

Cambridge ESOL follows the so-called VRIP framework for validation, which is based on the models presented by Bachman (1990) and Bachman and Palmer (1996).

Based on the traditional concepts of construct validity, content validity and criterion-related validity, Cambridge ESOL proposes a framework for test development and validation based on four qualities of test usefulness: validity, reliability, impact and practicality (or the VRIP system). These four test qualities overlap considerably with the six qualities of test usefulness proposed by Bachman and Palmer (1996), namely reliability, construct validity, authenticity, interactiveness, impact and practicality (Hawkey, 2006:18).

The four elements are regarded as essential qualities for test usefulness (Saville, 2003:65; O'Sullivan, 2006:95). Following Bachman and Palmer (1996), the VRIP elements are regarded as central qualities in validation (Shaw, 2006:8) because an instrument's usefulness for a particular purpose and context depends on them (Jones, 2001:2). These elements must be considered equally to develop valid assessment instruments successfully (Bachman & Palmer, 1996; Saville, 2001; Shaw & Jordan, 2002; Shaw, 2006).

In order to achieve a balance between the VRIP elements, a measurement instrument must be appropriate for the assessment purpose, produce very similar scores over repeated assessment occasions, have a positive influence on the general education process and individual stakeholder, and be practical to develop, produce and administer (Saville, 2003; Hawkey, 2006). ESOL assessment instruments are developed through an iterative, cyclical process, based on "the need to establish the 'utility' of a test [or scale] in fulfilling its intended purpose in a useful way" (Shaw, 2006:8).

ESOL test development practices support Messick's unitary view of validity. Shaw and Jordan (2002:11) describe validity as the most important of the VRIP assessment qualities. The ESOL framework presents "fitness for purpose" as the dominant principle of validity. Different types of validity evidence contribute to different levels of support to the central issue of validity, with construct-validity placed at the core (Saville, 2003). The ways in which validity is established must be relevant to the construct being assessed (Shaw & Jordan, 2002:11). Thus, content-related validity must be collected in order to indicate the degree to which items in an assessment represent the defined domain and construct.

O'Sullivan (2006:96) points out that the relevance of content validation becomes apparent when considering the authenticity of an assessment. The relationship between test input and response is an important feature of content validity. The authenticity of the content and of the test taker's interaction with the content must be considered in order to achieve high validity (Weir, 2002; O'Sullivan, 2006).

Reliability as presented in the ESOL framework concerns the degree to which test results are stable, consistent and free from bias and random error (Saville, 2003:69). Following Bachman (1990), the Cambridge framework classifies reliability as a necessary aspect of test validity, but one that is insufficient in itself to ensure validity. It therefore highlights the tension between the two complementary aspects – an increase in one does not necessarily result in an increase in the other (Jones, 2001:2).

Impact refers to "the effect of the new or revised test on its stakeholders and the feedback that they provide to test developers" (Shaw, 2006:8). The educational impact of assessment in the context of use must be monitored and investigated from a validation perspective. Saville (2003:74) suggests that developers strive to achieve positive impact, but that they at least

ensure that no negative impact results from the assessment. All stakeholders should benefit from assessment in terms of information provided by the results. The information that scores provide about learners should, for example, guide teachers and developers in adjusting teaching and assessment materials.

Both *a priori* and *a posteriori* procedures can be employed to achieve a more positive impact. *A priori* procedures include developing test specifications and using experts to develop the instrument and examine performances. *A posteriori* procedures include collecting data about who takes the test, who uses the results and for what purpose (Saville, 2003; O'Sullivan, 2006).

Saville (2003:76) is of the opinion that the importance of practicality as part of test usefulness is often overlooked. He argues that it affects many aspects of an examination. Practicality is mainly concerned with administration and refers to how economical the test is regarding time and other resources, e.g. security, money, equipment, staff etc. (Harrison, 1983:12-13).

The ESOL framework emphasises practicality and presents it as an element that contributes directly to validity. Falvey and Shaw (2006:8) explain that the focus of practicality is on whether testing and scoring objectives can be reached conveniently, without major restrictions or logistical problems. According to Weigle (2002:56), practicality is a "key limitation" for writing assessment.

Balancing the VRIP elements in test development is necessary to establish fitness for purpose (Weir & Shaw, 2005:10). However, the VRIP framework does not provide specific guidelines for dealing with the elements adequately and completely to present a validation argument. Saville (2004:2) points out that: "in order to develop a 'Cambridge ESOL validity argument', our test development model needs to be underpinned by theories ... in order to combine the *test development* process with *necessary evidence*".

He further notes that the concept of validity in the VRIP framework is not clearly defined in terms of its constituent parts. O'Sullivan (2006:96) notes that it is "difficult, if not impossible" to clearly distinguish between the concepts of construct and content-related evidence of an assessment's validity.

The interaction between the VRIP elements is furthermore not described explicitly. The relationship between validity and reliability is described as problematic, but these two elements are still considered separately and as opposing elements. Although validity and reliability have traditionally been conceptualised in this way, such a categorisation is "confusing and often misleading" (Marcoulides, 2004:183). Marcoulides (2004:183) criticises the conception of validity and reliability as two separate entities despite the fact that tradition presents it as such. He argues that such a split view impairs an understanding of intricate relationship between validity elements and creates a false impression of the concept of validity and validation (Marcoulides, 2004:183):

> In fact, this inherently fragmentary approach to measurement procedures also fosters a lack of appreciation of the interrelationships and interdependencies among various aspects of measurement theory. As a result of this fragmentary approach, some researchers are even led to believe that "reliability and validity are independent of each other" (Anderson, 2003), whereas others appear to follow Rozeboom's (1966:375) assertion that reliability is "the poor man's validity".

O'Sullivan (2006:195) also argues that a model of test development that considers validity and reliability as two (contributing) aspects of the unitary concept of validity would be more useful than a model that considers the pair separately.

Although consequences of assessment are addressed in this framework, impact is presented as an additional element to be considered outside of or in addition to validity. O'Sullivan (2006:97) raises concern about the lack of consideration of how reliability decisions impact inferences:

> The fact that no practical consideration of how reliability decisions impact on a test can be made without also considering the implications that these decisions might have on the validity of the inferences we can draw from performances on that test means that there is a limit to the lengths to which it is possible to go in order to achieve maximum reliability.

He further argues that internal consistency based on item variance measures and internal consistency estimates (e.g. Cronbach's alpha) alone are not suitable or accurate measures of reliability for criterion-referenced tests, such as the ESOL examinations (O'Sullivan, 2006:97). Such measures may give a false impression of a reliable or unreliable assessment instrument by exaggerating evidence of reliability either positively or negatively.

The VRIP framework also places much emphasis on practicality, although ease of use does not inherently affect validity. A valid instrument that cannot be implemented due to practical limitations, or needs particular resources (e.g. staff, time or finances) in order to be implemented, may not be of immediate use, but it remains a valid instrument.

Although all four VRIP elements may contribute to the usefulness of an assessment, it may be difficult to achieve the required balance between the elements presented in the ESOL framework. Weir (2005) recognises the value of considering the VRIP elements, but provides a more practical validation framework.

3.5.2 Weir's (2005) socio-cognitive framework

In an attempt to renew the VRIP approach to validation used for the Cambridge Main Suite Examination, Weir and Shaw (2005:10) look towards Weir's (2005) evidence-based approach for an enhanced validation framework. "Weir provides a theoretical, socio-cognitive framework for an evidence-based validity approach which accommodates and strengthens the existing VRIP approach" (Weir & Shaw, 2005:10). His temporal framework provides a useful "map" that guides developers as to what should be happening and when it should be happening in the process of validation (Weir, 2005:43; Weir & Shaw, 2005:11).

Weir (2005:12) sets a demanding task: "To improve test fairness we need an agenda for reform, which sets out clearly the basic requirements for sound testing practice". Proving an instrument's validity requires "multifaceted and different types of scores on a test".

In order to achieve this task, Weir (2005) advocates a socio-cognitive framework for assessment validation. In language testing, latent mental constructs are tested. Language use is regarded as a social phenomenon rather than a purely linguistic act. Therefore, both cognitive and social aspects such as the context and audience should be recognised in assessing abilities such as writing. Weir's (2005:19) framework helps to identify processing and contextual elements and the relationship between them.

Learners' mental processing, i.e. the cognitive dimension, demonstrates abilities to be tested. Test developers must be aware of the theories concerning underlying cognitive processes

necessary to perform in real-life communication situations. As noted, the social dimension of communication is emphasised, with language presented as a social phenomenon, rather than purely linguistic. The context in which language is to be performed must therefore also be investigated empirically and described. Descriptions of the underlying processes and the conditions of language use should simulate real-life situations as closely as possible. Such descriptions demonstrate context validity (Weir, 2005:18-19).

> [L]anguage processing does not take place in a vacuum, so testers also need to specify the context in which this processing takes place. They need to provide empirically-based descriptions of the conditions under which these language operations are usually performed. Such descriptions of both operations and performance conditions should match target situation use as closely as possible, i.e., they should demonstrate *context validity* ... In short, a socio-cognitive theoretical model is required which helps identify the elements of both context and processing and the relationships between them.

Weir's (2005) framework resembles VRIP, as it is based on the Cambridge framework, but validity is reconfigured as a unitary concept and the framework illustrates the interaction of the different elements (Weir & Shaw, 2005:10-11). Weir and Shaw (2005:11) describe the framework as "ostensibly concerned with specifying and inter-relating focus areas for the validation process rather than with how the validation case should be argued *per se*".

Weir (2005) supports the modern description of validity as the extent to which a test provides data in the form of test scores to indicate accurately a learner's ability levels of language knowledge or skills. The key elements of Weir's (2005) socio-cognitive framework comprise the various forms of validity discussed in Chapter 2 – context validity, theory-based validity, scoring validity, consequential validity and criterion-related validity – within a unified model. He accepts construct validity as a superordinate category describing the various elements that constitute validity. The elements of validity and the various kinds of evidence generated by each element to support the interpretation of test scores are described in the framework (Weir, 2005:13-14).

In contrast to the ESOL framework which continues addressing validity and reliability in a trade-off relationship, Weir (2005) regards this traditional polarisation of reliability and validity as unhelpful (2005:14). He suggests that reliability should be considered as one form of validity evidence and refers to this type of validity as scoring validity (Weir, 2005:23-24).

Scoring validity is introduced as an umbrella term for all issues that concern developers during the stage in the process when language performances are translated into test scores. The concept of scoring validity seems more helpful than reliability (O'Sullivan, 2006:185). According to Bachman (1990:163-166), scores are affected by factors such as communicative language ability, test method facets, test takers' personal attributes and other random factors. If reliability is conceptualised in terms of these factors, O'Sullivan (2006:185) suggests that consideration of reliability would improve greatly.

The concept of *scoring validity* refers to reliability and other statistical attributes such as item analysis, internal consistency, error of measurement, marker reliability. Although relatable to the factors proposed by Bachman (1990), including elements such as the rating scale, rating process and the rater makes the situation more complex. The rating process is relatively unexplored and little or no evidence is available about how the aspects of scoring validity as presented by Weir (2005) impact rating performance. O'Sullivan (2006:186) urges further research into these aspects and the effects they have on validity. He notes that just like test taker characteristics influence performances, so will similar characteristics of raters affect their performance in scoring assessment performances. Different raters may assign different scores to the same performance (cf. Engelhard, 1994; Lumley, Lynch & McNamara, 1994).

Scoring validity as descriptive term also serves to ground reliability firmly within the validation process. According to Weir (2005:44) the main concern at the time of scoring is marking reliability and internal consistency matters. Therefore, the component of scoring validity as projected in Figure 3.4 (below) is the main focus of this study (cf. Chapter 5).

Unlike the ESOL framework, Weir (2005) does not consider practicality as a necessary condition for validity. Thus, practicality is not included in the framework. Weir (2005:49) argues that practicality should only become an issue after sufficient validity evidence is available to justify interpreting scores as indicators of individuals' underlying abilities. The degree to which an instrument is convenient to use should not be allowed to overrule the fairness and accuracy with which the construct is being measured. Concerns with practicality often intrude too early in the process, causing the validity of the instrument to be threatened rather than enhanced. "We should not consider method before trait" (Weir, 2005:49).

Weir (2005) regards construct validity as inextricably linked in a symbiotic relationship between context validity, theory-based validity and scoring validity. The interaction between the assessment context, trait and score directly contribute to, and attempt to support the construct validity of an assessment (Weir, 2005:21). The symbiotic relationship is evident in, for example, how decisions about task context parameters influence the cognitive processing that takes place during task performance. Also, if learners know the scoring criteria before an assessment, their executive processes in planning and completing a task are influenced (O'Sullivan, 2005:1). This symbiotic interaction between the elements in the framework is discussed in Chapter 4.

The socio-cognitive framework proposed by Weir (2005) is presented graphically in terms of the four macro-skills, viz. reading, writing, speaking and listening. The premise behind Weir's model (2005) is that developers should work to provide evidence of a test's validity from a variety of perspectives (O'Sullivan, 2005:1). For the purpose of the present study, only the framework concerned with writing is relevant and presented below in Figure 3.4.

The figure shows how the different parts of the validation process fit together chronologically and conceptually. It illustrates the socio-cognitive framework of writing assessment validation.

Weir (2005) explains the unified approach of this framework: "The arrows indicate the principal direction(s) of any hypothesised relationship: what has an effect on what. The timeline runs from top to bottom: before the test is finalised, then when it is administered, and finally what happens after the test event" (Weir, 2005:43). Weir and Shaw (2005:10) note: "Within each constituent part of the framework criteria individual parameters for distinguishing between adjacent proficiency levels are also identified."

Although this socio-cognitive framework resembles the Cambridge ESOL framework, Weir (2005) rearranges the constituent parts of validity to illustrate how the parts interact. The various kinds of evidence are equal in the sense that they are complementary aspects of an evidential basis for the test inferences.

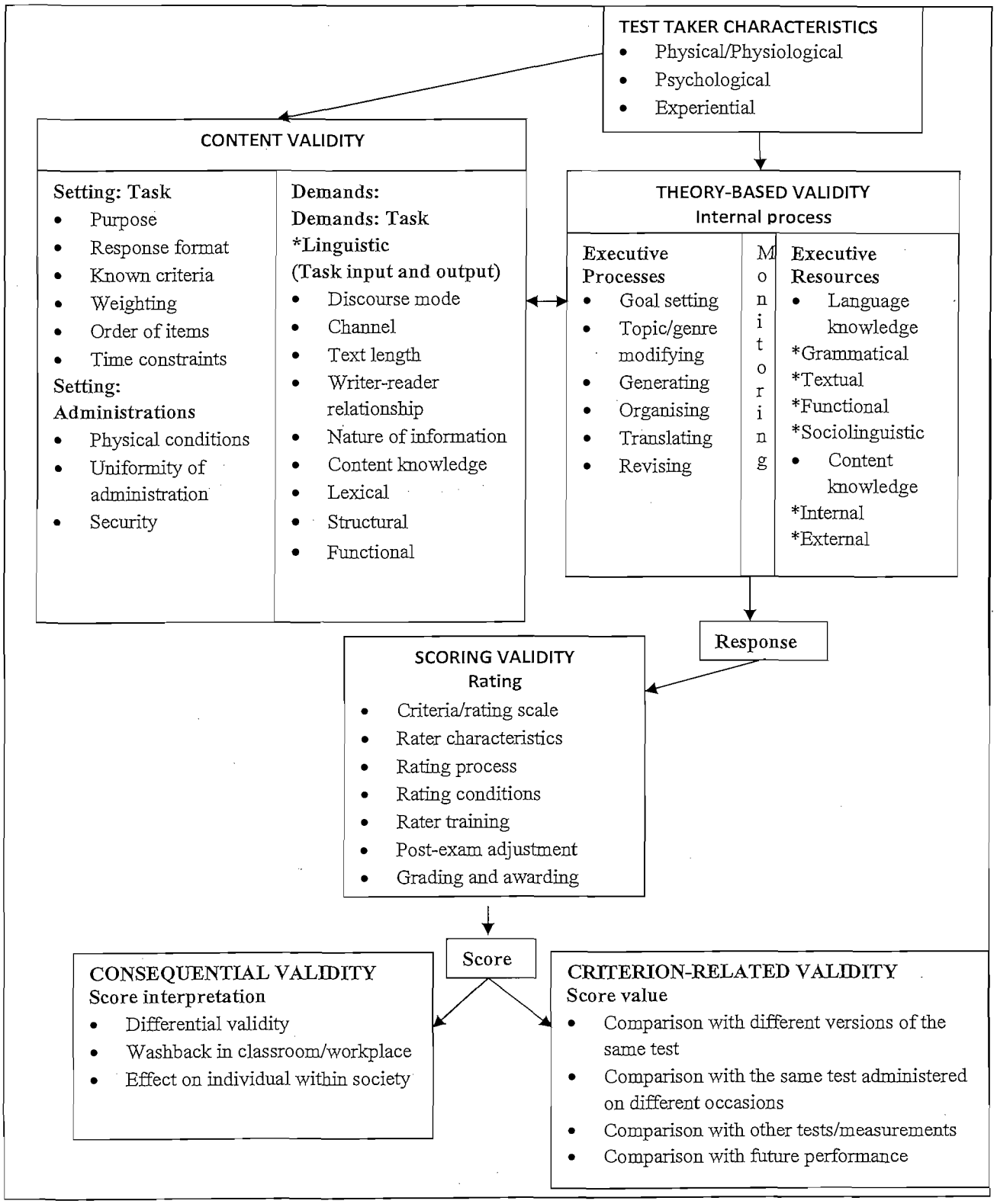

Figure 3.4 Socio-cognitive validation framework suggested by Weir (2005)

Both *a priori* and *a posteriori* components of validation are included in the framework. Construct validity cannot be illustrated only by means of *a posteriori* statistical validation to show that a test measures an underlying ability. Weir (2005:47) emphasises the need for *a priori* validation of the construct, stressing that we must specifically define the construct we intend to measure.

> The more comprehensive the approach to validation, the more evidence collected on each of the components of this framework, the more secure we can be in our claims for the validity of a test, The higher the stakes of the test the stricter the demands we might make in respect of all of these.

The socio-cognitive framework that Weir (2005) presents is clear, practical and follows an evidence-based approach to validation. According to O'Sullivan (2005:7), this comprehensive framework depicts validation as a step-by-step procedure in such a way that it serves as a practicum on validation methodology. Each element in the framework is discussed in detail with reference to when it should take place and the means of collecting the relevant forms of validity evidence in support of test score interpretations.

O'Sullivan (2005:7) notes that the framework is relatively easy to implement and has been implemented in Europe where it has been particularly influential. The framework has been applied to a range of examinations and context to describe existing examinations, to demonstrate the specificity of language for specific purposes and as basis for creating detailed test specifications (O'Sullivan, 2005:7).

However, despite its usefulness, O'Sullivan (2005:7) objects to the integration of consequential validity into the framework. He notes that it is not clear how consequential validity fits into an overall argument of validity. O'Sullivan (2005:7) suggests that consequential validity be regarded as a "more global aspect of test development which informs an ethical approach to all stages of development".

As mentioned above, Weir's (2005) framework addresses validation in terms of all four macro-skills. With the current study's focus on validation of specifically writing assessment, the framework proposed by Shaw and Weir (2007) is considered next.

3.5.3 Shaw and Weir's (2007) interactionist framework

Shaw and Weir (2007) adapt Weir's (2005) framework with a specific focus on validating writing assessment. Their interactionist framework reconciles the traditional "trait-based" and "task-based" approaches (Shaw & Weir, 2007:3).

The framework is based on a three-dimensional model of language proficiency. Shaw and Weir (2007) recognise the two dimensions of a construct as identified by Weir (2005), viz. the underlying cognitive ability and the characteristics of the context of use (task and situation in the test event). In addition, Shaw and Weir's (2007) identify a third dimension, namely the scoring dimension. The first two dimensions form core components of the construct definition. "The framework reminds us that language use – and also language assessment – is both a socially situated and cognitively processed phenomenon" (Shaw & Weir, 2007:xi).

Together with the scoring dimension, the three dimensions form the critical components of any language testing activity. According to Milanovic and Weir (2007:xi), these three internal dimensions of a language test constitute an "innovative conceptualisation" of construct validity. Theoretical, logical and empirical evidence in support of validity claims can be achieved by strongly focussing on these three elements and by analysing the test in relation to these elements.

The framework also accommodates and strengthens Cambridge ESOL's existing VRIP approach. It aims to establish similar evidence and additionally aims to reconfigure validity to show how the constituent elements interact with each other (Shaw & Weir, 2007:xii).

Following Weir (2005), validation is conceptualised in a temporal frame, including *a priori* and *a posteriori* elements of validation. Figure 3.5 illustrates the constituent elements of Shaw and Weir's (2007:4) writing validation framework.

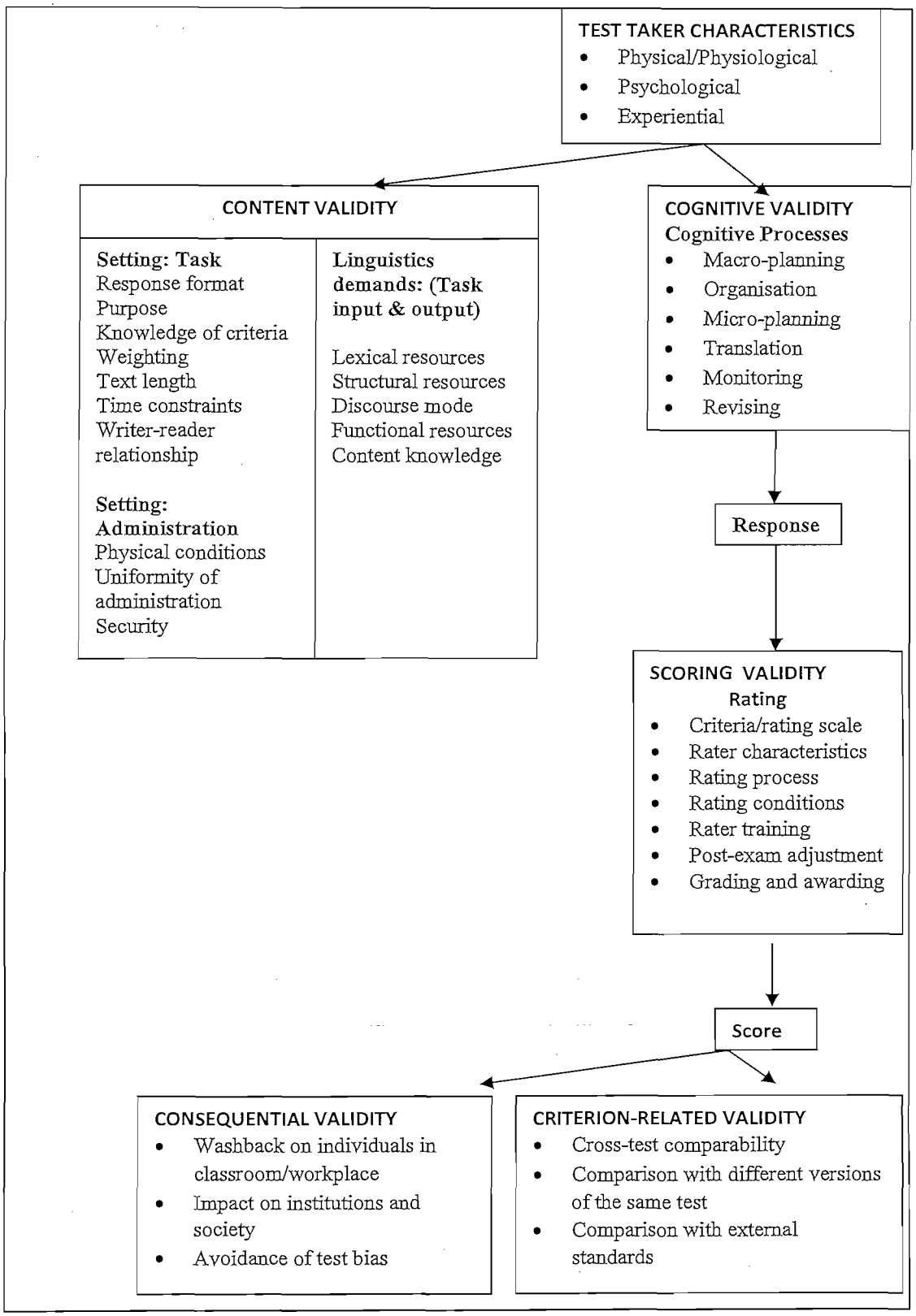

Figure 3.5 Framework for writing validation proposed by Shaw and Weir (2007)

Although the framework strongly resembles Weir's (2005) framework on which it is based, there are some changes. Weir's (2005) theory based validity is replaced by cognitive validity. Cognitive validity addresses the specific types of cognitive processes learners use to respond to writing tasks. Shaw and Weir (2007) also rephrase the factors of linguistic demand under context validity and adjust factors constituting scoring validity, consequential validity and criterion-related validity are adjusted and streamlined.

Following Weir (2005), construct validity is regarded as the product of the symbiotic relationship between context, cognitive and scoring validity. Various aspects of reliability are considered under the superordinate concept of scoring validity – one aspect of validity. Particularly in writing, scoring criteria describe the level of performance that is required and is therefore an important addition to context and processing to form the construct (Hawkey & Barker, 2004; Shaw & Weir, 2007).

Theoretically and practically, this socio-cognitive framework clarifies the constituent parts (test-taker characteristics, context validity, cognitive validity, scoring validity, criterion-related validity and consequential validity) of testing in relation to validity). Shaw and Weir (2007:4-5) pose a number of critical questions that are addressed when applying the validation framework:

- How does the test cater for the physical/physiological, psychological and experiential characteristics of candidates? (test taker)
- Do the test tasks elicit the appropriate cognitive processes? (cognitive validity)
- Are the test task characteristics and their administration appropriate and fair to the candidates who are performing the tasks? (context validity)
- To what extent can we depend on the scores that result from the test? (scoring validity)
- What are the effects of the test and scores on different stakeholders? (consequential validity)
- What external evidence is there other than the test scores are there to test the fairness of the test? (criterion-related validity).

Answering the questions listed above can help test developers to provide sufficient evidence to support the claim that test scores accurately represent the underlying trait being assessed.

Shaw and Weir (2007) offer a comprehensive and useful framework, but they do not claim to provide answers to all key questions of validation (Rimmer, 2008:2). In evaluating this framework, Rimmer (2008) finds that specific guidance about how to address the minimum requirements concerning vocabulary at various levels is still lacking. Shaw and Weir (2007:104) comment that currently available quantitative measures of vocabulary do not provide adequate descriptions of test-takers' output.

Rimmer (2008) also comments that Shaw and Weir's framework does not consider important aspects such as the language processing dimension, and ways that learners manipulate language resources in real situations: "[I]t is surprising that cognition is identified primarily with strategy use" (Rimmer, 2008:2). Cognitive processes, such as the translation process, are mostly discussed in behavioural terms. In addition, the roles of function and form are not considered explicitly. According to Rimmer (2008:2), a cognitive framework should offer a more detailed discussion of language function and form.

Rimmer recognises the need for elaborating cognitive validity through further research into the processing of specific language items during writing tests. For example, he refers to the role of formulaic phrases in language assessment, noting that the more predictable language is the more language may be prefabricated (Rimmer, 2008:2; cf. also Lewis, 1997:41). While everyday language use is lexically driven, grammatical range and complexity distinguishes creative language use. This has certain implications for task type and assessment criteria, since formulaic language is used differently at different levels of ability.

The present study adopts Shaw and Weir's (2007) framework for validating writing as a theoretical base for establishing a valid rating scale to assess writing performances in South Africa. The framework is comprehensive, allows for theoretically questioning of the elements of validity and provides guidance for practical implementation. Shaw and Weir's (2007) demonstrate application of the framework to writing assessment with reference to the validation of the Cambridge main suite examination. This application serves as a guiding example of the temporal validation process and is useful for other validation studies.

3.6 Conclusion

This chapter has discussed the validation process. Validity can only be accessed through a process of validation. Validation is a process of building an argument. It entails continuously collecting different types of validity evidence from different sources to support claims that the intended score interpretations and uses are fair and appropriate for the purpose and context of assessment.

Assessment instruments need to be validated for different purposes and situations, but many test instruments lack a sound validity argument. The current rating scale used to assess the FET Writing examination in South Africa is one such an instrument.

A theoretical framework is necessary to develop a scientifically sound validation argument that supports intended score interpretations. Based on Bachman's (1990) model of communicative competence, Shaw and Weir's (2007) socio-cognitive framework for validating writing assessment is adopted for the purpose of the present study. The following chapter examines the different aspects in the framework and the interaction between the aspects as proposed by Shaw and Weir.

CHAPTER 4

A Framework for Validating Writing Assessment

4.1 Introduction

A theoretical framework guides practical assessment instrument development procedures. Whereas a model (such as Bachman's (1990)) depicts a representation of a construct (cf. Chapter 3), a framework (such as Shaw and Weir's (2007)) focuses on salient components within the given context of assessment (Chalhoub-Deville, 2003:379). It provides information about the various aspects to be considered in order to put together a sound validation argument and serves as a map for evaluating existing instruments. In Chapter 3, Shaw and Weir's (2007) socio-cognitive framework was identified as the most suitable for the purpose of the present study because it is specifically designed for validating writing assessment instruments and can be implemented in practice.

This chapter examines the parameters addressed in Shaw and Weir's (2007) framework, viz. test taker characteristics, cognitive validity, context validity, scoring validity, criteria-related validity and consequential validity. Each parameter is discussed and then applied to the context of the National Senior Certificate (NSC) examination written in South Africa where relevant. The current scale used for this purpose is evaluated in terms of its adherence to requirements stipulated in the National Curriculum Statement (NCS, 2005), as well as guidelines provided in the Learning Programme Guidelines for Languages (LPG, 2008) and the Department of Education's (DoE) Examination Guidelines (2009).

4.2 Test taker characteristics

A community's communicative language use is tied to their position in time, space, social and historical relations and their social and emotional identity. Communicators are therefore central to the communicative process (Kramsch, 2005:249). Test takers are the central component of the assessment event and are therefore a crucial factor to consider when developing assessment instruments (Shaw & Weir, 2007:17).

If test takers are systematically defined, test developers can consider them comprehensively when designing tests. According to O'Sullivan (2000:82-83), the content of an instrument constructed with the test takers' characteristics in mind is less likely to be biased against any group. O'Sullivan (2005:3) notes, however, that test developers tend to develop assessments based on their own perceptions of the test population instead of evidence about the test takers.

Various factors such as age, interest, experience, knowledge and motivation affect test takers' performances. Ellis (1994) differentiates between social contexts and social factors that affect test takers in different ways. Social context refers to different settings in which formal or informal learning of a second language can take place. Social factors include age, gender, social class and ethnic identity. Different settings present different contexts in which combinations of social factors influence learning outcomes and learner performances (Ellis, 1994:197). Bachman (1990:146) distinguishes between systematic and unsystematic characteristics. Systematic characteristics always influence learners' performances in the same way and can be controlled to some degree. They include, for example, learners' content knowledge, cognitive style, physical disabilities, age and gender. Unsystematic characteristics are random, unpredictable and mostly temporary influences such as personal and emotional circumstances that cannot be controlled by test developers and administrators.

O'Sullivan (2000, 2006) distinguishes three categories of test taker characteristics, viz. physical/physiological, psychological and experiential characteristics. Information about physical characteristics allows validation officers to conduct bias analysis studies to prevent bias against a particular gender, age group or cultural/social/language background and accommodate learners with special needs. Physical/physiological characteristics include short-term illnesses, such as headaches or colds, speaking, hearing, and vision disabilities, as well as age and gender. Psychological characteristics are those related to test takers' personalities, memory, cognitive style, concentration, motivation and emotional state. According to O'Sullivan (2006:9), these characteristics are issues to be considered for research in relation to the test design, rather than for consideration of validating a test instrument. Experiential characteristics relate to factors such as learners' former education, examination preparedness and experience, and their experience communicating in a first language environment. Shaw and Weir (2007) follow O'Sullivan's distinction and address physical/physiological, psychological and experiential characteristics in their framework.

It is very difficult to cater for all the specific needs of individuals and adhere to the requirements of fair assessment with such a wide range of variables that could influence test takers. This is particularly the case in large-scale assessment situations with heterogeneous groups comprising individuals from multi-ethnic backgrounds, a wide age range, and speakers of different varieties of English. Therefore, Shaw and Weir (2007:19) advise that, at best, no learner should be disadvantaged by the socio-cultural content of an assessment.

In order to achieve this, Shaw and Weir (2007:19) suggest that tasks should reflect real-life communication situations. They also recommend using neutral topics and trialled materials, discarding any inaccessible or inappropriate content, collecting feedback from test takers about the accessibility of topics, and considering possible influences from learners' native languages during training. Weir (2005:54) urges that "every attempt should be made" to make sure test takers are familiar with the task types, demands of the assessment and environmental features to be presented during assessment. Class exercises with the same format as that in the examination, and specimen papers can be used to familiarise learners with the content and response format. If learners are used to the type of tasks, negative effects of previous assessment experiences may be overcome.

In 2008, 592 000 Grade 12 learners were the first to write the new FET examination, which included the re-introduced Writing Paper (Paper 3) (cf. Chapter 1). This is a high-stakes, large-scale examination. Learners who write the Department of Education's National Senior Certificate examination at the end of Grade 12 all write the same paper. Results from this examination are widely used as indication of learners' abilities.

In order to ensure that all learners are familiar with the format and typical content of the FET examination, previous and exemplar examination papers are available. In addition, the outline and details of the September examination for Grade 12 learners are the same as that of the final examination at the end of the year (NCS, 2005:22). Learners should therefore know what they can expect in the examination.

Additional measures are taken to accommodate physically disabled learners when writing the examination through, for example, Braille papers, verbal explanations and oral examinations, or using a 'buddy-system' (Language Programme Guidelines, 2008:32-33). The arrangements

made depend on the disability of the learner, but regardless of how these learners are accommodated, they should not have an unfair advantage over the normal test takers.

However, various factors that are more difficult to control influence learners' performances. The Grade 12 group is a very heterogeneous and multi-ethnic one. Learners have very different educational and L1 backgrounds (South Africa recognises eleven official languages) and display different skill levels. They also have diverse socio-economic and cultural backgrounds. Learners come from different life histories, diverse learning circumstances and have access to varying resources and capacities. McCabe (2007:26) notes the high proportion of over-age learners who are generally much older than the average learner.

Backgrounds range from the previous "white" model C schools to those from rural and very poor communities. Proficiency levels range from highly proficient to virtually illiterate. Many Grade 12 learners display poor levels of proficiency in English First Additional Language (FAL). Learners display "chequered patterns" of academic achievement with individual learners showing "enormous discrepancies between very good and very bad marks ... or have overall poor results" (Blommaert, Muyllaert, Huysmans & Dyers, 2005:8).

The causes of poor performance in English at Grade 12 level are complex. Learner performances are influenced by external factors such as poverty, malnutrition, and absenteeism (Van der Westhuizen, Mentz, Mosoge, Niewoudt, Steyn, Legotlo, Maaga & Sebego, 2002:113; Mafisa, 2008:8). Grade 12 learners' low levels of English proficiency are also partly attributable to factors such as shortages of resources and qualified teachers, poor infrastructure, disadvantaged backgrounds and poverty (Güles, 2005; Tailor & Prinsloo, 2005; McCabe, 2007).

English Second Language speakers in South African speak different varieties of English, including Black South African English (BSAE), Afrikaans South African English (ASAE) and Cape South African English (CSAE). Many non-native English speaking learners attend schools where English is the medium of instruction. English is also the dominant second language taught at primary and secondary school levels and the dominant medium of instruction in most higher education institutions (Dyers, 2008:112). However, even in English-medium schools, English is often spoken and taught as an additional language (Bekker, 2004; Blommaert et al., 2005; Osman, Cockcroft & Kajee, 2008). Most South African learners are second or third language speakers of English and many teachers are not fully proficient in English either. Teachers' poor

language use and approaches to teaching may be a major contributing factor to poor matriculation results in English (Güles, 2005).

Furthermore, many learners have little exposure to English outside the classroom. In many rural and township communities, English is mostly used to communicate with the teacher in the English class. In these communities learners often do not have English books to read at home either (McCabe, 2007:17; Dyers, 2008:119).

Kapp (2004:247) reports that South African learners' attitudes to the implications of learning English vary. On the one hand, many South African teachers and learners regard English as the most important language to acquire in order for them to achieve higher social and economic goals. Research (e.g. Dyers, 2004, 2008; Blommaert et al., 2005) indicates that secondary school learners believe acquiring English will enable them to rise from poverty and guarantee them better employment and upward social mobility. English is highly valued as "an index of spatial-social mobility" (Dyers, 2008:119; 123) and many regard it as "the language of liberation", adding a politically emotive connotation (McCabe, 2007:23). This attitude seems entrenched and unlikely to change.

On the other hand, many learners are loyal to their first language and their culture: "some ... regarded English as playing a part in creating social distance between members of the same family. They felt that people who were better off largely spoke English, and did not live in the township ... English was largely for use with outsiders" (Dyers, 2008:123). Speaking English removes them from their native background. This attitude influences learners' motivation to become proficient in English. They may experience English examinations as negative and perhaps marginalising.

Many South African ESL learners come from a history of educational deprivation, leaving them educationally unprepared (Moore, 1996; Osman et al. 2008). Mafisa (2008:7) notes that socio-economic background significantly influences learners' performances. Many learners come from low socio-economic backgrounds and their parents may not be well-educated. There is often a shortage of qualified English teachers in poor areas. Many of the poorer communities also struggle with a high rate of absenteeism for both learners and teachers. According to Taylor and Vinjevold (1999:228), the most debilitating malfunction in the educational system is widespread

absenteeism of teachers and teachers refusing to work after school hours. They use up to two months of teaching time for other activities, such as preparing and marking examination papers.

Classes are generally large, containing up to 50 or more learners per class. In schools with poor infrastructure, such large classes often cause a lack of motivation, class discipline and attention during classes and poor academic performance.

Kapp (2001), Blommaert et al., (2005), Barkhuizen (2005), McCabe (2007) and Mafisa (2008) all describe the dire learning and living circumstances of many South African learners. Their accounts show that some schools offer a stable learning environment with adequate resources, good discipline and supporting staff. Other schools, however, have poor infrastructure, inadequate resources (e.g. not enough textbooks for all), dilapidated buildings, and environments characterised by violence, gangsterism and vandalism.

Poor environmental circumstances such as these may affect learners' confidence and self-esteem levels and, in turn, their test performances negatively. Osman et al. (2008:7) report that many learners are not confident in their own abilities to write well in English. Learners find it stressful to generate and communicate knowledge in writing, and fear being evaluated on the product. "To a great extent, ESL students appear to be working in a subtractive environment, which impacts on their self esteem and self worth" (Osman et al., 2008:7).

4.3 Cognitive validity

Cognitive validity in a test of writing refers to the degree to which test tasks activate the same cognitive processes than writing in a real-life context. It is therefore important to select tasks that simulate real-life writing tasks in order to ensure that the relevant cognitive processes are activated. Cognitive processes involved in writing are influenced by factors such as the text type, purpose and the function of writing. Test takers use resources such as their content knowledge, which may be existing background knowledge or provided by the task input (cf. 4.4.1.1; 4.4.2.5) when responding to a task. These factors determine how learners approach, plan and execute a task.

In order to choose tasks that will activate the relevant processes, the construct that is being assessed must be well defined. The construct definition serves as starting point for all decisions

made about the assessment. A detailed definition indicates the criteria that are important for assessing the construct, in other words, what constitutes skilled writing in the particular context for the relevant purpose (Shaw & Weir, 2007:43). Paper 3 of the FET examination is concerned with assessing communicative writing in different contexts. All decisions regarding the content of the scale should therefore be made to produce an instrument that assesses this construct most effectively. (Chapter 7 discusses the procedure followed in this study to identify the features of the construct evident in learner writing at Grade 12 level.)

According to Field (2004:331-332), skilled writers spend more time on planning and monitoring than less skilled writers. In second language writing, additional cognitive demands are made on learners in that they have to pay attention to different aspects such as spelling, syntax and lexical retrieval.

Shaw and Weir (2007:34) point out that it is difficult to demonstrate cognitive validity. In order to establish it in writing assessment, the executive resources and cognitive processes activated by the task must be evaluated (Weir, 2005:110). The executive process is similar to strategic competence identified in Bachman's (1990) CLA model of language ability. The cognitive components identified in Bachman's (1990) model (discussed in Chapter 3) are subsumed within the cognitive processes involved in the cognitive validity parameter. Strategic competence or metacognitive strategies are specifically relevant here. According to Bachman and Palmer (1996:70), strategic competence refers to a "set of metacognitive components or strategies, which can be thought of as higher order executive processes that provide a cognitive management function in language use, as well as other cognitive activities". Various models have been proposed to demonstrate cognitive processing involved in writing, such as those of Hayes and Flower (1980), Bereiter and Scardamalia (1987), Grabe and Kaplan (1996) and Field (2004).

Hayes and Flower (1980, 2002) suggest that the writing process is influenced by the task environment and the writer's long-term memory. They represent writing as a non-linear, non-sequential, but recursive process. The writing-as-a-process approach entails goal-directed writing and interactive composition processes. The process of writing involves extensive planning, writing, revising and editing texts. These activities are recursive, interactive and they may happen simultaneously. An executive control monitor oversees the whole process (Hayes & Flower, 2002:25).

Hayes (1996) identifies two main focuses of the writing process: the social audience for whom the writing is intended, and the physical text must be considered. Facets of the writing process, such as the social and physical environment, motivation and working long-term memory are regarded as determining factors in writing performance. Hayes and Flower discuss the cognitive processes that link these elements generally and not in any specific order. Their model does not distinguish skilled writing from unskilled writing and is too imprecise to be used for predicting what writers might do in real life (Grabe & Kaplan, 1996:92; Weir, 2005:109).

Bereiter and Scardamalia (1987) distinguish between writing processes of skilled and less skilled writers by referring to knowledge telling and knowledge transformation. Knowledge telling involves writing without much planning or revision. This is more 'natural' writing and involves merely putting ideas down on paper. All writers do this type of writing. Knowledge transformation is demonstrated by more skilled writers and includes planning, organising and revising the ideas. Skilled writers spend more time planning and revising their writing, whereas novice writers are more concerned with providing content from their internal resources. Novice writers tend to focus more on telling what they remember about a topic. The interaction between developing knowledge and the text is continuous and both the text and writer's ideas can change (cf. Hyland, 2002:28).

Bereiter and Scardamalia do not consider the produced text as the main concern. They focus on cognitive processes taking place during writing, but ignore performance conditions and its potential effect on theory based-elements. Thus they do not address the contextual factors that affect the writing process (Weir, 2005:110).

Grabe and Kaplan (1996) integrate the social, cultural and/or cognitive aspects of writing to form more comprehensive models of writing. They advocate a socio-cognitive approach to writing and portray writing as a form of communicative language use. Based on a model of L1 writing, they illustrate the cognitive demands of L2 writing in consideration of contextual factors that influence the writing process. Their model is driven by the question: "Who writes what to whom, for what purpose, why, when, where and how?" (Grabe & Kaplan, 1996:203).

They describe text as a multi-dimensional construct (Grabe & Kaplan, 1996:80; 202-203). Through writing, information is communicated within certain accepted linguistic, psychological

and sociological principles, which in turn influence the organisation and structure of a text. These principles include the need to be informative, conventions for conveying status, intent, situation and attitude, and mechanisms to point out new information.

According to Grabe and Kaplan (1996:226-230), writing involves goal setting as a cognitive process, which activates "verbal working memory". Goal setting involves assessing the context, drafting the text, evaluating possible problems, considering the required genre and constructing an organisational plan. Verbal working memory comprises three elements, viz. language competence, world knowledge and metacognitive processing which is used to assemble language competence and world knowledge. They suggest that an internal goal always mediates the effect that context has on the verbal working memory. The writer considers the topic and required genre and generates ideas using existing schemata and/or the provided input. Ideas about the topic are organised and translated into appropriate and logical language. Metacognitive awareness and conscious monitoring of the text, whether written or in the writer's mind, are also important activities in the writing process (Grabe & Kaplan, 1996:229-230).

Grabe and Kaplan's (1996) model is useful in that it describes how a number of processing elements interact, but they do not discuss the sequence in which the process develops.

Field (2004) describes the stages of processing that a writer engages in, as well as the operations within the levels, based on psycholinguistic theory and information-processing principles. He describes writing as a process of recurring looping back between interactive stages in the process. Field (2004) distinguishes three elements of planning: macro-planning, organisation and micro-planning.

Shaw and Weir (2007:37) regard Field's (2004) model as more "accessible, detailed and structured" than those of Hayes and Flower, Bereiter and Scardamalia, or Grabe and Kaplan. They build on Field's (2004) model, postulating that writing is a form of communicative language use and that text is communicative discourse (Shaw & Weir, 2007:62). Shaw and Weir (2007) identify the following cognitive stages involved in writing: macro-planning, organisation, micro-planning, translation, monitoring and revising in their socio-cognitive model. Macro-planning refers to gathering ideas and identifying major constraints such as genre, readers and goals. Organisation involves ordering ideas and identifying those ideas that are relevant for the purpose and audience. Micro-planning entails planning at paragraph level. The goal and the

content of each paragraph are aligned with the goal of writing. Translation takes place when abstract mental ideas are converted into linguistic form. Monitoring involves checking the text for accuracy in terms of mechanical features as well as how clearly it reflects the writer's intentions (Shaw & Weir, 2007:38-39).

Shaw and Weir (2007:36) argue that the processes of macro-planning, micro-planning and organisation happen separately and in sequence. They suggest that elements of goal setting, as identified by Grabe and Kaplan (1996), are first gathered together before organisation takes place. These planning stages are what distinguish good writers from weaker ones. Decisions made about task settings and linguistic and contextual demands of the task (discussed later in this chapter) influence cognitive processing and the resources learners need in order to complete a task successfully. Metacognitive strategies play an important role as the mechanism that initiates interaction between contextual demands and the demonstration of an ability. In other words, it provides access to executive resources (Chapelle, 1998:43).

As mentioned above, assessment tasks must make the same cognitive demands on learners' abilities than a real-life task would make, but it is difficult to assess which cognitive processes are activated in the end. However, Cooper (1986:371) explains that "by focusing our attention on what goes on in an author's mind, it forces us to conceive all significant aspects of writing in terms of mental entities". O'Sullivan (2005:3) points out that test developers can only hope to impact the cognitive processes directly in the first stages of development by stipulating criteria and ensuring that task rubrics are unambiguous. These criteria should also be reflected in the rating scale to ensure a valid assessment of the task response.

Since the criteria specified in the task rubric and assessment scale depend on the definition of the construct, it is imperative that the definition be clearly defined and specific. Without a specific definition of the writing construct, the development of valid assessment instruments is jeopardized, because cognitive validity relies on explicit criteria based on the construct definition in order to avoid construct under- or over-representation. Developers therefore need to specify the features that constitute the construct. The content of both the examination paper and assessment scale should be based on this specific construct definition in order to ensure that the relevant cognitive processes are stimulated.

"Learning Outcomes" in the NCS state the aims of teaching and learning at different grade levels (NCS, 2005:6). These statements therefore describe the construct of assessment for listening and speaking, reading and viewing, writing and presenting, and language. The Learning Outcome for Writing and Presentation states: "The learner is able to write and present for a wide range of purposes and audiences using conventions and formats appropriate to diverse contexts" (NCS, 2005:13; LPG, 2008:19). It further describes writing as "a powerful instrument of communication that allows learners to construct and communicate thoughts and ideas coherently" and encourages "frequent writing across a variety of contexts, tasks and subject fields" to help learners become competent writers who can communicate functionally and creatively (NCS, 2005:13). This definition and description may imply a broad range of specific features of writing, but mainly emphasises the communicative nature of the skill, while emphasising the appropriateness of the type of discourse necessary to fulfil the communicative goal.

The Writing and Presenting Learning Outcome applies to all three sections of the FET Writing examination (Paper 3), although each section assesses a different type of writing with different genres, text types and purposes. As mentioned above, different types of texts, functions and purposes require different cognitive processing. The construct definition should be a reference point for developing assessment instruments to avoid construct under- or over-representation. It should indicate those aspects of writing that are most important in each of the three sections and would constitute a successful performance if achieved. It should, therefore, indicate criteria for assessing the relevant features for each particular type of writing. Currently the definition provided in the NCS (2005) does not specify individual aspects (criteria) to be achieved in each section of the Writing paper, but rather provides general assessment guidelines. By focussing on functional and creative communication, the Learning Outcomes suggest a socio-cognitive construct of writing, but provide little guidance with regards to what should be assessed in terms of the cognitive aspects of writing.

In advocating the communicative approach to learning and teaching, the DoE regards writing as a process, which involves stages such as planning, drafting, revising and editing (Mafisa, 2008:44). According to the NCS, assessment standards embody the skills, knowledge and values that learners need in order to achieve the learning outcome. Assessment Standards are "criteria that collectively describe what a learner should know and be able to demonstrate at a specific grade" and "collectively show how conceptual progression occurs from grade to grade" (NCS,

2005:7). Although the Learning Outcome pertaining to Writing and Presenting (NCS, 2005:13) does not refer to particular cognitive processes of writing or writing as a process, the Assessment Standards (NCS, 2005:13) make reference to planning and using writing strategies.

The NCS (2005:7) indicates the Language Programme Guidelines (LPG, 2008) as reference for the specific scope for learning and assessment. It demonstrates a plan for achieving learning outcomes to assists teachers in planning and designing "quality learning, teaching and assessment programmes" (NCS, 2005:7).

Both the Subject Assessment Guidelines (SAG) (2008:17-19) and the LPG (2008:19) stress that writing is a process, comprising brainstorming, research (if appropriate), planning and organising ideas, using various strategies and techniques to provide a first draft. Learners should be able to reflect on their first drafts, review and edit the draft until they can present a final one. Test takers are required to demonstrate these skills and provide evidence of planning and proof-reading (SAG, 2008:19; NCS, 2005:33-39). The SAG (2005) provides an example of how each Learning Outcome could be assessed. According to the SAG (2005:19), creative writing should be assessed for planning and the process of writing. Five marks will be allocated for planning and editing the first draft in Paper 3 of the end-of-year NSCE in addition to marks allocated to the final product. The 2008 SAG, however, only provides an example for assessing the first Language Outcome, namely Listening and Speaking. This may cause confusion amongst teachers whether they need to consider the five marks for planning or not.

Assessing aspects such as evidence of planning or editing and proofreading a first draft is problematic. No specifications are provided in the NCS (2005), SAG (2008) or the LPG (2008) as to how learners are meant to demonstrate planning skills, writing strategies and techniques, or reflection on their own work. Nor are criteria specified for assessing these skills. For example, no criteria are specified for allocating the five marks for planning in addition to the mark for the final draft, as mentioned above. It is unclear if or how learners' use of techniques such as mind maps and flow charts to plan and organise their ideas coherently (NCS, 2005:33) should be incorporated into the rating scale. Should five marks be added to the final total, or should the additional 5 marks make up the final total? For example, if an essay totals 50 marks, do raters score 45 marks and add the five marks for planning, or do they score 50 marks and add five marks to total 55 marks?

Furthermore, the present scale ambiguously addresses coherence under both the language and content criterion (cf. Appendix A). The content criterion addresses evidence of planning in conjunction with coherence, the degree of insight into a topic, imaginative ideas, development of details and a critical awareness of language and the impact of language. The language criterion also addresses coherent sentences and paragraphs, as well as the degree to which a text is error-free following proof-reading. Although the scale refers to planning and coherence, it does not allow raters to assess these features explicitly and individually. Thus, the current scale fails to guide educators and raters in implementing the requirements of the NCS (2005) and SAG (2008) and assess aspects of cognitive processing involved in writing.

Matriculation standards of subjects such as English FAL have long been a topic of concern for the public and newspapers. Results of a study conducted at the University of Pretoria in 2001 indicate that one third of first-year learners are at the same level of proficiency as Grade 8 level learners. Results indicated that the Grade 12 language results do not cot correlate with actual proficiency. In 2005, a research forum of Umalusi investigated whether expectations as expressed by the level of complexity of the examination questions are reached, and whether the level of cognitive demand declined from previous years. Results of this investigation were unambiguous: "In nearly every case the researchers found that the level of cognitive demand had declined, as public opinion had surmised" (Muller, 2005:2).

The Learning Programme Guidelines (LPG) (2008:10) state that "additional language speakers need to use and discuss language in an academic way (metalanguage)". This entails evaluating information and using it to substantiate their points of view in a coherent discussion. Learners at the end of Grade 12 are therefore expected to posses the necessary skills to perform adequately at tertiary level. The NCS (2005) also specify that the learning outcomes for FAL indicate the threshold requirements necessary for effective learning across the curriculum, particularly since many learners are not taught in their native language. "This includes the abstract cognitive academic language skills required for thinking and learning" (NCS, 2005:11).

However, research indicates that many learners do not have the ability to comply with this requirement. According to McCabe (2007:17), learners may have basic interpersonal communicative skills (BICS), but they lack cognitive academic language proficiency skills (CALP). Learners should acquire the necessary competence at secondary level, but many do not. Learners from low socio-economic backgrounds do not seem to use effective learning strategies,

such as metacognitive, cognitive and social or affective strategies (Mafisa, 2008:13-14). Kapp (2004:260) notes that many South African ESL learners do not possess the conceptual framework, metalinguistic tools or vocabulary that writing tasks require. Some teachers only vaguely engage in meta-talk about the language system and structuring of text, or do not discuss it at all. Kapp (2004:251) notes that classroom writing tasks tend to focus on transactional writing, defined by the Department of Education's guideline document (WCED, 1997:1) as factual, real-life writing such as dialogues, business and formal letters and creating information brochures, which are assessed in Section B of Paper 3.

Section A of the NSC Writing paper aims to assess extended communicative writing ability through essay writing. This entails that learners need to demonstrate analytic writing abilities, viz. to interpret information, compare viewpoints and argue a point of view (cf. 4.4.1.2). These skills must be demonstrated in a way that is appropriate in a variety of contexts and acceptable for various audiences.

Kapp (2004:251) points out that transactional writing formats directly oppose the objective of seeing writing as a process that aids cognition, as these documents have set formats that can be learnt, memorised and reproduced under examination circumstances. However, these skills are assessed in instrumental ways that focus on the state of knowledge as opposed to the way of knowing, despite the fact that the DoE encourages developing learners' critical analytic skills and promotes an approach to writing as a process (Kapp, 2004:251, with reference to Bernstein, 1990:98). Thus, perhaps too little time is spent in some classrooms on developing learners' extended writing skills, as assessed in the composition section of Paper 3.

Kaplan (1972:63) explains that each language and culture has a unique paragraph structure. "[P]art of the learning of a particular language is the mastering of its logical system". Kramsch (2005:245) describes an English paragraph as represented by a straight downward arrow, whereas an Oriental paragraph circles towards the centre in a spiral. She notes that the main problem may not so much be that the one language is, for example, more direct than the other, but that different educational systems value different styles. Culturally-related preferences for organising and expressing a message differ and learners are likely to fall back on their L1 ways, particularly less proficient learners. With the low level of proficiency amongst South African Grade 12 learners, it can be expected that learners will borrow much from their first language.

In summary, it is clear that the test-taker's cognitive abilities are at the heart of cognitive validity. In the present study this ability can be defined as the ability to construct a sustained piece of writing in order to communicate with a reader in a second language at Grade 12 level. Learners need to demonstrate their ability to produce discourse in which thoughts and language are related in a way that produces an appropriate structure of meaning. The writing process involves planning (where ideas are generated on a topic, taking a specific genre into account, and setting goals for the writing), organising the piece of writing (using the appropriate conventions in support of the discourse and type of text, planning the sections and their logical flow) and presenting an intelligible text (using accurate and appropriate grammar, vocabulary, spelling and punctuation).

4.4 Context validity

Context validity refers to features of the task setting that describe the required performance and the linguistic and content demands that must be met in order to complete a task successfully (Shaw & Weir, 2007:63). Writing is never an isolated event, nor is learning a second language. Both take place in a social context and the relationship between writing and/or learning a language and the context is a complex one (Barkhuizen, 2005:552-553). The cognitive processes involved in writing are not activated in a vacuum, but in a specific social context. In terms of assessment, the wording of test tasks sets the contextual parameters in which the task must be performed, which activate certain cognitive processes.

Contextualisation cues are important in order to make sense of conversation (Gumperz, 1992; Kramsch, 2005). Gumperz (1992:230) describes contextualisation cues as those features of speech (or text) that "relate what is said at any one time and in any one place to knowledge acquired through past experience, in order to retrieve the presuppositions [learners and raters] must rely on to maintain conversational involvement and assess what is intended". The audience and the purpose of writing are important contextual factors that influence which cognitive processes are activated directly. Test takers have to consider features such as the status of participants, degree of formality and expressing different attitudes, and linguistic contexts.

Different types of talk are appropriate in different contexts and test takers must demonstrate that they are able to cope with various contexts in different situations (Fulcher, 2000:490). Context plays a paramount role in determining communicative language ability (Shaw & Weir, 2007:63).

According to Weir (1993:28), it is important that "the context must be acceptable to the candidates and expert judges as a suitable milieu for assessing particular language abilities". It is therefore necessary to consider any interaction between cognitive processes activated by tasks and the performance context.

Stakeholders generalise test takers' performances as indicative of their ability to perform in real life. A representative sample of communicative ability increases the generalisability of performances. Although it is not possible to have fully authentic tasks in assessment situations (Shaw & Weir, 2007:64), test designers should consciously try to simulate the normal conditions under which the abilities would be performed in real life to ensure credible inferences about learners' abilities to function in real-life situations (Weir, 1993:28-29). Therefore, tasks must be selected according to general descriptive parameters of the target situation, particularly with regard to the skills needed to participate successfully in the target situation. These tasks should be presented in a context that resembles real-life communicative contexts.

The NCS (2005:46) emphasises that teachers must consider and incorporate appropriate content and context when assessing whether learners have achieved Learning Outcomes. Context is a valuable aspect of a communicative, text-based approach to language teaching and learning, as advocated by the Department of Education (DoE).

Communicative language teaching and assessment intends to provide information about learners' ability to perform in the target language situations in particular context-specific tasks (Miyata-Boddy, 2000:75). Communicative assessment tasks should be as authentic as possible and should involve realistic discourse processing (Weir, 1990:12). The NCS (2005:47) views the communicative approach to learning a language as follows:

> The communicative approach means that when learning a language, a learner should have a great deal of exposure to it and many opportunities to practise or produce the language by communicating for social or practical purposes. Language learning should be a natural, informal process carried over into the classroom where literacy skills of reading/viewing and writing/presenting are learned in a "natural" way – learners read by doing a great deal of reading and learn to write by doing much writing.

Following a communicative approach, the DoE recommends providing learners with extensive opportunity to practise or produce language in order to solve problems and interacting in social

situations (LPG, 2008:10). The communicative approach has various implications for classroom practice, as noted in the Learning Programme Guidelines (LPG, 2008:10). For example, language should be taught in an integrated way and language structures should be taught in context. Errors are seen as part of the learning process but the focus is on the effective communication of meaning rather than on using the correct form of language. If teachers recognise that learners struggle with tenses, they can adapt classroom practices and provide additional exercises to help learners master the particular structure.

Following a text-based approach, educators should work towards "enabling learners to become "competent, confident and critical readers, writers, viewers and designers of texts" (NCS, 2005:47). A text-based approach entails exploring how texts work, in other words how they are produced and what their effects are. Different genres of texts are produced to fulfil functions for particular purposes and with certain audiences in mind, while implementing the appropriate conventions and formats for each specific context.

According to the LPG (2008:9-11), a text-based approach involves critical interaction with texts to understand how they are produced and what effects they have. The approach is informed by an understanding of how texts are constructed and explores the interaction between learner and text. Learners need to recognise that texts are produced for particular contexts (in other words for a certain audience and purpose), and that texts reflect the context in which it was created (e.g. political, social, or cultural). This type of critical interaction develops learners' ability to evaluate texts, which entails producing a variety of texts for particular purposes, audiences and contexts.

The NCS does not provide specific guidelines in this regard and encourages teachers to choose content and context based on their personal views: "[T]eachers should be aware of and use local contexts, not necessarily indicated here [in the NCS] which could be more suited to the experiences of the learners" (NCS, 2005:46). Furthermore, vague descriptions provide limited guidance and unclear requirements regarding content selection. For example, learners must be exposed to "rich and appropriate social, cultural and historical settings that develop understanding of the heritage of the language" and "challenging and stimulating themes that develop critical understanding of values and appreciation of the important socio-cultural and ethical issues which are relevant to the lives of South African learners".

These contexts should serve to embed the content suggested by the NCS in meaningful situations to assist learning and teaching. The NCS (2005:48-49) suggests that narrative, descriptive, reflective, discursive, expository and argumentative compositions, as well as responses to literature, make up the content of assessments for creative writing (cf. also SAG, 2008:12; DoE Examination Guidelines, 2009:5). Although the unspecific guidelines may accommodate learners from different socio-economic and cultural environments, less experienced teachers may find it difficult to interpret and implement these guidelines successfully.

Test takers base their decisions on the contextual parameters established in the wording of tasks. In order to promote context validity, all production demands placed on the writer in assessment situations must be made very clear to the test taker. Task rubrics should include parameters of both content and context, and operationalise as many parameters as possible. The contextual parameters should be presented explicitly in the task input and the test rubric. Parameters such as the task setting and linguistic demands activate the relevant cognitive processes to accomplish the task. The three main types of contextual parameters that are most likely to influence test performance are task setting, administration setting and linguistic demands (Weir, 2005:47; Shaw & Weir, 2007:64). Each of these parameters is discussed below.

4.4.1 Task setting

Task setting, as described in Shaw and Weir's (2007) framework, refers to task-related features. Test takers should be informed about the task setting before the examination. Task specifications and instructions should describe the task setting in terms of the response format, purpose, knowledge of scoring criteria, weighting of different tasks, the required length of the response, time limitations and the writer-reader relationship. Bachman and Palmer (1996:121) suggest that test instructions should be simple and easy to understand, short, concise, yet detailed enough so that test takers know exactly what they are expected to do.

The NCS Assessment Standards (2005:33) state that learners must be able to decide on and apply an appropriate style, point of view and format of text. The task rubric should guide test takers in this regard and the rating scale should assess performances accordingly.

4.4.1.1 Response format

Shaw and Weir (2007:61) point out that the technique or format of assessment has direct implications for context validity, as it does for cognitive validity. The format determines whether knowledge telling or knowledge transformation is required from test takers. It is better to assess a range of formats, including essay type questions and transactional writing, in order to gain a complete indication of learners' communicative writing abilities (Alderson et al., 1995).

Each section in the FET Writing paper assesses a different response format, viz. an essay in Section A, one longer transactional text such as a *curriculum vitae* or letter in Section B, and one shorter text such as a diary entry or invitation card in Section C.

The essay is a traditional form of direct writing assessment, which requires learners to produce samples of connected writing (Weir, 1990:60). It provides an opportunity to test cognitive strategies that are not involved in transactional writing. Weir (1990:60-61) summarises the disadvantages and advantages associated with essay tasks. Learners from different cultures and environments have different background knowledge and approach open-ended questions differently. This may result in a great variety of responses that are difficult to score, particularly when different types of texts are produced. Learners also have to perform under abnormal time constraints (Weir, 2005:164).

Essay writing as response format has traditionally been a popular means of assessment. Essay tasks accommodate learners from different circumstances, offering them equal opportunity to present their best performance, since a variety of interests can be accommodated. In addition, essay tests provide a suitable means for assessing skills such as the ability to develop a logical argument. According to Weir (1990:60-61), the major advantage of essay tasks, however, is that they require learners to produce a sample of writing that serves as a tangible point of reference for future comparisons. The perceived content validity of 'job sample' tasks makes a very strong case for including writing examination. "It tests important skills which no other form of assessment can sample adequately" (Weir, 1990:61). Omitting writing tasks when writing is an essential communicative tool in real life will severely lower the validity of an assessment of communicative language abilities.

In order to enhance equality in assessing communicative writing, the FET examination usually presents written stimuli which can vary from one word to several sentences. Poems, quotes and pictures are also sometimes used. Very general topics are used and test takers have to use their resources such as background knowledge, cultural knowledge and topic knowledge, as well as their imaginations. In the 2008 FET paper, learners were provided with some ideas on how to address the topic and what their essays should offer. These ideas may guide learners in planning and structuring their response in an appropriate format. However, learners are not required to include the ideas provided as guidelines.

4.4.1.2 Purpose

Real-life written communication always takes place for a reason. In communicative testing, learners have to demonstrate that they are able to recognise the communicative purpose and respond to this purpose appropriately (Fulcher, 2000:490). Thus the task rubric must clearly and unambiguously specify the reason for completing the task and the rating scale must guide raters to assess the degree to which the purpose was successfully achieved.

A clearly stated purpose is critical to test takers' macro-planning, because a reason for writing serves to motivate test takers and helps them to plan and structure their thoughts (Bachman, 1990:124; Bachman & Palmer, 1996:181). Weigle (2002:10) categorises text types according to the following main intentions: to learn, to persuade or convince, to convey feelings, to entertain, and to keep in touch. These intentions can be realised through three levels of cognitive processing: reproducing information, organising known information or generating new ideas. The communicative purpose determines which of these processes needs to take place.

The choice between writing tasks in Section A of the FET Writing paper offers learners a variety of communicative purposes to achieve. Each task's purpose is determined by the genre required, for example describing or interpreting pictures or situations, arguing a point of view on a topic or entertaining through storytelling. In the FET paper written at the end of 2008, for example, learners could write an essay to persuade the reader, to convey their feelings and viewpoints on a particular issue, or to entertain through story telling or giving a description. Offering learners such a wide choice of communication purposes allows learners to demonstrate their best performance. However, it may make consistent scoring of essays problematic.

The Subject Assessment Guidelines of the NCS (2005) do not provide guidelines for assessing learners' achievement of communicative purpose and the current rating scale does not address the issue explicitly or implicitly. Raters receive guidelines, in addition to the rating scale, for scoring different topics each year. The instructions differ for each essay topic according to its purpose and genre.

4.4.1.3 Knowledge of criteria

Learners' knowledge of assessment criteria influences their planning and monitoring of the texts they produce (Shaw & Weir, 2007:77). The criteria used to assess performances must be communicated to the learners before they write the examination so that test takers can spend all their time planning, organising and monitoring their output with regard to the assessment criteria.

The NCS (2005:53) stipulates that assessment should be based on the criteria indicated in the Assessment Standards. This implies that the assessment instruments must reflect the Assessment Standards and be accessible to and appropriate for learners at different levels. As mentioned above, the Assessment Standards state certain requirements in order to achieve the Learning Outcomes, but do not indicate what constitutes success or proof of achievement in terms of the requirements. Thus, there may be some confusion regarding specific assessment criteria to achieve outcomes.

The two criteria (language and content) in the current rating scale are fairly general and somewhat vague (cf. Appendix A). It is, for example, not clear whether the language criterion should be interpreted in terms of grammatical accuracy. It is also not clear which aspects the content criterion represents, since a combination of features that may not be clearly related such as length, coherence and idiomatic language, is addressed under this criterion.

The NCS also stipulates a separate Learning Outcome related to Language: "The learner is able to use language structures and conventions appropriately and effectively" (NCS, 2005:38). It is not specifically stated that these guidelines are relevant in the assessment of the Writing paper, but it can be assumed. The rating scale, therefore, has to reflect these guidelines as well as those stipulated for the Writing and Presenting Outcome. The language criterion in the current scale considers some features that are specified under the Learning Outcome for Language, such as

critical language awareness. However, neither the Learning Outcomes nor Assessment Standards indicate how learners should demonstrate their critical awareness of language in writing. Without a clear definition, it is difficult to standardise raters' application of the rating scale to assess such features.

According to the DoE's guidelines for setting examination papers (DoE Examination Guidelines, 2009:6), FAL essay performances for examination purposes must be assessed in terms of content and planning, language, style and editing; and structure. Raters are also provided with a different scale than the one used to assess performances for the FET examination (DoE Examination Guidelines, 2009:14). Using two different scales for scoring essay writing performances during different examinations may influence the scores raters assign to performances (cf. 5.2.1). It may therefore be that test takers perform better on one examination than on the other. Their higher scores may, however, be a result of the rating scale used to assess performances, rather than the learners' own abilities.

4.4.1.4 Weighting

As mentioned above, information regarding the task setting (cf. 4.4.1), including information regarding the weighting of tasks, is generally communicated through the test instructions and task rubric. Test takers should know the weighting of each section, question or aspect of a question so that they can prepare and plan their responses accordingly (Shaw & Weir, 2007:80). Different weightings within an assessment must be based on a clearly defined rationale. The marks assigned to different sections, tasks or items must reflect the relative contribution and importance of that aspect of the construct being assessed (Weigle, 2002:101-102; Shaw & Weir, 2007:77-81).

The total for the FET Writing Paper is 100. Section A, the essay, counts 50 marks of the total, while Section B, the longer transactional text counts 30 and Section C, transactional/ reference/ informational text, counts 20 marks. The Writing paper as a whole makes up 100 marks of the total mark of 300 for the final end-of-year English FAL external examination (SAG, 2008:31).

The current rating scale allocates equal weighting to language and content (cf. Appendix A). It addresses each criterion in terms of a number of features. Raters are expected to consider each of these features individually before deciding on a mark for the criterion. However, the format of

the current scale does not allow raters to differentiate the features of writing addressed under each criterion, because raters have to choose one score to reflect learners' achievements on all features listed under Language and Content. It is particularly important when assessing second language writing that raters be able to assess features individually because different features develop at different rates (cf. 5.2.1). Raters may intentionally or unintentionally focus more on one or a combination of the features, rather than consider all features related to a criterion equally.

4.4.1.5 Text length

Weir (2005:74) notes that the length of text that test takers are required to engage with and produce must be realistic in terms of the requirements of the target situation. Different lengths of texts require different cognitive processes. Shorter tasks do not make the same cognitive demands as longer texts and may not require the necessary resources required to perform the construct in real-life situations, compromising cognitive validity (Weir, 2005:74; Shaw & Weir, 2007:81).

The NCS (2005:21) and Examination Guidelines (2009:5) recommend that Grade 12 FET test takers produce an essay of 250-300 words. The required length is stipulated in the instructions on the examination paper. Learners gain experience in writing essays within time constraints during the mid-year examination, which has the same limits as the final examination.

4.4.1.6 Time constraints

Time allocation is related to the type of writing tasks. Shaw and Weir (2007:83) note that time constraints make longer process-oriented tests impractical in most situations, since responses in real-life situations are usually not timed as strictly as in examinations, despite deadlines that have to be met. According to Weigle (2002:101-102), time constraints have direct implications for essay writing, although time limits do not differentially affect particular groups. Rather, the effects of time limits on test takers are culturally dependent, based on what test takers are used to in their normal environments. Learners do not necessarily perform better when given more time to write.

Weir (1990) estimates that thirty minutes should be enough for the average learner to produce approximately 150 words. However, since writing is a process, the time limit in examination

settings is often an unrealistic constraint because it does not allow learners to write and re-write several drafts (Weir, 1990:61).

The Subject Assessment Guidelines (2005:12) stipulate that three hours should be allowed for all three sections of the Writing paper. The 2008 FET paper advised learners to spend approximately 80 minutes on Section A.

4.4.1.7 Writer-reader relationship

Both the reader and writer are involved in written communication. Grabe and Kaplan (1996:207-208) regard the audience as critical in creating meaning in a text. Test takers can either write for an addressed (real) audience, or for an imagined one. In order to create the appropriate impact, the test taker must gauge the degree to which the reader will be able to interpret the message, and the reader's probable reaction to the message (Hyland, 2002:72; Shaw & Weir, 2007:86). Grabe and Kaplan (1996) identify five factors related to the audience that influence a writer's decisions about how to create a text, viz. the size of the audience, degree of formality, the status of the reader, the extent of their shared background, and the extent of the their topic knowledge. Hyland (2002:69) notes that, as proficiency levels increase, there is a progressive need for writers to address the wider social and discursive practices in terms of the context, purpose, audience and genre.

Examiners usually constitute the audience for whom FET test takers write. Learners as well as raters have to consider that their cultural, economic or language backgrounds may be different to that of the person who rates, or wrote, the text. These differences may influence them to interpret the text differently than intended. In the NCS (2005), both the Assessment Standards and Learning Outcome emphasise that learners must demonstrate sensitivity for and consideration of various audiences in their writing. Test takers must show that they are sensitive to social concerns such as human rights, social, cultural, environmental and ethical issues such as gender, race, disability, age, status, religion, and diseases such as AIDS (NCS, 2005:43).

Cultural differences have particular implications for the relationship between the reader and the writer. The examiners and test takers involved in the FET examination often do not share the same background. McCabe (2007:22) points out that individuals from different backgrounds may read a text differently. Different language groups may also vary in how they use rhetoric,

organise arguments, use cohesion, cater for readers and use secondary sources, which may not always be acceptable in another language such as English (Grabe & Kaplan, 1996:239). Both the test takers and the raters have to keep these differences in mind.

However, many teachers are not aware of or do not consider the economic backgrounds of learners and tend to respond to learner performances in ways that alienate learners (Smith & Liebenberg, 2003:2; Mafisa, 2008:8). These teachers often act as examiners in the FET examination, which means that learners may be penalised unfairly due to raters' lack of understanding of their background.

In the current rating scale, social awareness is listed as a feature to be demonstrated through learners' choice of vocabulary and language. However, it is not clear what kind of evidence would indicate this awareness. Learners may address the audience politely and in an acceptable manner using inoffensive language, but it is difficult to score this feature in terms of different levels. Furthermore, the scale does not specifically address the appropriateness of genre and particular conventions associated with certain functions, audiences or genres.

4.4.2 Linguistic demands (task & input)

In order to generalise from test results to predict language use in target language situations, test tasks should make the same the same linguistic demands on test takers than those made by communicative tasks in real life at the targeted performance level. Fulcher (2000:490) points out that for tasks to be authentic, input and prompts should not be simplified for learners. In order to understand input and function productively in writing, learners must have an adequate knowledge of vocabulary, of basic grammatical items and of their appropriate use for different functions (Weir, 1990:58). Shaw and Weir (2007:91) base their interpretation of linguistic knowledge on a communicative approach to modelling language ability (Bachman, 1990; Bachman & Palmer, 1996).

Shaw and Weir (2007) distinguish a number of linguistic aspects or resources that learners use when responding to a task, viz. lexical resources, structural resources, discourse mode, functional resources and content knowledge. Linguistic features – such as lexical and structural resources, discourse mode, functional resources and learners' content knowledge – play a vital role in the level of performance learners are able to produce in an assessment. Linguistic

demands are made on both receptive and productive skills of learners at lexical and grammatical level.

4.4.2.1 Lexical resources

Lexical items used in task input and required in task output should demand an appropriate level of skill from test takers. However, it is difficult to select which lexical items should be assessed at a particular level and the best way to assess them. This is the case particularly when test takers study a variety of subjects. According to Weir (1990:58-59), it is easier to select appropriate items for the purpose of assessment in cases where there is an identifiable, agreed register, than for more generalised subject matter. However; the task remains challenging. Weir also raises the question of how to establish the frequency and the importance of lexical items intended for the purpose of assessment.

Empirically-based word lists compiled from learner writing corpora can be used to determine the appropriate level of lexical complexity. Based on an analysis of writing task prompts, Schmitt (2005) found that the average number of words, the number of different types of words and the mean length of words used in task input increases progressively with higher levels. Shaw and Weir (2007:95) suggest that, even at higher performance levels, input and prompts with longer texts should contain vocabulary that is common, general, non-specialist, without cultural reference and at a language level that is fully comprehensible for learners at the target level. The DoE stipulates that "topics should be concise and in language that is accessible to candidates" (Examination Guidelines, 2009:5). The input provided in the FET Writing examination paper therefore consists of non-specialist, general vocabulary in short phrases or at most a few sentences. The Assessment Standards stipulated for the Writing and Presenting and the Language Outcomes (NCS, 2005) provide some general guidance regarding the lexical requirements for learners at Grade12 level.

In terms of output requirements, topics, tasks and functions that only require simple language should be avoided at higher levels. Learners should demonstrate an appropriate level of lexical command in the output they generate in response to the prompts while performing the functions required (Shaw & Weir, 2007:107). Learners must be familiar with vocabulary relevant to the language functions they have to fulfil, such as description, comparison or arguing. More advanced learners could be expected to use longer words of lower frequency. Shaw and Weir

(2007:94) note that the lexicon learners produce should reflect both depth and breadth of vocabulary knowledge, although breadth of knowledge is much more difficult to capture.

Shaw and Weir (2007:94) furthermore distinguish between 'known vocabulary' and 'used vocabulary'. Learners may know a wider range of vocabulary than they actually choose to use in their communication. Learners at more advanced levels, such as Grade 12, should be able to use vocabulary to perform the required language function effectively. Much of the vocabulary will be frequent, but some words may not be. In addition, learners at advanced levels should be able to discuss a topic in detail and even provide more precise information. This would, for example, require a wider range of adjectives for description purposes (cf. Schmitt, 2005).

Rating learners' output in terms of lexis is very problematic, particularly in cases where they produce extended texts. It is difficult to control the target words learners produce and evaluating the vocabulary produced is even more problematic (Shaw & Weir, 2007:102). For example, Schmitt (2005) found that lexical variation increases over lower performance levels, but is not a reliable measure of lexical mastery at higher performance levels. Determining the difficulty of a word is a complex process, and therefore the frequency of words is generally used to rank vocabulary. Higher frequency words are generally acquired before low frequency words (Schmitt, Schmitt & Clapham, 2001). Shorter words tend to be more frequent than longer words (Zipf's law, 2008; also cf. Zipf, 1949, 2006). The mean word length also increases across input for more advanced performance levels. Shaw and Weir (2007:96) suggest that a general marking scheme should make reference to elements such as sophisticated and/or fluent use of vocabulary, collocation and expressions that are appropriate to the task, since it is very difficult to specify requirements concerning lexical resources explicitly in the rating scale.

In the FET examination, learners should therefore demonstrate an extended vocabulary. The Assessment Standards of the Writing and Presenting Learning Outcome do not address lexical features explicitly, but some guidelines are provided in relation to the Language outcome (NCS, 2005:40). These include, for example, that learners should demonstrate the ability to use a range of common abbreviations and acronyms correctly, distinguish between commonly confused polysemes, homophones and homonyms and use them correctly, and use synonyms and antonyms correctly (NCS, 2005:39).

Owing to a lack of resources, materials and exposure to English, many learners may not have reached an adequate level of lexical development. In circumstances where learners do not have access to books at school and where reading outside the classroom is not highly valued, learners' reading skills tend to be underdeveloped, which may affect the development of their lexical resources negatively. These learners are likely to have under-developed vocabularies (Cummins, 2000; Balfour, 2002; McCabe, 2007).

It is difficult to address each lexical feature specifically in a rating scale. The current rating scale (cf. Appendix A) refers to "choice of words" under the content criterion, but whether this refers to range, accuracy or both is unclear.

4.4.2.2 Structural resources

Grammar is continuously identified in *a posteriori* validations as a significant factor in interpreting scores (McNamara, 1990; Rimmer, 2006). In addition, Rimmer (2006:497) notes that grammar correlates highly with overall proficiency and is indicative of performance levels. Thus, structural resources make up an important part of assessing writing. A major requirement – the substantive requirement (Bachman, 2004:7) – for validity is that assessment is linked to a coherent description of language use. Yet, many rating scales tend to relate scores to statements of grammatical competence without a sound theoretical base for doing so. Syllabi tend to be based on intuition and tradition rather than on empirical data, which does not make for very reliable grounds to determine grammatical complexity. What is needed is a "principled and comprehensive grading and sequencing of grammar items that can be operationalised in testing" (Rimmer, 2006:499). However, such information is not currently available and it is therefore difficult to determine which structural resources should be assessed at different levels.

Two dimensions of grammar can be measured, namely accuracy and range. Of the two, range is more problematic to assess than accuracy, but accuracy alone is an insufficient measure of learners' ability to use grammar appropriately in order to communicate. Learners must demonstrate that they are able to utilise the necessary structures appropriately to perform the function successfully. Some grammatical structures may be considered more fundamental than others. If this is the case, these structures could be specifically highlighted in a rating scale.

As with the selection of lexical items, determining which grammatical items are appropriate and required at a particular level, and how to measure success, are problematic. Alderson (2000:69) considers it naive to think that certain grammatical structures are intrinsically more difficult than others. Weir (1990:59) notes that it is impractical to expect test constructors to conduct quantitative surveys to determine which structural items occur in receptive and productive materials that test takers will have to cope with in future target situations. He advises a more pragmatic, subjective way of making decisions, such as examining the content of tests and course books at an equivalent level.

Shaw and Weir (2007:109) suggest not restricting or forcing learners to use specific grammatical structures at different levels. Instead, learners should have the opportunity to demonstrate their structural resources as best they can. At advanced performance levels, grammatical specifications may fall away as learners are expected to have mastered all the basic grammatical structures, aiming to improve their fluency and accuracy (Shaw & Weir, 2007:114). The structural resources that are important for assessment purposes are those necessary for coping with the particular functions of writing set by the task. Different functions may be appropriate at different levels. More advanced students can be expected to use a wide range of structures to express themselves appropriately and accurately, while showing sensitivity to register and to consideration of the audience (Shaw & Weir, 2007:109).

Formal aspects of writing, such as grammar, are considered in the text-based approach advocated by the DoE, but these are viewed in terms of the effect they have on communication, rather than analysed in isolation (LPG, 2008:10-11). In other words, Grade 12 learners need to demonstrate the ability to use grammatical structures appropriately and effectively to perform a communicative function. This aligns with Shaw and Weir's suggestion on the matter.

Aspects related to grammar are addressed only in relation to the Language Learning Outcomes. The Learning Outcomes for Writing and Presenting and its Assessment Standards do not refer explicitly to issues related to language structures that must be considered in order to achieve the outcome. The Writing and Presentation Outcomes call for accurate and appropriate use of writing conventions (NCS, 2005:130).

The main aim is to communicate meaning effectively, and accurate grammar is one means of achieving this aim, in addition to choosing appropriate structures to fulfil the communicative

function. As mentioned above, a text-based approach involves critical interaction with texts to understand how they are produced, the effects they may have, and to be able to produce such texts. Subsequently, learners have to have an accurate knowledge of language structures and conventions in order to produce these texts.

The Language Outcomes and related Assessment Standards provide more explicit information on the structures that Grade 12 learners should have mastered. Learners need to demonstrate assessment standards, including the following, to achieve the Language Learning Outcome:

- Use structurally sound sentences in a meaningful and functional manner;
- Use a wide range of figurative language such as idioms, idiomatic expressions and proverbs appropriately;
- Develop critical language awareness (NCS, 2005:39-43).

In elaborating on these criteria, the Assessment Standards explicitly call for accurate use of a variety of grammatical features, sentence types and lengths, vocabulary, and experimentation with format and style for creative effect. Also, punctuation and concord, word order and direct and indirect speech must be used correctly for the specific communicative purpose in order to achieve the Language Learning Outcome (NCS, 2005:41).

As noted, the degree to which the rating scale should reflect these features as part of the creative writing construct remains unclear. It is unclear whether – and if so, how – the criteria for the Language Learning Outcome should be integrated with criteria for Writing and Presenting and incorporated in the assessment of writing in the FET examination. As mentioned earlier (cf. 4.4.1.3), it can be assumed that both outcomes should be considered in assessing writing. This may affect the content of the rating scale.

The language criterion in the current scale (Appendix A) addresses grammatical accuracy, but not range. Additional features addressed as language-related are idiomatic expressions, coherent organisation and text length. Criteria from the two separate Outcomes for Language and for Writing and Presenting seem to be integrated under the language criterion in the scale. Thus, it appears that the criterion does not refer only to grammatical accuracy and range. However, there is no clear description of what demonstrating the relevant features entails.

4.4.2.3 Discourse mode

Discourse mode and genre provide the writer with supportive resources for communicating appropriately in the context of language use, whether political, social or cultural (Ivanic, 2004:222-223). Urquhart and Weir (1998:141ff) argue that test developers must generate empirical evidence to determine appropriate discourse modes at different grade and achievement levels. Determining which discourse type is appropriate for particular levels is problematic for two reasons. First, there is little agreement on the terminology used to classify different texts. Second, not much research has been done to determine the effect that the required text has on the difficulty level of the task (Shaw & Weir, 2007:15).

Weigle (2002:62) regards discourse as multi-faceted, including genre, rhetorical task and patterns of composition. The rating instrument must therefore address the appropriateness of the output in terms of discourse mode and genre produced by learners within the particular context. The wide variety of schemes to analyse discourse mode complicates the situation even further.

Beck and Jeffery (2007:75-76) note that not all genres need to be represented at all performance levels and suggest making argument genres a priority in high-stakes assessment situations:

> Argument ... serves an important function as an organising macrostructure for the presentation of one's interpretive position: the introductory paragraph or paragraphs present an interpretation, and textual or historically factual examples with commentary serve as support for that interpretation ... [A]rgument might be the most pedagogically useful genre to assign as a task for high-stakes writing assessment.

Beck and Jeffery (2007) call for greater clarification and consistency in describing genre expectations, particularly in high-stakes assessment situations. They argue that teaching practices are heavily influenced by high-stakes examinations. Therefore, clarifying the expectations would enable teachers to prepare learners appropriately for writing in the required target situations, such as writing at college-level.

The DoE lists a range of genres considered suitable for assessment at Grades 10 to 12 in the LPG (2008). The same genre may be listed for different grades. In the Examination Guidelines (2009:5), the DoE indicates that the types of essays to be set are narrative, descriptive, reflective, argumentative and discursive. The LPG (2008:11-12) specifies that these genres should

progressively become more challenging from one grade to the next. Learners are expected to produce progressively more sophisticated and challenging texts.

The DoE Examination Guidelines document (2009:5) suggests that the type of essay that candidates should write on a topic should not be prescribed. This means that Grade 12 learners may choose their own discourse mode and raters are faced with the task of comparing vastly different performances on the same scale.

As Weir (1990:60) points out, it is particularly difficult to compare performances on different topics, especially when different types of tasks are being compared. As mentioned above (cf. 4.4.1.2), the memorandum each year stipulates conventions typical to the genre associated with each prompt. The chosen genre must also be appropriate to fulfil the required function. The NCS document (2005:33) indicates that learners should produce genres appropriately in the context of use and for the intended audience. However, this is difficult when using a general rating scale to assess responses covering various genres, as is required for assessing the FET Writing paper.

4.4.2.4 Functional resources

As noted above (cf. 4.3), a communicative, text-based approach emphasises functional language use. The main focus of assessment is therefore to determine what learners are able to do with language. Basic functions, such as telling a story or giving an opinion, are assessed in terms of the sophistication with which learners perform the function and by the range of exponents they use in performing the function. Other functions, such as hypothesising, can only be expected from learners at higher levels. Learners at an advanced level can also be expected to act as independent language users who are able to operate in a range of social situations. Functions such as arguing, narrating and describing must be assessed in terms of elements that indicate sophistication in performing functions. At Grade 12 level, learners should be able to demonstrate more depth and breadth in performing these functions.

In the essay writing section (Section A) of the FET Writing Paper, the communicative function of each task is determined by the genre learners are required to produce. Some topics specify the function, for example, stating that learners must discuss their point of view, tell a story or provide a description. However, some tasks are open to interpretation and learners are instructed to write an essay on "a topic that comes to mind". In such cases learners can use their own

judgement to decide which function is most appropriate in the context presented by the task. The rating scale therefore needs to be applicable to assessing various functions (across genres).

4.4.2.5 Content knowledge

Test takers' topic or content knowledge is a significant variable in test performance (Read, 1990:78). Particularly in writing examinations, choosing an appropriate topic that is realistic, feasible and reasonably familiar is very important (Hamp-Lyons, 1990:53). This is a major responsibility of a test developer (Weir, 2005:76). In this regard, it may be better to have all test takers write on the same topic, because a choice between topics in writing assignments elicit measurably different responses (Read, 1990:78), and allows too much uncontrolled variance into the assessment (Jacobs et al., 1981:1).

The relationship between the writer's background knowledge, topic knowledge and the topic knowledge required by the task determines how test takers will deal with the task (Weir, 2005:75). Test takers use their background knowledge, subject knowledge and the content of the input provided to produce a text of appropriate discourse and to perform a function effectively. Different topics elicit performances that are measurably different in quality. The interaction between test takers' executive resources and demands of the task with which the learner interacts emphasises the symbiotic nature of context- and theory-based validity (Douglas, 2000:19). Alderson (2000:29) urges that "every attempt should be made to allow background knowledge to facilitate performance rather than allowing its absence to inhibit performance".

Weir (2005:75) stresses the importance of providing texts with content that is sufficiently familiar to test takers to ensure that learners at the level of assessment have enough existing schemata to apply appropriate skills and strategies in understanding the text. Providing a text as input offers learners the same access to content knowledge and reduces potential bias to individuals. However, it does not improve the quality of learners' performances. The stimulus text may provide learners with a range of ideas, but they may tend to borrow heavily from the source text.

As mentioned above (cf. 4.4.1.1), the 2008 FET paper provided guidance for learners by providing some ideas on a topic in a bulleted list format. Learners could therefore use these ideas (in other words, borrow from the source text) as presented in the question paper, but they

would still need to rely on their own ability, content knowledge and imagination in order to complete the task successfully.

As mentioned in the discussion of test taker characteristics (cf. 4.2), many South African learners come from poor socio-economical backgrounds. Their circumstances affect their background and subject knowledge. Privileged learners have access to information from various sources such as the internet, television, magazines and books. They are therefore likely to have more comprehensive background and topic knowledge than learners form poor, rural environments. Test developers cannot assume that all learners share the same level of topic knowledge (McCabe, 2007:21). Test developers need to consider the differences in socio-cultural and educational backgrounds of South African learners when developing materials and selecting writing prompts and the input provided in examinations. The content should be accessible and familiar with the content provided pertaining to their cultures, customs and the areas where they live and go to school.

The NCS (2005:46) therefore specifies that teachers should select tasks and present content in a context that is familiar to learners. The Examination Guidelines (2009:5) stipulate that FAL learners should have a choice between eight different topics, and a minimum of two and a maximum of three should be visual stimuli. The various topics in the FET examination are open to interpretation and serve to accommodate learners from various backgrounds.

The current rating scale does not address the appropriateness of a response to the topic or content. This may present raters with problems when dealing with off-topic scripts.

Task setting and the linguistic demands of tasks are factors within the assessment instrument that influence learners' response in the context of assessment. However, the administration setting also influences test takers' performances in terms of context validity.

4.4.3 Administration setting

Administration setting concerns factors such as physical examination conditions and procedures, requirements, uniform administration and security. Circumstances under which the assessment is administered can influence the validity of the assessment (Weir, 2005:82; Shaw & Weir, 2007:133). A poorly administered test is likely to produce unreliable scores.

4.4.3.1 Physical conditions

Physical conditions such as background noise, lighting and air-conditioning should be controlled to ensure that test takers are comfortable.

South African learners write their matriculation examination under varying circumstances. Some learners write the examination in examination halls with good lighting, enough desks and chairs and a disciplined environment with limited noise. However, not all FET testing centres have adequate facilities and favourable testing circumstances. Many write in locations close to or next to highways, with poor lighting or no electricity and insufficient chairs and tables (cf. Blommaert et al., 2005). Such varying circumstances are not ideal because learners who write the examination in more challenging and disruptive circumstances may be affected negatively, causing them to under-perform.

4.4.3.2 Uniformity of administration

Test invigilators should have a set of clear and precise instructions for testing procedures and all administrators should adhere to these instructions. If a test centre breaks the uniformity rule, for example, by allowing more time, cognitive validity is threatened because the decisions learners make and their cognitive processes will be influenced by such changes (Shaw & Weir, 2007:139).

The administration procedure pertaining to the FET examination is well-established in South Africa. Administration at schools takes place under conditions that are specified in the *National policy on the conduct, administration and management of the assessment of the National Senior Certificate* (Subject Assessment Guidelines, 2008:5). Instructions regarding the structure of the examination are clear and transparent, as presented in the SAG (2008:18-22). The DoE presents the Examination Guidelines (2009) in conjunction with the National Curriculum Statement (2005) and the Subject Assessment Guidelines (SAG, 2008). They stipulate requirements regarding the setting of examinations in all the official languages in respect of the number of sections, length and type of texts, types and levels of questions, mark allocation and assessment rubrics. They therefore serve to standardise examinations (Examination Guidelines, 2009:3).

The examination papers are set by a national panel, assigned by the Department of Education (NCS, 2007:21). The panel comprises representatives of the various provincial educational boards in South Africa.

4.4.3.3 Security

In order to protect the content of assessments from being made public, security measures must be set. Access to the tests should be limited and items of secure tests should not be published or copied by any stakeholder. If security is breached, some learners may only be tested on their abilities to memorise pre-prepared answers.

Strict security measures are followed for the FET examination. Papers are sealed and transported to examination centres under strict security. A back-up paper is set and used if a leak should occur. All teachers are familiar with the security regulations as well as with the procedures during and after examination sessions.

4.5 Scoring validity

Elements that influence scoring validity include scoring criteria or the rating scale, rater characteristics, the rating process, rating conditions, rater training, post-examination judgement, grading of papers and awarding scores. Scoring of the FET examination performances takes placed in allocated venues where a large number of teachers come together for scoring. Each script is only scored by one rater, and ten percent of the scripts are moderated externally.

Scoring validity is of critical importance in the construct of writing (Shaw & Weir, 2007:1; cf. Chapter 5). As discussed in Chapter 3, the construct to be measured is a result of the symbiotic relationship between the underlying ability, the context in which the task is performed, and the scoring process (the "constructed triangle"; cf. Shaw & Weir, 2007:x). Context validity, cognitive validity and scoring validity contribute equally to overall construct validity. The construct therefore ultimately resides in the interaction of these dimensions (cf. Chapter 3).

A detailed discussion of the elements of scoring validity follows in the next chapter. This concept is considered as one aspect that contributes to overall validity, as opposed to the traditional view of reliability as separate from validity.

4.6 Criterion-related validity

Criterion-related validity is an external requirement established *a posteriori* in predominantly quantitative ways. The question is if – and to what degree – an assessment correlates with another criterion that measures the same construct as the instrument in question. The comparison is expressed in terms of a correlation coefficient. A correlation coefficient above 0.9 indicates a strong relationship between the instrument and the criterion. However, a single correlation coefficient with an external criterion is not sufficient evidence for criterion-related validity (Taylor, 2004a; Weir, 2005; cf. Chapter 5 for further discussion). Criterion-related validity concerns the following parameters: cross-test comparability, test equivalence and comparison with external standards.

4.6.1 Cross-test validity

Cross-test comparability centres on the degree to which scores of two different measurement instruments produce similar results for the same candidates at roughly the same time (Weir, 2005:207). However, Shaw and Weir (2007:229-230) echo the warning extended by Bachman, Davidson, Ryan, and Inn-Chull Choi (1995) that equivalent scores alone are insufficient evidence for comparable tests (also cf. Chapter 2). In addition to the scores, the equivalence of test content and performance must also be considered when instruments are being compared.

FET examination results are generally not compared to results from any particular measurement. Some tertiary institutions, courses or working environments may require additional assessment of particular skills which may be compared to matriculation results.

4.6.2 Test equivalence

Test equivalence refers to the relationship between different administrations of parallel forms of the same test conducted at different times. Both administrations have to be conducted for the same population and under similar conditions (Weir, 2005:208). The tests must be based on the same specifications and measure the same construct. The two administrations can be regarded as equivalent if they are equal in aspects such as raw score means, mean difficulty, variance and covariance when administered to the same person (UCLES, Multilingual Glossary of Language Testing terms, 1998; AERA, APA & NCME, 1999; Taylor, 2004a).

In addition to correlations, other statistical procedures that can be used to investigate test equivalence include ANOVA and multi-faceted Rasch measurements. Reliability can be investigated using conventional quantitative methods as well as verbal protocol analyses of expert raters' judgements (Weir, 2005; Bachman, 2004; Stemler, 2004; Shaw & Weir, 2007).

The DoE's mid-year examination could serve as a parallel test against which the results of the final examination could be compared, since the formats and requirements are similar. Such comparisons are currently not officially made to ensure scoring validity.

4.6.3 Comparison with external standards

Different assessment instruments and results can only be compared meaningfully against a shared reference criterion. Taylor (2004b) advocates the use of a comparative framework against which various assessments can be compared. Such a framework would be useful for a variety of test stakeholders. An external criterion should be used according to which assessment levels are benchmarked and descriptions standardised to present stakeholders with such a useful reference tool.

However, Shaw and Weir (2007:239) warn against oversimplifying and misinterpreting the framework and comparisons. They stress the need for evidence that the instrument addresses context, cognitive and scoring parameters of validity that are relevant to the level of language ability under the particular circumstance. Furthermore, the external criterion should not be used as a prescriptive tool, but as an informative one only.

4.7 Consequential validity

Consequential validity concerns the impact that an assessment has on stakeholders in educational and other contexts. According to Messick (1989:18), " [t]he questions are whether the potential and actual social consequences of test interpretations and use are not only supportive of the intended testing purposes, but at the same time are consistent with other social values ... social values and social consequences cannot be ignored in consideration of validity". Assessment should enhance teaching and learning practices and materials by providing information about learners' progress and needs. Outside the educational context, measures must be taken to ensure that any negative impact is not a direct result of error in the assessment instrument. Stakeholders must be able to trust that scores give an accurate indication of learners'

abilities and that they can confidently use the scores to guide their decisions concerning test-takers, such as whether to accept learners for further training or appoint learners in a particular position.

4.7.1 Washback

According to Shaw and Weir (2007:218), washback is considered to be part of the impact of assessment. Green (2003:6-8) regards washback as the effect that assessment has on individual participants, which mainly includes educators and test takers. In particular, the term "washback" refers to the effects of tests on teaching practices, for instance influences on teaching, teachers and learning (including curriculum and materials), whereas "impact" refers rather to the wider influences of tests on the community at large (McNamara, 2000; Hamp-Lyons, 2000; Shaw & Weir, 2007).

Washback may be intended or unintended and have positive or negative effects on the participants in terms of teaching or learning. Communicative tests should ideally have a positive effect on classroom practices, but tests are often experienced as negative by learners. Teachers can use the information from assessments to structure their practices and learners get an indication of which areas they still need to work on before mastering the skill. Detailed rating scales are useful for this purpose.

Alderson (2004:ix) points out, however, that washback is not merely as simple as whether a test has a positive or negative influence on teaching. It is a "hugely complex matter". Assessment may work as extrinsic motivation for learners to work harder, but it may also cause anxiety, which may cause learners to perform poorer than their actual ability. Teachers may also tend to teach to the test in fear of poor results (Alderson & Wall, 1993:115-129; Fulcher & Davidson, 2007:222-224). Hughes (2003:53-57) provides the following guidelines to achieve positive washback:

- The abilities being tested should be those you want to encourage;
- Sample widely and unpredictably;
- Use direct testing;
- Use criterion-referenced assessment;
- Make sure test takers and teachers are familiar with and understand the content;

- Assist teachers where necessary (by means of training and providing support materials).

Teachers may use test results to alter their teaching practices or focus in order to address learners' particular needs as identified through the assessment. This washback may entail teachers giving feedback to learners on their strong and weak points. The feedback, in turn, may help learners to improve their performance.

Schools that follow the national curriculum are required to provide feedback to parents on learners' progress using a 7-point scale. The level descriptions range from "not achieved" (0-29%) to "outstanding achievement" (80-100%) (SAG, 2008:6). The scale is not explicitly used as feedback tool in classrooms to make test takers aware of their strong and weak points. Learners' abilities are indicated by a single overall score. Owing to the lack of explicitly and distinctly defined criteria, the current rating scale is not very useful in providing detailed feedback to teachers and learners, which may inhibit the potential washback effect that could result from more detailed information. This conclusion is in line with the concern raised in the Organisation for Economic Co-operation and Development Forum (OECD) Reviews of National Policies for Education regarding the status of the South African curriculum, learning materials and examination practices in South Africa:

> Pass rates on external examinations provide only a partial measure of the quality of primary and secondary education. Nevertheless, they serve as "report cards" that are used by the education system and the general public as proxy indicators of quality. In their present state, however, the South African examinations cannot provide valid information about the performance of a particular school, teacher, or provincial department of education, because they do not reflect the powerful effects of non-educational factors like culture, family background, economic status, or urban/rural location. Value-added measurement has at least the potential of identifying the unique contribution of a particular school or teacher, by separating these contributions from other factors that affect learner performance (OECD, 2008:200).

4.7.2 Impact on institutions and society

Impact refers to the influence of assessment at social and institutional level (Hamp-Lyons, 1997:299). According to Weir (2005:214), it is difficult to measure the effects of assessment on the broader society. The effects are easily overlooked because developers are required to look further than the immediate stakeholders towards potential employers and educational institutions. Hamp-Lyons (1997:299) urges test developers to consider the consequences of

assessment on the broader society in addition to the washback towards individuals. She notes that assessment instruments should not only be evaluated with the test-setter in mind, but from the perspective of other stakeholders, including learners, teachers, parents, government and official bodies, and the marketplace.

Shaw and Weir (2007:38) note the value of impact studies (including washback studies) in ensuring ethical language testing practices. According to McNamara (1997:566), the "centrality of the notion of social identity to current work on language learning reflects renewed theoretical and political concern for the social dimension of language learning". Tests influence people's lives and can therefore potentially be used as power instruments to manipulate and control.

At its extreme, the critical language testing view (Shohamy, 2001) regards tests as a political tool of power and control, whether intentional or unintentional. Tests are by definition regarded as undemocratic, unethical and unfair. Followers of the critical language testing view, such as Shohamy (2001), see the impact of assessments as mainly constricting and prescriptive, used as disciplinary measure and to implement political agendas.

Ndaba (2005:2) is concerned about the focus on assessment as agent for political, economic and cultural change in South Africa. He comments on politicians' insufficient consideration of the social function and outcomes of assessment: "[D]ebate rarely addresses more fundamental issues concerning the social functions and outcomes of assessment". Politicians are in a position to use high-stakes assessment programmes to control what learners are taught and how they are taught. Such agendas are difficult to control.

The impact of the high-stakes FET examination is prominent in South Africa. Matriculation results are typically used as a measure to evaluate candidates for potential jobs, courses and tertiary education.

4.7.3 Avoidance of test bias

Before a test is administered, cognitive and context validity should be established to ensure that bias does not occur in the test. However, once the administration has been completed, developers must check whether any bias occurred in the assessment despite pre-test efforts. Various stakeholders could be consulted to evaluate the extent of the impact on stakeholders,

post-test performance and practices (Shaw & Weir, 2007). Based on stakeholders' feedback, assessment developers should revise the instrument to correct any bias and enhance the instrument's usefulness as teaching and learning tool.

As pointed out, learners have a choice between various topics on which to write an essay in the FET examination. As mentioned above in the discussion of context validity (cf. 4.4), this range of topics serves to provide learners from heterogeneous backgrounds with an equal opportunity to produce their best performance on a topic they feel comfortable with, thus avoiding bias towards any particular group. However, no investigation takes place after the examination to determine whether any bias occurred despite the selection of topics.

Performances are moderated to control potential bias of individual raters. FET examination scripts are moderated both internally and externally.

4.8 Conclusion

This chapter has discussed the parameters identified in Shaw and Weir's (2007) validation framework. The NCS (2005) requirements regarding writing were considered in relation to the parameters and the current writing rating scale was evaluated in terms of the framework.

It was established that the Subject Assessment Guidelines document of the NCS (2005) does not provide a clear-cut definition of the construct on which to base assessment instruments such as the rating scale. Nor are explicit guidelines and definitions provided pertaining to the criteria for assessing Section A of the Writing paper. The current scale does not always clearly and specifically address all features specified in the Learning Outcomes and Assessment Guidelines (2005). Based on the discussion in this chapter, it seems that the current rating scale used to assess writing in the final NSC examination is not an adequate measuring instrument for the purpose of the high-stakes examination.

An empirically validated rating scale will not solve all the problems that learners and the educational system in South Africa face. However, it may contribute to fairer and more accurate assessment in addition to providing explicit guidance and feedback for teaching and learning purposes. A valid rating scale must be supplemented by the adequate training of raters. These are issues related to scoring validity, which is considered in detail in the next chapter.

CHAPTER 5

Scoring Validity

5.1 Introduction

This chapter continues the discussion in the previous chapter on the parameters identified in Shaw and Weir's (2007) framework, but with a specific focus on scoring validity. Scoring validity forms the central focus of the present study, since it is concerned with the validation of a rating scale. It is therefore considered in more detail than the other parameters in the framework.

An assessment instrument is valid if it consistently produces accurate measurements, implying that it is then reliable (Hughes, 1989:42). Weir (2005:23) suggests that reliability should not be seen as separate from validity, but rather as an element that contributes to the overall validity of an instrument. He proposes the concept of *scoring validity* (cf. Chapter 3), which "concerns the extent to which test results are *stable over time, consistent in terms of the content sampling* and *free from bias*" (Weir's italics). Scoring validity is said to account for "the degree to which examination marks are free from error of measurement and therefore the extent to which they can be depended on for making decisions about the candidate" (Weir, 2005:23).

O'Sullivan (2006:184-185) warns against a simplistic view of reliability and suggests that reliability should be understood in terms of all factors that influence test scores. He therefore also regards the concept of scoring validity as a more viable option than the traditional understanding of reliability (cf. Chapter 2). Such a conceptualisation would improve reliability measurements considerably.

Alderson et al. (1995:105) note that without a reliable instrument, scoring procedure and trained raters, other attempts to ensure validity (e.g. choosing relevant and representative test tasks) are wasted. Various factors influence scores, such as the communicative ability of the test taker, test method facets, personal characteristics of test takers, the rating scale, and unsystematic factors that influence performance (Bachman, 1990:163-166; cf. also Weir, 2005). This chapter discusses each of these parameters, along with typical strategies for

controlling each factor. Practices implemented to ensure scoring validity in relation to the National Senior Certificate examination in South Africa are considered where relevant.

5.2 Rating scales

The first feature addressed in Shaw and Weir's (2007) framework in relation to scoring validity is the rating scale and assessment criteria. Two main aspects are important to consider when choosing or developing a rating scale. The first is the type of scale, which should be appropriate to the purpose of the assessment. The type of rating scale influences raters' approach to and interpretation of the text. The second feature is the format of the scale. Once the type of scale has been chosen, developers must consider which criteria best represent the construct for the assessment purpose, and the number of band levels to distinguish.

5.2.1 Types of scales

The rating scale represents the most concrete statement of the construct that is being measured. It is directly relevant to the validity of test results, as it "forms both our discussions of writing as a set of conventions, and it is used to assess the writing" (Inoue, 2005:220). Therefore, choosing the most appropriate type of scale is an important validation consideration and a major assessment decision (Hamp-Lyons, 1991b; Weigle, 2002; Barkaoui, 2007). According to Weigle (2002:108), this decision is critical because the scores are ultimately used to make decisions and inferences about learner abilities.

The type of scale influences the way raters assign scores. Barkaoui (2007), for example, investigated the degree to which rater severity depends on the type of rating scale used and the dimensions being assessed. He found significant rater-by-scale interaction effects, but it is not clear precisely how different rating scales affect raters and their scoring tendencies (Barkaoui, 2007:86; 103).

The construct, the purpose of assessment and the degree of task specificity should inform the choice of scale used for a particular purpose to ensure that a scale is suitable, and provides relevant, useful and accurate information about learners' abilities. In addition, the situation and context of administration may also influence the decision (Shaw, 2002:10; Weigle, 2002:72; 191).

Hudson (2005:208) regards the underlying nature of the scale as another fundamental issue that developers must consider when choosing the best type of scale for the purpose of assessment. Some scales indicate progression along a continuum, whereas other scales indicate a specific point or step that has been reached in the continuum. In addition, the type of comparisons (norm-referenced or criterion-referenced) that will be made with the scale is an important issue to consider.

Finally, examples of learner writing should be analysed to determine which features are typical of their writing at the relevant level. The development and evaluation of the content of the scale should be based on features that actually occur in learner performances (Upshur & Turner, 1995:208; Hudson, 2005:208).

Three forms of scales can be distinguished on the basis of the intended audience or primary user, namely user-oriented, constructor-oriented and assessor-oriented scales (Alderson, 1991; Council of Europe, 2001; Hawkey & Barker, 2004). User-oriented scales provide information about typical learner behaviour at different performance levels, including what learners are able to do, or sometimes cannot do, at different levels. Generally, these scales provide only one descriptor per level. These scales are constructed with potential employers in mind and are useful in circumstances beyond the educational environment.

Assessor-oriented scales are designed for raters scoring language performances and to guide the rating process. Statements describe how well learners perform and descriptions are often phrased in terms of what learners cannot do. "Assessment" refers to summative assessment of a specific performance. Assessor-oriented scales can be used to consider an overall performance or different aspects of a performance, such as fluency, accuracy and range. These scales describe typical performances of learners at different performance levels. This type of scale can be described as "diagnosis-oriented" because it can be used to describe current positions (e.g. current ability), identify target needs related to relevant categories, and accordingly diagnose which areas need to be covered or which skills need to be developed to reach the goal proficiency level (Council of Europe, 2001:38).

As the name suggests, constructor-oriented scales are designed to guide the construction of a test at an appropriate level. Constructor-oriented scales describe the tasks learners can do at

different performance levels. Statements describe the communication tasks that learners will have to be able to perform. These scales resemble lists of specifications and often describe what learners can do. Descriptors are holistic and can be broken down into one-word categories, such as 'description' or 'conversation'.

Scales are also distinguished on the basis of the raters' approach towards the text, in other words the text is considered as a whole, or as constituent parts contributing to a whole. Three main approaches to scoring are holistic (or impression) scoring, primary trait scoring (PTS), and multiple trait scoring (MTS). These three types of scales are characterised by two distinctive features. First, the scales are either intended for a single specific task, or can be generalised to a class of tasks. Second, the scales either report a single or multiple scores (Weigle, 2002:109).

Holistic scoring is also referred to as impressionistic scoring. It involves awarding a single score, based on raters' overall impression and subjective interpretation of the text (Davies, Brown, Elder, Lumley & McNamara, 1999:75). Specifications and performances are ranked in comparison to others of the same administration. A committee of experts usually first defines the construct based on their personal opinions (*a priori*), before writing definitions to describe manifestations of the construct at different predetermined levels (Fulcher & Davidson, 2007:96).

The major advantage of holistic scoring is that it is fast and therefore relatively inexpensive (Shaw, 2002a:11). Holistic scoring methods focus raters' attention on what learners can do well in their writing, as opposed to their weak points, and reward them accordingly. These scales provide information on the aspect of writing considered to be most important in the particular context (Huot, 1990:208; McNamara, 2000:41; cf. also Weigle, 2002).

However, holistic scales do not give very detailed information about learners' abilities. Holistic scoring is a rank ordering process and does not provide diagnostic information or specific feedback to learners and teachers. Nor does it account for learners who perform better on certain features or skills than on others. This is particularly problematic for second-language writers because all L2 learners do not develop all aspects of writing ability at the same rate (Weigle, 2002:114). Weir (2005:189) points out that holistic scoring fails in practice because learners who perform at different levels for different criteria are not

accounted for. Holistic scales seem to resemble levels of proficiency, but an acquisition order of different aspects of writing ability has not been established.

Owing to the *a priori* method typically used to develop holistic scales, these scales only provide a tentative indication of learners' abilities. Therefore, we cannot assume that the scales give an accurate indication of a learner's writing ability.

Weigle (2002:114) adds that holistic scales have been criticised for focusing on high inter-rater reliability at the expense of validity.

The holistic approach to assessment lacks a firm theoretical foundation. The general theoretical connections made in support of holistic scoring are not sufficient to claim that the procedure of holistic scoring is theoretically based. For example, holistic scoring assumes that relevant aspects of writing ability develop at the same rate and can be reflected in a single score. Raters cannot differentiate features such as depth and range of vocabulary, organisational aspects and syntactical control when they assign a single score to represent all these features. Shaw (2002b:11) warns that a single score may disguise an uneven writing profile and be misleading. Conclusions drawn about the validity of holistic scoring are therefore questionable (Huot, 1990; Shaw, 2002b).

Current research and development procedures place greater emphasis on the theoretical validity of assessment procedures, and holistic scales fall short in this regard. For this reason, Hamp-Lyons (1991b, 1995) suggests using analytic rating scales (discussed below) rather than holistic scales for assessing L2 writing.

Primary trait scoring is exclusively aimed at measuring one specific task (or at least task type) and cannot be generalised to other types of tasks (Fulcher & Davidson, 2007:97). It also involves allocating a single score, but the rating scale is only useful for one task or a specific task type. A primary trait scoring procedure typically involves a description of the task, a statement of the primary trait or construct being measured, a rating scale with level descriptors, samples of performances to illustrate each performance level, and explanations of why samples were scored the way they were (Weigle, 2002:110).

The benefit of a primary trait scoring procedure is that it requires a specific and clearly stated link between claims and evidence. Multiple experts have to agree on and specify those features that make one performance better than another. These features have to be specified

to guide other raters to come to the same conclusion about the level of a performance (Fulcher & Davidson, 2007:97).

However, the scores only give information about learners' abilities on a particular task. A separate scale containing a specific set of rating protocols has to be designed for each task. If a test comprises multiple tasks, raters have to score each task using a separate scale. This makes primary trait scoring very expensive, complex and time consuming. It is usually only implemented in research situations or situations where information about a specific skill of learners is needed. If raters have to score more than one task, they have to use different scales for each task, which makes scoring difficult. This makes primary trait scoring very expensive and time-consuming. Fulcher and Davidson (2007:97) note: "What we gain in explicitness and a stronger validity claim has to be offset against a reduction in the generalisability of the meaning we may wish to invest in the score".

Multiple trait scoring (MTS) is a popular means of assessing writing. Raters have to consider various features related to the construct individually and assign individual scores to each feature. It focuses raters' attention on evaluating the most salient features of a construct presented in a performance. Raters judge performances in terms of a number of particular criteria. They have to identify the extent to which a composition shows a number of different features that are regarded as important to a specific genre, function, or good writing as a whole. Fulcher and Davidson (2007:97) explain that each of the multiple scores assigned by raters represent "a separate claim about the relationship between the evidence and the multiple underlying constructs". A single combined score or multiple scores can be reported.

In some cases, raters may even be required to count the actual occurrences of particular features, such as errors pertaining to the features identified in the scale. This variation of multiple trait scoring is referred to as analytic scoring (Fulcher & Davidson, 2007:98; cf. also Cooper & Odell, 1977).

Multiple trait rating scales are popular means of assessing writing, but there are certain aspects that need to be considered regarding the construction and use these scales. For example, multiple trait scoring is more expensive and may take longer than holistic scoring. Criteria and scale level descriptors are also not always clearly phrased and distinctly defined, which may confuse raters and lead to inconsistent scoring. Bands may also provide an

oversimplified description of language proficiency, which may not accurately represent the nature of language proficiency (Spolsky, 1995:350-353; cf. also Alderson & Windeatt, 1991).

However, if the complexity of both writing and measuring writing ability is considered during the development and use of the scale, multiple trait scales offer many benefits (Spolsky, 1995:353). Tests results are easy to present and understand, which make this an attractive assessment option. The scales can also be used for a variety of writing tasks, as long as the test specifications are the same.

Multiple trait scales provide a common standard and meaning for judgments, which improves the reliability of scores in writing assessment. Focussing raters' attention on assessing the same features in performances ensures a reasonable degree of consistency and agreement between raters.

MTS scales have higher scoring validity when assessing L2 performances than holistic scales, because they provide raters with the opportunity to consider separate aspects in the performance that are related to the construct. The fact that raters assign multiple scores to a single performance in order to arrive at one combined overall score also improves the reliability of the final scores (McNamara, 2000:44).

These scales are functional because they provide diagnostic information about learners' abilities, which holistic scoring does not. Using multiple traits scores gives a more specific indication of learners' strong and weak points (Weir 2005:183). MTS thus considers and rewards learners who are more advanced in some aspects of an ability than in others, as is typical of L2 writers (McNamara, 2000:44; cf. also Shaw, 2002b).

In addition, these scales reflect what raters do when rating. Raters must be able to motivate the scores they assign. Multiple trait scales require raters to do so by assigning scores to various aspects of a construct. In this way, both rater bias and variability are controlled. In addition, multiple trait scales are useful tools for training and standardising examiners (Grabe & Kaplan 1996:405; Shaw, 2002b:11; Weir, 2005:190). Davies et al. (1999:126) note that if raters are well trained, MTS may improve the reliability of scores.

Different types of scales necessarily influence the outcomes of tests (Weigle, 2002). When choosing between holistic, primary trait or multiple trait scales, there are certain factors that one has to keep in mind, such as the reliability-time trade off and the purpose of the assessment. As mentioned above, primary trait scoring instruments are usually the most expensive and time-consuming to develop and use, followed by multiple trait scales. Holistic scoring instruments may be less expensive and sometimes easier to use, but are generally considered to be less reliable than primary trait or multiple trait scoring instruments.

Primary trait scoring is mostly used for research purposes, whereas holistic and multiple trait scales are generally used for general teaching practices. Holistic and multiple trait scales are accepted widely and valued in the field of writing assessment (Shaw & Weir, 2007:149). Therefore, Weigle (2002:72) recommends choosing between these two types of scales.

According to Carr (2000:228), the difference between holistic scoring and MTS can be summarised as one of focus: "holistic scores provide an assessment of a single construct, whereas composite scores from [multiple trait] rating scales conflate the information from several constructs". MTS is regarded as one of the main procedures for directly assessing writing ability and as a more valid measurement of L2 writing than holistic scoring (Hamp-Lyons, 1991b; Huot, 1990; Sasaki & Hirose, 2004).

Owing to the high-stakes nature of the matriculation examination in South Africa, a multiple trait rating scale seems to be the most suitable option for the assessment of writing. Holistic scoring remains more subjective and raters may easily be influenced by test takers' choice of topic or point of view. A primary trait scale may be too specific and restrictive, considering that the choice of tasks offered in the NSC examination also includes a choice in genre (cf. Chapter 4). Thus, the present study proposes to develop and validate a multiple trait rating scale for the assessment of Grade 12 writing performances.

As noted in Chapter 4, it is difficult to cater for the variety of individual test takers' needs in assessment situations. It is equally impossible to develop analytic scales with criteria that are suitable to and useful in all contexts for all populations, and that all raters interpret exactly in the same manner (Weir, 2005:185). At best, the scale criteria and descriptors should be clearly defined to support raters in scoring and increase reliability and validity. Important decisions include establishing the nature and number of criteria and levels, and defining

baseline performances students must produce in order to pass (Weir, 2005:187; cf. also Grabe & Kaplan, 1996).

Scales can be developed intuitively, by means of quantitative or qualitative procedures, or by using a combined approach (cf. Council of Europe, 2001; Fulcher, 2003; Hawkey & Barker, 2004). An intuitive approach relies exclusively on subjective expert judgements. Intuitively developed scales may sometimes be appropriate for assessment in low-stakes contexts with a familiar population, but should not be employed in high-stakes situations such as national assessment ones (North & Schneider, 1998:220). Holistic scales are typically developed intuitively by a panel of judges, although multiple trait scales can also be developed in this way (Fulcher & Davidson, 2007:96).

Quantitative approaches entail analyses to produce data-based, empirically-derived scales (cf. Upshur & Turner, 1995; Fulcher, 2003). The Council of Europe (2001:207) describes quantitative methods as those that entail quantifying "qualitatively pre-tested material". Quantitative data include, for example, test scores, responses to questionnaires and self-ratings. These data are analysed by means of statistical procedures, viz. correlations and Item Response Theory (IRT) models (such as Generalisability or Rasch procedures). An empirical approach entails examining actual samples of learner writing to identify the features that typify learners' writing at a particular grade level. These methods are usually applied when developing PTS, MTS and analytic scales.

Qualitative methods are those requiring "the intuitive preparation and selection of material and the interpretation of results" that requires intuitive interpretation of results (Council of Europe, 2001:207). According to Fulcher (1996:214), qualitative approaches are becoming more popular in test design because they lead to applied linguistic descriptions that make test scores more meaningful. Qualitative procedures include expert judgements, interviews and questionnaires, or assessments to provide data such as such as observations, verbal self-reports, or samples of language.

Qualitative interpretations can be used in addition to quantitative approaches, and can contribute to the process of theory development. Bachman (2004:6) notes that both quantitative and qualitative procedures should ideally be used to collect evidence to establish an instrument's usefulness and suitability for a particular assessment purpose and situation.

Criteria and band level descriptors can be developed by using *a priori* or *a posteriori* qualitative methods. A *priori* methods entail defining an ability first, based on an explicit theory of language and language use, and then describing the construct in terms of band levels and salient features. According to Bachman and Palmer (1996:212), this approach allows for making inferences about learner abilities based on an absolute scale (ranging from 'not mastered at all' to 'fully mastered'), rather than in comparison to other performances.

A posteriori methods entail empirical and statistical analyses to establish salient features of a construct. Empirical data are gathered from a sample of actual learner writing. Criteria and descriptors are then generated based on the data. In other words the salient features identified in the analysis of the sample indicate the most appropriate criteria. If criteria are identified based on actual scripts, the content of the scale can be developed to address features that occur in actual learner writing, as opposed to what developers may think should occur. This makes the scale more applicable in the particular assessment situation (Alderson et al., 1995; Weir, 2005; Hawkey & Barker, 2004).

Various researchers such as North and Schneider (1991), Alderson (1991), Fulcher (1987, 1993, 2003), Upshur and Turner (1995), Douglas (2001), Turner (2002), Weigle (2002), and Weir (2005) encourage empirical, data-based *a posteriori* procedures to develop valid rating scales, as opposed to an intuitive or *a priori* approach.

Weigle (2002:126) explains that the choice between an *a priori* approach and an empirical approach to scale descriptor development may be a philosophical question, regarding whether one wants to evaluate the degree to which a learner 'has' an ability, or the degree to which a learner 'can do' the ability. The "mastery approach", proposed by Bachman and Palmer (1996), is useful for making inferences about learners' underlying abilities, for instance whether a learner "has" ability X. According to Weigle (2002:126), if the construct is an inherent ability (a student 'has' ability X), the mastery approach may be a more appropriate choice. A pragmatic approach lends itself more toward predicting how learners will perform in future situations. Thus, if the construct is seen in terms of what a learner can, or will be able to do, without reference to the exact nature of the underlying ability, an empirical approach is a better option (Weigle, 2002:126).

As stated in Chapter 1, the current South African writing scale was developed intuitively by a panel of judges following an *a priori* approach. No empirical evidence is available to verify the validity of the scale. The FET Writing paper aims to evaluate whether learners are able to communicate effectively in writing for a wide variety of audiences and contexts (cf. Chapter 4). A pragmatic, empirical approach to scale development therefore seems to be a more appropriate option for the current project.

Another option is to follow a combination of *a priori* and *a posteriori* methods to enhance the development of the scale content. *A priori* methods alone may not be sufficient to ensure validity, but they enhance the meaning of statistical procedures. Weir (2005:18) emphasises the fact that statistical data in themselves do not provide conceptual labels. It is therefore necessary to circumscribe a construct in as much detail possible before trying to measure it in order to make statistical analyses more meaningful. "We can never escape from the need to define what is being measured, just as we are obliged to investigate how adequate a test is in operation" (Weir, 2005:18).

North and Schneider (1991, 1998) describe five methods of establishing criteria and generating descriptors empirically that comprise different degrees of combining quantitative and qualitative methods to draft and scale descriptors into a grid. The methods vary from mainly intuitive to a combined empirical approach which involves empirically selecting and calibrating level descriptors. North and Schneider advocate the latter option, because such a procedure should clearly indicate relevant criteria for the particular assessment purpose, and the number of band levels that should be distinguished.

The basic combination procedure for empirically deriving assessment criteria and descriptors involves using expert judgements to identify important criteria relevant to a particular assessment purpose. Experienced raters score a number of performances using the scale to determine the range of the levels that are necessary and will actually be used by raters when scoring performances. Including expert raters in the development or adjustment of a rating scale is crucial, because the raters have to accept the scale content as relevant and correct for it to be of value. This will ensure that they use and interpret the scale appropriately (Davidson, 1991; Hamp-Lyons, 1991b; Barkaoui, 2007).

Empirical data and statistical analyses may then be used to strengthen the empirical base of scale construction. For example, experts' descriptions of specific behaviours or typical features observed in written performances can be collected and analysed. The identified features are categorised and scaled down into more general criteria. Raters trial the scale after which the way they apply descriptors are analysed statistically. In addition, the level of difficulty that descriptors represent is determined and the scale is revised or adjusted accordingly (North & Schneider, 1991:222).

Both quantitative and qualitative procedures will be used in the present study to develop a rating scale. These include expert judgements, questionnaires and feedback reports, correlations and IRT model (Rasch) calibrations. In each phase, appropriate *a priori* and *a posteriori* validation procedures will be conducted, in accordance with the aim of the particular phase. A detailed discussion of the procedures involved in the present study follows in Chapters 6 and 7.

5.2.2 Criteria and band levels

Multiple trait scales comprise a number of levels that represent a range of scores. Each performance is assigned a level based on raters' assessment in terms of particular criteria relevant to the situation. According to Bachman (2002:471), developers should identify the aspects which best represent the construct to serve as criteria.

Multiple trait scales address a number of criteria, indicating the most important aspects of the construct to be assessed. Each criterion may be sub-divided into features related to the specific criterion. Band level descriptors are short statements formulated to describe the appearance of salient features (or sub-scales) addressed under criteria at each performance level (Underhill, 1987:98; McNamara, 2000:40; Hudson, 2005:207-208).

Different scales may be developed to assess performances at each level for a particular purpose, target group and context. Individual scales will not necessarily comprise the same criteria or distinguish the same number of levels, even though they may be designed to assess the same grade or proficiency level. These aspects are mainly determined by the construct of assessment. This section first considers the number of band levels that a scale should comprise, followed by the selection of criteria and, finally, the formulation of band level

descriptors to describe the manifestation of criteria at different performance levels distinguished in the scale in terms of the salient features.

Scales generally comprise between three and nine band levels. Various factors may influence the number of band levels, such as the purpose of the assessment, the range of performances expected from the test takers, and the background and experience of the raters (Bachman & Palmer, 1996; Weigle, 2002:123).

Alderson et al. (1995:111) recommend that a scale should not consist of more than seven points or band levels, as raters may struggle to distinguish more finely between the levels. Distinctions that are too fine may confuse raters, resulting in inter- or intra-rater variability. The full range of the scale should reflect the full range of learners' performances at that performance level. In discussing the common reference levels for the Common European Framework (CEF), the Council of Europe (2001:21) advises that "[t]he number of levels adopted should be adequate to show progression in different sectors, but, in any particular context, should not exceed the number of levels between which people are capable of making reasonably consistent distinctions".

Well-known and widely implemented scales such as the International English Language Testing System (IELTS) Writing Test (Fig. 5.1), Jacobs et al. (1981) (Fig. 5.2), Cambridge Main Suite Examination and Test of English for Educational Purposes (TEEP) Writing scale differ in the number of band levels distinguished. The IELTS scale distinguishes nine levels, whereas the Jacobs et al. (1981) distinguish four band levels. The various scales used for the Cambridge Main Suit Examination (Key English Test, Preliminary English Test, First Certificate in English, Certificate in Advanced English and Certificate of Proficiency in English) distinguish six different levels (Hawkey & Barker, 2004:131) and the TEEP Attribute Writing Scale (Weir, 1990) distinguishes four performance levels.

IELTS Band Scale

Band 9 – Expert User
Has fully operational command of the language: appropriate, accurate and fluent with complete understanding.

Band 8 – Very Good User
Has fully operational command of the language with only occasional unsystematic inaccuracies and inappropriacies. Misunderstandings may occur in unfamiliar situations. Handles complex detailed argumentation well.

Band 7 – Good User
Has operational command of the language though with occasional inaccuracies, inappropriacies and misunderstandings in some situations. Generally handles complex language well and understands detailed reasoning.

Band 6 – Competent User
Has generally effective command of the language despite some inaccuracies, inappropriacies and misunderstandings.
Can use fairly complex language, particularly in familiar situations.

Band 5 – Modest User
Has partial command of the language, coping with overall meaning in most situations, though is likely to make many mistakes. Should be able to handle basic communication in own field.

Band 4 – Limited User
Basic competence is limited to familiar situations. Have frequent problems in understanding and expression. Is not able to use complex language.

Band 3 – Extremely Limited User
Conveys and understands only general meaning in very familiar situations Frequent breakdowns in communication occur.

Band 2 – Intermittent User
No real communication is possible except for the most basic information using isolated words or short formulae in familiar situations and to meet immediate needs. Has great difficulty in understanding spoken and written English.

Band 1 – Non user
Essentially has no ability to use the language beyond possibly a few isolated words.

Figure 5.1 IELTS band scale level descriptors (IELTS, 2007:4)

Band levels can be used to regulate the weighting of criteria in assessment. As noted, they represent a range of scores rather than one specific scoring point. This is illustrated in Jacobs et al.'s (1981) scoring profile illustrated in Figure 5.2.

In this scale (Fig. 5.2), various weights are assigned to the different criteria. Each criterion is scored in terms of the four band levels, but the score ranges differ for each criterion. Scores assigned for content range from thirteen to thirty marks. This criterion thus represents a range of eighteen scores, divided across the four band levels. Organisation covers a range of fourteen marks from seven to twenty. Vocabulary covers a range of twenty-two from seven to twenty-eight. Finally, mechanics is weighted five marks ranging across the four band levels.

ESL COMPOSITION PROFILE			
STUDENT		DATE	TOPIC
SCORE	LEVEL	CRITERIA	
CONTENT	30 - 27	EXCELLENT TO VERY GOOD: knowledge; substantive; thorough development of thesis; relevant to assigned topic	
	26-22	GOOD TO AVERAGE: some knowledge of subject; adequate range; limited development of thesis; mostly relevant to topic, but lacks detail.	
	21-17	FAIR TO POOR: limited knowledge of subject; little substance; inadequate development of topic.	
	16-13	VERY POOR: does not show knowledge of subject; non-substantive; not pertinent; OR not enough to evaluate	
ORGANISATION	20-18	EXCELLENT TO VERY GOOD: fluent expression; ideas clearly stated/supported; succinct; well-organised; logical sequencing; cohesive.	
	17-14	GOOD TO AVERAGE: somewhat choppy; loosely organised but main idea stands out; limited support; logical but incomplete sequencing.	
	13-10	FAIR TO POOR: non-fluent; ideas confused or disconnected; lacks logical sequencing and development.	
	9-7	VERY POOR: does not communicate; no organisation; OR not enough to evaluate.	
VOCABULARY	20-28	EXCELLENT TO VERY GOOD: sophisticate range; effective word/idiom choice and usage; word form mastery; appropriate register.	
	17-14	GOOD TO AVERAGE: adequate range; occasional errors of word/idiom form, choice, usage but meaning not obscured.	
	13-10	FAIR TO POOR: limited range; frequent errors of words/idiom form, choice, usage; meaning confused or obscured.	
	9-7	VERY POOR: essentially translation; little knowledge of English vocabulary, idioms, word form; OR note enough to evaluate.	
LANGUAGE USE	25-22	EXCELLENT TO VERY GOOD: effective complex construction; few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions.	
	21-18	GOOD TO AVERAGE: effective but simple construction; minor problems in complex construction; several errors of agreement, tense, number, word order/ function, articles, pronouns, prepositions but meaning seldom obscured.	
	17-11	FAIR TO POOR: major problems in simple/complex constructions; frequent errors of negation, agreement, tense, number, word order/ function, articles, pronouns, prepositions and/or fragments, run-ons, deletions; meaning confused or obscured.	
	10-5	VERY POOR: virtually no mastery of sentence construction rules; dominated by errors; does not communicate; OR not enough to valuate.	
MECHANICS	5	EXCELLENT TO VERY GOOD: demonstrates mastery of conventions; few errors of spelling, punctuation, capitalisation, paragraphing.	
	4	GOOD TO AVERAGE: occasional errors of spelling, punctuation, capitalisation, paragraphing.	
	3	FAIR TO POOR: frequent errors of spelling, punctuation, capitalisation, paragraphing; poor handwriting; meaning confused or obscured	
	2	VERY POOR: no mastery of conventions; dominated by errors of spelling, punctuation, capitalisation, paragraphing, handwriting legible; OR not enough to evaluate.	

Figure 5.2 Jacobs et al.'s (1981) scoring profile, illustrating five criteria with varying weights and four band levels

Both McNamara (2000:42) and Luoma (2004:80) note that choosing the number of band levels is a matter of what is practical rather than of what is theoretically valid. "There is no point in proliferating descriptions outside the range of ability of interest. Having too few distinctions within the range of such ability is also frustrating, and the revision of rating scales often involves the creation of more distinctions" (McNamara, 2000:42). Distinctions can also be made within a particular level. The IELTS writing examination scale, for

example, now recognises half-levels, allowing raters to distinguish more precisely between "upper-" and "lower-level" performances within level.

Scale developers should work towards a balanced scale that gives adequate feedback for both teachers and learners, while being as practical as possible. Weigle (2002:127) considers empirical procedures such as pre-testing or piloting a scale as the best means of determining exactly how many levels to distinguish in a scale.

A seven-point scale is recommended for assessing writing and report scores for Grade 12 certification (SAG, 2008:5). The current two-dimensional grid used to score FET examination performances distinguishes these levels. Level 1 indicates the lowest score range and Level 7 the highest score range (cf. Appendix A). The scale levels are not equally distributed. Level 1 represents a range of fourteen marks (0-29%), while Levels 2 and 6 each represents a range of five marks (10%). Level 3 represents only 4 marks (8%), while Levels 5 and 6 each represents a range of six marks (12%). Level 7 represents a range of ten marks (20%).

The rationale for the unequal distribution of score ranges in the current SA rating scale is not clear. The large ranges represented by the first and seventh levels in particular make it difficult to distinguish accurately between very poor and very good performances. Van der Walt (2009: personal communication), an external moderator at the time the scale was originally devised, indicated in an interview that the cut-off points were determined intuitively without any empirical evidence to support the decisions (cf. Chapter 1). For the present study, scores will be equally distributed across scale levels (cf. Chapter 7).

It is widely recognised that explicitly stated and clearly described criteria are one of the key factors in successful scale development and application (Weir, 1990, 2005; Hamp-Lyons, 1990; Alderson, 1991; Bachman & Palmer, 1996; Brown, Glasswell and Harland, 2004; Elder, 2005). Clear criteria and band level descriptors in scales are also critically important to the validity of an assessment since the rating scale either explicitly or implicitly represents the theoretical basis of the assessment (Weir, 2005:109).

Weir (1990:67-68) notes that developers have paid too little attention to assessment criteria for judging written performances in the past, with "too much room for idiosyncratic

interpretation" of what constitutes the criterion to be applied. He attributes the fact that many marking schemes fail in practice to the poor choice and delineation of criteria.

Criteria should help raters focus on those features of writing that are important to realise the relevant communicative function for the purpose of the assessment. Writing may typically be assessed in terms of, for example, topic development, organisation, and accuracy and range pertaining to vocabulary and grammatical features. Criteria should guide raters' judgments on learner performances. Poorly-focused criteria and ill-phrased descriptors (discussed below) may contribute to rater variability and may also slow down the scoring process (Weir, 2005:180-198).

According to McNamara (1996:121, 2000:38), clearly specifying criteria is central in defining the assessment construct. It illustrates the notions of the skills and abilities being assessed. Scales reflect those skills or abilities that raters want to measure and should do so explicitly. A scoring instrument that does not specifically reflect the construct will result in less reliable scores (cf. McNamara, 1996; Weir, 2005). Therefore, the construct determines both the nature and number of criteria addressed in a scale.

It is important for the eventual validity of the instrument that the content of the scale must represent the construct being assessed (thus avoiding construct irrelevant variance or construct over-representation) (cf. Weir, 1993, 2005; Bachman, 2002). In order to ensure scoring validity, scale developers need to reach consensus on the criteria and related salient features that must be assessed, and what they signify (Turner, 2000:2).

There is no consensus on the perfect number of criteria that should be addressed in a scale. The construct definition and purpose of assessment generally indicate which criteria are most important to consider during scoring. Scales should consist of enough criteria to provide a detailed picture of learners' abilities, while being practical to use. Luoma (2004:80) suggests five to six criteria as an appropriate maximum number, since criteria can remain conceptually independent. Seven criteria are regarded as the psychological limit (Pollit, 1991; Council of Europe, 2001; Luoma, 2004; Weir, 2005).

Typical scales used in practice include the Jacobs et al. (1981) scale, the Cambridge Main Suite, International English Language Testing System (IELTS) Writing Assessment scale and

the TEEP Attribute Writing Scales (Weir, 1983). These scales are used in different contexts and for different purposes for assessing writing. They accordingly distinguish different criteria. Table 5.1 summarises the criteria distinguished in each of these scales.

Scale	Number of criteria	Nature of criteria
Jacobs et al. (1981)	5	• content, • organisation • vocabulary • language use • mechanics
Cambridge Main Suite Examination	4	• fulfilment of the task set • communicative command of the target language • organisation of discourse • linguistic errors
International English Language Testing System (IELTS) Writing Assessment scale	4	• task achievement/response • coherence and cohesion • lexical resource • grammatical range and accuracy
TEEP Attribute Writing Scales (Weir, 1983)	7	• accuracy • fluency • interaction • coherence and organisation • task fulfilment • language control and linguistic range • communicative effectiveness • register

Table 5.1 Summary of criteria distinguished in four current scales widely used in practice (Hawkey & Barker, 2004:123)

Standards of writing are communicated through the criteria, which therefore need to be comprehensive and based on empirical evidence from actual sample scripts (Weir, 1990:68). Bachman (1990:36) stresses that band level descriptors must be precise to help raters distinguish levels and assign the most appropriate score. Therefore, descriptors should be phrased so that they are distinct from each other and clearly communicate the differences between levels.

Weigle (2002:125) suggests that descriptors should not be phrased in terms of each other, for example, ones such as "poorer than...", "better than level 2" or "very good". Such descriptors may be difficult to distinguish precisely and consistently, particularly by inexperienced raters. The descriptions must be unambiguous and give raters a specific indication of how criteria are manifested (in terms of salient features) at each performance level.

In addition, exemplar scripts should be used to clarify the meaning of criteria and descriptors and illustrate the relevant standard and performance level and. Criteria that are not exemplified by actual performances of writing at the appropriate level can be interpreted at different levels of proficiency (Wolf, 1995:76). Exemplar scripts help to illustrate the standard of writing at each performance level in order to clarify band level descriptors, providing raters with a more concrete conception of how they should apply the scale (Weir, 1990:68).

Table 5.2 and Table 5.3 present examples of scale level descriptors for two widely accepted scales. The Common European Framework of Reference for Language (CEF; developed by Council of Europe, 2001) distinguishes six scales for different proficiency levels. Table 5.2 provides examples of the descriptors for the general proficiency levels representing two of the six scales, viz. levels C1 and A2 (Hudson, 2005:217).

Proficient User	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic, and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors, and cohesive devices.
...	...	....
Basic User	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need

Table 5.2 Common European framework – global scale (Council of Europe, 2001)

These are broad level descriptors that serve as framework for developing instruments to assess learners at different proficiency levels. They describe the overall level at which learners are to be assessed. Individual scales can be developed accordingly to determine the degree to which learners have achieved, for example, a C1 level or an A1 level.

Table 5.3 presents an excerpt from the IELTS scale illustrating the descriptors formulated to explain the four criteria at the two highest levels of achievement (Bands 8 and 9).

Band	Task achievement	Coherence and cohesion	Lexical resource	Grammatical range and accuracy
9	• fully satisfies all the requirements of the task • clearly presents a fully developed response	• uses cohesion in such a way that it attracts no attention • skilfully manages paragraphing	• uses a wide range of vocabulary with very natural and sophisticated control of lexical features; rare minor errors occur only as 'slips'	• uses a wide range of structures with full flexibility and accuracy; rare minor errors occur only as 'slips'
8	• covers all requirements of the task sufficiently • presents, highlights and illustrates key features/ bullet points clearly and appropriately	• sequences information and ideas logically • manages all aspects of cohesion well • uses paragraphing sufficiently and appropriately	• uses a wide range of vocabulary fluently and flexibly to convey precise meanings • skilfully uses uncommon lexical items but there may be occasional inaccuracies in word choice and collocation • produces rare errors in spelling and/or word formation	• uses a wide range of structures • the majority of sentences are error-free • makes only very occasional errors or inappropriacies

Table 5.3 IELTS writing band level descriptors for bands 8 and 9

Each level is described in terms of a number of bulleted points. Raters should consider each of these points to determine the degree to which a performance meets the requirements of that level for a particular criterion. Both the labels and the salient features need to be clear to raters in order to ensure that raters know how to apply the scale and do so consistently. The IELTS examination is widely used to assess second language writing. However, raters may – intentionally or unintentionally – focus more on one of the features listed than on the others in order to place a performance. For example, a learner may demonstrate "fully satisfies all the requirements of the task", but may not succeed in presenting a "fully developed response". Such scenarios may complicate scoring for raters.

5.3 Rater characteristics

Accurate scoring of written performances is especially important in large-scale national assessment contexts, although this is not easy to achieve. Writing assessment has long been "plagued by concerns about the reliability of rating", which is, in fact, a concern about the reliability of raters (Hamp-Lyons, 2007:1; cf. also Huot, 1990; Brown et al., 2004). Schaefer (2008:466) points out that "essay rating is a complex and error-prone cognitive process which introduces systematic variance in performance ratings". Rater variability has been a concern in language testing for many years and, as mentioned above, various studies have reported considerable inter- and intra-rater differences (e.g. Huot, 1990; Engelhard, 1994; Lumley, Lynch & McNamara, 1994; Myford & Wolfe, 2003).

Although rater variance is traditionally regarded as 'error' (Engelhard, 1994), not all research regards rater variance as such (Weigle, 1994; Wolfe, Kao & Ranney, 1998; Kondo-Brown, 2002). Rather, raters are recognised as people with values and beliefs. Hamp-Lyons (2007:2) suggests that test developers and validators should try to use these values as best possible to assesses the effectiveness of written performances without constraining raters too much. In so doing, raters may score writing more validly by expressing the 'true value' of the composition. However, raters' personal characteristics may influence scores negatively and should be controlled.

Some raters may consistently or randomly be more severe or more lenient in their scoring than others. According to McNamara (1996:121), three aspects that may influence rater variance during scoring are the candidate, the task/s and the rater. The candidate in this case refers to the relative abilities of a learner, in other words how their performance on one task can be correlated with performance on another task. When learners are given a choice of tasks, the tasks are never exactly the same. Different test takers will find different tasks more or less difficult than the other. Finally, individual raters may differ from other raters in four ways:

- Some raters may overall be more lenient than others;
- Some raters may be biased against certain tasks or certain groups of learners;
- Raters may vary in their degree of intra-rater consistency;
- Individual raters may interpret and apply the rating scale in different ways (McNamara, 1996:123-125).

Rater characteristics is a crucial influential factor in how raters score writing (Shaw & Weir, 2007:168). Barkaoui (2007:86), for example, investigates the effects of different types of scales on raters' behaviour and reports that "raters were the main source of variability in terms of scores and decision-making behaviour". Alderson (1993:47) argues that "so-called professional judgements are frequently flawed", or in serious conflict with other professional judgements, and urges that these professional judgements must be corroborated and validated. However, little information is available on what the exact effects of raters' characteristics are on scores.

Characteristics such as professional, linguistic and cultural background, extent of training and scoring experience, gender, the amount of exposure to L2 writing, and external pressures have been found to influence rater variability (cf. for example Hamp-Lyons, 1990; Elder, 1993; Weigle, 1994; Kondo-Brown, 2002).

Lumley (2002) emphasises that it is the rater, not the scale, who is at the centre of the scoring process. "It is the rater who decides which features of the scale to pay attention to, how to arbitrate between the inevitable conflicts in scale wording; and how to justify her impression of the text in terms of the institutional requirements represented by the scale and rater training" (Lumley, 2002:22-23). Shaw and Weir (2007:168) identify two experiential characteristics as particularly influential, viz. language experience, and the effect of professional experience. Language background influences raters' behaviour and values. Raters tend to be more sympathetic in scoring L2 written performances that demonstrate rhetorical patterns similar or identical to their own than performance that contain clearly different rhetorical patterns (Kobayashi & Rinnert, 1996; Grabe & Kaplan, 1996).

Raters' professional background and training also influence how they go about scoring and evaluating individual criteria. Teachers with similar experiences and backgrounds tend to have similar views about the nature of L2 proficiency and are likely to base their judgements on the same construct of writing (Erdosy, 2004:57; Barkaoui, 2007:102).

Raters' expectations have an influence on the severity with which they score performances. For example, raters may expect more from compositions written by honours level students than from first-year learners and therefore score the same performance more strictly at honours level than at first-year level. Raters may have higher expectation of, for example, electronic performances than hand-written versions and therefore be harsher in scoring the first than the latter. Therefore, a hand-written version of an essay may receive higher marks than a computerised version of the same essay, because raters have higher expectations of formatting, grammatical and spelling accuracy for word-processed compositions (Weigle, 1999:145).

Despite the variety of influential measures that may cause measurement error, research (e.g. Jacobs et al., 1981; Alderson, 1991; Hamp-Lyons, 1991b; North & Schneider, 1998; Weigle, 1994) shows that the reliability of writing assessment can be improved by controlling rater

variance through a combination of factors, viz. using appropriate scales with clear scoring criteria (cf. 5.1), standardising scoring procedures (cf. 5.4) and training raters (cf. 5.6) to interpret and apply scales consistently (Brown et al., 2004:106; Elders, 2005:176; Shaw & Weir, 2007:147). Alderson et al. (1995:188) urge developers to use as many ways as possible to enhance scoring validity and ensure fair and accurate results.

Grade 12 English FAL teachers score the Writing paper performances for the National Senior Certificate examination. They come from different socio-cultural, language, educational and professional backgrounds. Raters' scoring may be influenced by any of these factors that distance the writer and the reader.

As pointed out in Chapter 4, raters and test takers often do not share the same background, which may cause raters to interpret performances differently than intended by the test takers. Different cultures may, for example, prefer different paragraph organising strategies or implement rhetorical devices differently. Raters who do not share the test taker's cultural background may be less familiar with other cultures' preferences in this regard and penalise the learner – possibly unfairly so. The average Grade 12 learner in South Africa is not highly proficient in English FAL (cf. Chapter 4). Learners with lower ability levels may be more likely to revert to their L1 resources and writing strategies in order to try to express themselves. Raters may therefore be faced with different writing conventions to what they are used, which may influence the scores they assign to a performance.

5.4 Rating process

Different scoring procedures and raters' individual scoring strategies may also affect inter-rater consistency (Lumley & McNamara, 1995; Lumley, 2000; O'Sullivan, 2000). The rating process determines how scales are applied to performances. It addresses issues such as whether all tasks are scored using the same scale, whether all criteria are applicable to all tasks, or whether raters should report separate scores for different sections and/or features. Luoma (2004:171) suggests that these practical questions help to clarify criteria and may help developers identify potential problem areas regarding the application of the scale. Hamp-Lyons (1991b:242) argues that raters must pay conscious attention to all the essential elements of good writing to establish a reasonable balance when scoring performances. She

notes that a detailed scoring procedure which requires raters to consider the multi-dimensionality of ESL writing is essential to establish this balance.

A well-established, systematic scoring process is regarded as key to valid assessment (Brown et al., 2004:106; Elders, 2005:176). An established and standardised rating procedure can help raters make decisions in a more systematic way and assign scores consistently, thus increasing inter- and intra-rater reliability.

Establishing a rating procedure entails specifying the role of the rating scale and specifying how scores should be assigned. It also involves establishing moderating procedures to help raters remain consistent in their interpretation and application of the scale. Monitoring during scoring procedures and scoring by at least two persons can help stabilise an individual rater's marking (Huot 1996:559).

According to Weir (2005:192), it is essential to consider how assessment criteria can best be applied to performances once appropriate tasks have been selected and criteria established. He emphasises that "tasks cannot be considered separately from the criteria that might be applied to the performance they result in".

A complicating matter is the fact that individual raters may not always use the same internal processing to make decisions while scoring. Since the rating process is concerned with how raters apply scales and report scores, raters' internal behaviour during scoring has a major impact on results. Various studies (e.g. Cumming, 1990; Huot, 1990; Weigle, 1994; Wolfe et al., 1998; Cumming, Kantor & Powers, 2001; 2002; Lumley, 2002; Schoonen, 2005) provide insight into how raters use scales as guidance, raters' internal criteria and individual expectations of compositions that influence their scoring. It is important for scoring validity that test administrators manage the procedures effectively and help raters remain consistent in their decision-making (Lumley, 2002:247).

Examining the cognitive decision-making process of raters when they assign scores may provide insight into their behaviour and reasons for their varying in the scores they assign. However, the scoring process is still fairly unexplored, and too little is known about what goes on in raters' minds while scoring, about the effects of rater training, and the value of standardisation (O'Sullivan, 2006:186). Huot (1990:258), Lumley (2002:246) and O'Sullivan

(2006:186) comment on how little is known about how raters arrive at their decisions or the role that scoring procedures play in the reading and rating process. The process is "still not well understood", as Lumley (2002:246) points out, despite numerous studies investigating the issue (cf. Cumming, 1990; Huot, 1990; Weigle, 1994; Wolfe et al., 1998; Cumming, et al., 2001, 2002; Lumley, 2002; Schoonen, 2005).

Wolfe et al. (1998) found that more proficient raters tend to follow a read-then-evaluate approach, compared to less proficient raters' more iterative reading-refer-to-scale-evaluate-read-refer-evaluate approach. According to Lumley (2002) and Barkaoui (2007), raters generally seem to follow scoring procedures that are fundamentally similar. However, although they seem to share a general understanding of the scale content, they do not always apply the scale content in the same way. Raters' expectations may also influence their overall judgements during scoring (Shaw & Weir, 2007:172).

Scoring procedures is a complex issue (cf. Lumley, 2002). Raters not only have to interact with the performance, but also with the task. There is an assumption that raters find the best fit, in other words they decide which scale descriptor best matches the text, based on a common interpretation of the scale content. Contrary to this assumption, Lumley (2002) reports that raters seem to base their judgements on an indefinable feeling about the text, as opposed to basing their decisions on the actual features and descriptors in the scale content. Raters form a "uniquely complex impression independently of the scale wording" but ultimately manage to refer to the scale in each instance (Lumley 2002:263). Raters seem to use the scale descriptors as a tool to articulate and justify their scoring decisions, rather than offering descriptions of the text.

Raters, who are at the centre of the scoring process, use the scale as "a set of negotiated principles that the raters use as a basis for reliable action, rather than a valid description of language performance" (Lumley, 2002:268). Scales and rules for scoring provide guidelines for raters on dealing with variety in performances, but they do not cover all possible problems that raters may encounter in performances. Raters are therefore forced to develop their own strategies to deal with problematic situations in the rating process.

Raters resolve tension between the scale and their own impression by means of an intermediate process to reach a compromise between their own opinion and the scale

descriptors. Lumley (2002:246) argues that "despite this tension and indeterminacy, rating can succeed in yielding consistent scores provided raters are supported by adequate training, with additional guidelines to assist them in dealing with problems".

One way of controlling such tension, Huot (1996:559) suggests, is by basing assessment practices in a specific context. This will force raters to make judgements that are practical, pedagogical, programmatic and interpretative. Raters are also spared the effort of having to define abstract values such as the quality of writing.

Raters' internal scoring behaviour can be investigated by means of verbal protocol analyses, which require raters to describe their thoughts and procedures while scoring. A verbal protocol analysis entails raters providing a verbal report of how they go about scoring a performance. Raters are trained to verbalise their thoughts while performing a complex cognitive activity such as scoring (Bachman, 2004:276; cf. also Chapter 6).

According to Bachman (2004:276), verbal protocol procedures are the most commonly used indirect means of empirically investigating internal procedures related to taking tests and scoring performances. Weir (2005:244) also considers qualitative research such as verbal protocol analyses as a potential rich source of information on the processes involved, as the recordings provide insight into underlying mental processes of raters. Verbal report measures are considered to be "more illuminative with regard to strategy" (Weir, 2005:244; cf. also Green, 1998) than other methods.

Verbal protocols can be collected either during or directly after scoring. Concurrent reports may be more likely to be complete and, perhaps, accurate than retrospective ones, but raters may find concurrent reports intrusive. Alternatively, verbal reports can be recorded directly after scoring, as successfully achieved by Anderson, Bachman, Perkins and Cohen (1991). Alderson et al. (1995:176-177) suggest collecting the reports directly after raters have finished scoring the performances. Memory prompts such as the test scripts can then be used to help raters provide detailed reports. Raters may find this method less intrusive than simultaneously reporting their thoughts and scoring. In order to avoid bias as result of any single research method, different methods and sources – such as questionnaires, interviews and verbal reports – can be triangulated. However, interview or questionnaire questions

should lead the respondent as little as possible to a particular answer (Urquhart & Weir, 1998; Phakiti, 2003; Weir, 2005).

Although verbal reports may be strongly subjective in nature, Weigle (1999:149) notes that protocol analysis can provide valuable data if the protocols are collected and analysed systematically, driven by theory, and substantiated by other evidence. Bachman (2004:278) also emphasises the usefulness and necessity of this procedure in investigating test performances:

> The use of verbal protocol analysis has provided many valuable insights into the ways in which test takers [and raters] process different kinds of language test tasks [test performances]. I thus consider this methodology to be an indispensable tool for collecting information about test performance as part of the test try-out phase before tests [or scales] are used operationally to make decisions.

However, in some cases raters think they do one thing, while actually and unintentionally doing another. Verbal protocol analyses will not reveal such cases. McNamara (1996:216) suggests using more sophisticated methods such as multi-faceted Rasch measurement (MFRM) (cf. 5.7) as a useful tool to investigate rater behaviour and reveal underlying patterns in scoring data and fundamental questions of validity. Various studies (e.g. McNamara, 1996; Lumley, 2002; Kondo-Brown, 2002) have used MFRM to investigate rater variability. For example, McNamara (1996) used MFRM to investigate rater behaviour in scoring responses to tasks with a communicative focus. He found that raters were significantly influenced by grammatical accuracy, despite the fact that grammar was downplayed, and that the raters themselves were not aware of this bias.

A well-established scoring process is in place for marking the NSC examination performances. Each script is scored by an individual examiner who reports a single score for each section in the writing paper. Scores for the different sections are added to make up the total score out of one hundred. In order to reach a score for the essay writing paper, raters have to determine the level of performance in terms of language and content, and then read a combined score from the scale (cf. Appendix A). Raters select a score for content on the horizontal scale read in rows from left to right, and a score for language on the vertical scale read in downward columns. The point where the two scales cross indicates the overall score

that raters report on the answer sheet. The assessment is moderated externally before results are processed and announced (cf. Chapter 4).

5.5 Rating conditions

The conditions – be they temporal, physical and/or psychological – in which performances are scored may also influence the validity of the scores. Various external factors (such as heat/cold, noise and time limits) and internal factors (such as emotions and physical well-being) may influence the consistency with which raters score. For example, if raters are irritated by noisy conditions, they may tend to score more harshly than necessary or than they normally would in a quieter setting. Weir (2005:200) compares scoring in a shady tranquil landscape to scoring in the busy London tubes during rush hour. Raters are likely to score more consistently in favourable conditions than in uncomfortable settings.

Raters who score the performances for the final NSC examination may be exposed to different scoring conditions and environments, in the same way that test takers are exposed to different examination conditions and environments when writing the examination (viz. noise, poor lighting; cf. Chapter 4). FET examination examiners, for example, score *en mass*, which may influence individual raters in different ways. Scoring circumstances may therefore contribute to or cause anxiety for some raters.

5.6 Rater training

As noted above (cf. 5.2), training raters is considered an important means to improve scoring validity. Shaw (2004a:3) regards sound judgment as fundamental to valid and successful assessments. Training raters helps them to make more consistent and accurate judgements about performances. Trained raters who use a valid scale with clear descriptions are more likely to make sounder judgements than untrained ones using the same scale. Therefore, training is crucial to the validity of language testing.

Weir (2005:198) notes that improving inter- and intra-rater consistency involves developing appropriate rating scales and standardising the rating to these scales. However, Brindley (1998:65) warns not to take it for granted that raters will interpret a scale similarly, even if the scale is clear, valid and the scoring process well-managed. Shaw and Weir (2007:147)

argue that definitions of performance levels are not only conveyed through the written criteria and descriptors, but also through a process of training and standardising scoring procedures.

Thus, detailed rating scales alone are not sufficient to guarantee reliable scoring. Raters must receive adequate training to ensure that they interpret the scale content correctly so as to score consistently and allocate reliable scores. Training helps raters develop a sense for the "institutionally sanctioned interpretations" of task requirements and scale features and how other raters relate their personal impressions of a text to the scale (Lumley, 2002:267).

Rater training is an ongoing process with raters being trained and re-trained to familiarise them with scoring criteria and clarify any confusion regarding the interpretation or application of criteria. Trainees must also demonstrate the extent to which they arrive at a common understanding and application of criteria through rating a number of performances (Knoch, Read & Von Randow, 2007:27).

The aim of training is generally regarded as getting raters to agree with each other. Wolfe et al. (1998:485) suggest that training, particularly in large-scale scoring situations, helps to create a group of like-minded raters who focus on similar features of writing when making decisions about scores. Groups of people who think alike about certain features are more likely to assign consistent scores. White (1985) argues that establishing a community of readers is important for the successful assessment of writing. Training raters is not "indoctrination into standards determined by those who know best", but rather the establishment of a harmonised community "that feels a sense of ownership of the standards and the process" (White, 1985:69). Lumley (2002:246) also notes that rating requires some restricting procedures such as training and guidelines in order to produce reliable measurement.

During training, scoring criteria have to be clarified and discussed extensively. Anchor papers or sample scripts must be considered to demonstrate features as they appear in learner writing (Barkaoui, 2007:104; cf. also Davidson, 1991; Erdosy, 2004). Training usually involves a series of scoring and discussion sessions. The content of the scale is explained to the raters, who then score a number of scripts as a group and compare their results. Differences and difficulties are discussed, followed by more individual scoring sessions and group discussions. The final scoring session is an individual scoring session after which the results

are compared to determine the degree to which raters assign similar scores. The scripts used during training sessions are either chosen to exemplify a particular level or levels of the scale, or to illustrate specific problematic situations that may arise during scoring (Elder, 2005:176).

Various effects – such as severity effect, halo effect, central tendency effect, inconsistency, and the bias effect – may cause raters to mark either too leniently, too harshly or inconsistently. Mulqueen and Baker (2002:4), for example, note that the choice of topic may influence raters' consistency. They explain that some topics may be more difficult for raters to score than others. Raters' experience also affects the accuracy with which they score. Weigle (1999) compared the ratings of experienced and inexperienced raters on two different prompts (a choice essay and an interpretation of a graph or "graph essay"). Results showed that the inexperienced raters rated more severely than the experienced raters and particularly so for a graph essay. Rater effects such as these must be controlled in order to achieve reliable ratings. The most important way to address these issues is through rater training.

Various studies report mixed results regarding the effects of training on rater variance. Lumley (2002) reports that, despite explicit training and guidelines for dealing with certain problematic features, some raters still struggled with aspects that were not directly addressed by the scale. "Components of training sessions may go unheeded, partially unheeded, or may take on proportions unintended by the trainer" (Lumley, 2002:267). Eckes (2005) reports strong differences in severity after training in a study about rater effects in scoring writing and speaking. Raters displayed inconsistencies in their application of rating scale criteria. Kondo-Brown (2002) also found that raters differed in the severity with which they scored certain candidates and criteria. Furthermore, each rater showed a different bias pattern.

However, distinctly positive effects have resulted from training raters. Hamp-Lyons (2007:1) notes that rigorous rater training has "consistently shown to be pretty successful". Alderson et al. (1995:105), Lumley (2002:248) and Knoch et al. (2007:27) also note that training can help to eliminate extreme differences in rater severity, increase inter- and intra-rater consistency and reduce individual bias. Weigle (1999) found that, after training, the differences between raters' severity (mentioned above) were no longer detectable to the same degree, although the graph essay was still rated more severely by inexperienced raters. Elder (2005) reports that after receiving feedback on their scoring, raters' levels of severity remained statistically significant, but the degree to which the raters differed in their severity decreased

significantly. Raters scored more consistently, although only slightly. Raters also responded positively to the feedback they received during training.

Training assists in making raters aware of their personal scoring tendencies. They can then adjust their marks accordingly, in terms of the characteristics of test takers and the task demands (Elder, 2005:176; 183). Results from Lumley's (2002) study confirm the positive impact that training has in creating a like-minded community of raters that follows a similar process during scoring. Lumley (2002:22-23) states that training "almost certainly" forges common understanding, interpretation and agreement amongst raters. Clarifying criteria during training helps raters to gain a better consensual understanding of the scale levels and terminology used, which leads to higher intra-rater and inter-rater consistency (Weigle, 1994:249).

Thus, although training may not always eliminate all aspects of rater variance, it does have a positive impact on rater variance as it increases agreement amongst raters. Hamp-Lyons (2007) calls for more information on how training and standardisation procedures are managed by administrators and test developers. She suggests that a better understanding of what happens during standardisation sessions might help using the sessions more appropriately to shape raters' decisions without restricting or undermining their skills and perceptions. Her findings show that raters do not share the same discourse when discussing performances. They follow a wide variety of approaches when scoring performances, and sometimes have strong internal expectations which may differ from the scale criteria (Hamp-Lyons, 2007:2-4). These differences could be addressed more effectively during training if more information were available on the matter.

Knoch et al. (2007:27) emphasise the importance of ongoing training, particularly since training may not always be equally effective for all individual raters. It is therefore not guaranteed that training will eliminate variability one hundred percent. Computerised self-training programmes can, for example, be used to help raters retrain off-site.

FET examiners in South Africa usually receive standardisation training before scoring the final matriculation examination. Raters are retrained annually.

5.7 Post-exam adjustment

Despite using clearly defined descriptors and training raters, researchers and testers recognise that scoring written performances involves an element of subjectivity and that the judgements of raters for the same performance may often vary (Schaefer, 2008:466). McNamara (1996:127) suggests accepting rater variability as an element to be compensated for. Adjustments can be made after scoring has taken place to balance the effect of raters.

One way of compensation is using multiple raters to score each performance and then averaging their scores. However, this method is time-consuming and expensive. Schaefer (2008:470) advocates using modern statistical methods to control rater variability instead of, or in addition to, averaging multiple scores.

Modern statistical measures offer a more sophisticated means of conducting post-exam adjustments (Lumley & McNamara, 1995:57; Kondo-Brown, 2002:4). Particularly in large scale assessment, various statistical methods can be used to investigate inter- and intra-rater agreement. Three types of statistical methods distinguished are consensus estimates, consistency estimates and measurement estimates. Each of these categories offers a number of procedures that can be used to base post-exam adjustments on.

Consensus estimates demonstrate the degree to which raters give similar scores to a performance. Once high consensus has been reached, scores from different raters can be averaged. These estimates are used when raters are trained to use a scale assumed to represent a linear continuum of progression in a construct. Consensus estimates are easy to calculate and they easily identify rater disagreement. These measures indicate the exact percentage of adjacent agreement within one scoring level above or below. However, individual calculations have to be made for each pair of raters and for each task or performance (Stemler, 2004; Brown et al., 2004).

Consensus estimates include statistical percent agreement such as Cohen's Kappa (1960). This statistic indicates the "degree to which consensus agreement ratings vary from the rate expected by chance" (Brown et al., 2004:106). Procedures such as the multi-rater version of the Kappa procedure can be used to investigate how strongly raters agree with one another in

terms of the scores they assign to learner performances. This procedure averages the proportion of agreement between raters and then adjusts for any chance agreements.

Consistency estimates provide an indication of the general agreement between raters (inter-rater reliability). According to Brown et al. (2004:106-107), these estimates do not indicate whether raters give the same scores, but rather indicate whether raters give high or low scores in a similar pattern. Consistency estimates have the functional purpose of getting raters to apply a rating scale consistently. These measures are useful for determining the exact degree of agreement between raters for training purposes. They show whether there is a noticeable pattern of assigning high/low scores across a number of raters. Consistency measures can be used to determine the extent to which one factor overlaps or correlates with another. For example, one can determine whether two different tests actually measure the same ability.

Consistency measures also serve as means to sum scores for each performance. However, some statistical adjustments for differences in rater severity have to be made before the scores can be summarised. Different raters may have different mean scores. Calculating necessary adjustments to rater severity may take effort, but without such adjustments, the validity of final results may be threatened. In addition, if little variance between raters is calculated the correlation value will decrease, which may falsely give the impression of poor consistency (Brown et al., 2004:107; Stemler, 2004).

Pearson's coefficient (r), Spearman's rank coefficient, and Cronbach's alpha are typical consistency measures. Pearson correlation coefficient can be used to investigate pairs of raters, while Spearman's rank coefficient is useful for unevenly distributed data. Cronbach's alpha is useful to investigate multiple raters, indicating the degree of consistency between them (cf. Chapter 6).

Measurement estimates are considered to be the strongest indicator of inter-rater agreement (Shavelson & Webb, 1991; Brown et al., 2004). Rater severity is measured independently on a linear measure. One statistic is provided, so rater severity on all and any items can be compared directly. Measurement estimates incorporate as much information as possible about each rater to incorporate into a model. Differences in rater severity are considered at individual and group level.

Measurement estimates are useful to investigate the degree to which scores can be attributed to consistent scoring rather than error (Stemler, 2004; Brown et al, 2004). These measures identify variance in scores as related to tasks, raters, errors or interaction components. They can be used to identify rater bias and inconsistencies empirically and identify raters that may need retraining. Measurement estimates can therefore make assessing writing performances fairer and more equitable (Wigglesworth, 1993; Knoch et al., 2007; Schaefer, 2008).

Special computer software is required to calculate measurement estimates, however, and the structure of data files for analyses is initially complicated and counterintuitive to set up (Brown et al., 2004; Stemler, 2004).

Typical measurement estimates include procedures such as Generalisability theory (G-theory) and multi-faceted Rasch measurement (MFRM). G-theory is used to investigate the extent to which results of one exercise, using a sample of a population, can be trusted to be true for the larger population as much as it is true for the sample. Results will indicate the extent to which raters interpret and use a scale in the same way. For example, G-theory can be used to investigate the reliability of the rating procedure applied by a single rater can (Shaw, 2004b:5) (cf. Chapter 6).

MFRM is an application of Item Response Theory (IRT). It is a logistic latent trait model of probabilities that calibrates different facets independently of each other, within a common frame of reference. All facets are measured on one logit scale. Weigle (1999:148) explains the underlying theory of the model:

> The theory underlying this measurement model is that the probability of a certain performance being given a particular score ... can be seen as a function of the examinee's ability and several other facets of the scoring situation, such as the severity of the rater, the difficulty of the task, and the threshold of difficulty between the points on the scale ... By estimating these other facets it is possible to obtain a more accurate estimate of the examinee's ability in the skill being tested.

Thus, different facets, viz. task difficulty, test takers' ability and rater severity, can be compared to one another on a logit scale. MFRM is valuable in performance assessment and rating scale validation because it can analyse other facets besides task difficulty and ability. For example, the interaction between raters and scales, as well as person-item interaction can be measured (McNamara, 1996:121; Schaefer, 2008:466).

MFRF allows for differences in rater severity, inter-rater variability and variation in item difficulty (Linacre, 2006a). Schaefer (2008:470) notes that MFRM has an advantage above other analysis methods for investigating rater inconsistency and subjectivity because raters can be investigated individually or as part of a group. MFRM analysis is useful for identifying individual outlier cases (misfits). Investigating these individual cases may provide information on problems related to interpreting the marking scheme. MFRM can be used to investigate raters' tendencies to be biased and score performances higher or lower than is warranted. Such a so-called "bias analysis" reveals systematic sub-patterns of raters interacting with particular items, criteria or other aspects of the rating situation (Wigglesworth, 1993:309). Bias analyses can provide researchers with a better understanding of what causes rater bias, which will contribute to improved rater training and scale development (Schaefer, 2008:466).

Owing to the fact that MFRM provides specific information about individual elements of different facets, each rater can receive specific feedback about their performances. In addition, trainers can get a better idea of which problem areas to focus on during training, and how to train or retrain raters to help them use a scale more consistently. Increasing consistency through bias analysis therefore also helps to reduce measurement error. Reducing measurement error increases fairness and accuracy in norm-referenced assessment such as summative evaluation (Wigglesworth, 1993; Schaefer, 2008).

According to Engelhard (1992:98), MFRM improves the fairness of writing assessment, because raw scores alone may be an under- or over-rated view of the performance due to different degrees of rater severity. MFRM adjusts for rater variability and therefore gives a more accurate indication of test takers' abilities.

It is important to note, however, that measures such as MFRM should be used in combination with rater training. Although MFRM accepts that raters judge independently from each other, they are still expected to judge from the same point of view. MFRM indicates whether this condition is adhered to or not (Linacre, 1998:630) (cf. Chapter 6 for further discussion on MFRM).

Post-exam adjustment procedures for the FET examination involve external moderation and statistical adjustments. Group averages are compared to those of previous years and scores, since examination papers are never the same two years in a row. In order to retain consistent standards, group averages are adjusted, using statistical measures (Van der Walt, 2009: personal communication). Information about the statistical adjustment, viz. measurements used and procedures followed, are confidential and not available to the public or for research purposes.

5.8 Grading and awarding

Following scoring and post-exam adjustments, papers are graded and a final score is awarded. The final scores are recorded and reported to the learners and other stakeholders. These scores are then used as indication of learners' abilities. The decisions made based on these scores have certain consequences for learners, teachers and other stakeholders. After an assessment has been administered, the consequential validity of the assessment must be investigated.

The announcement of the final NSC matriculation results is a major event in South Africa. Results are published in the major newspapers across the country. Learners receive report cards with their individual final scores for each subject and a matriculation certificate to confirm their achievements. Potential employers and tertiary institutions usually use the reported scores and matriculation certificate as indication of their abilities to perform certain skills for a variety of purposes.

5.9 Summary

The discussion in this chapter and the previous one identified the need for revising the current rating scale, based on various points of criticism which indicate that the scale is not valid in the present context and for the purpose for which it is being used. The following points present the rationale for an empirically derived and validated rating scale to assessing writing for the FET examination:

- The scale was designed and developed intuitively by a committee. No empirical data are available to support claims that the scale is valid.
- It was constructed over a decade ago and has not been revised to suit the current assessment context. It should not be assumed that the current scale is appropriate for

assessing the new NSC examination, particularly considering the changes in the educational system over the last decade and a half.

- Only two criteria are addressed, namely Language and Content. Two criteria do not adequately represent a complex construct such as second language writing. The current scale therefore does not provide an adequate and comprehensive indication of learners' writing abilities.
- Not all features specified in the National Curriculum Statement (2005) and Subject Assessment Guidelines (2008) are addressed under the criteria and are therefore neglected in scoring.
- The criteria are not conceptually distinct or clearly defined. Language is addressed in terms of features such as grammatical accuracy, punctuation, vocabulary, style, tone, register and length, whereas content is addressed in terms of features such as critical awareness of the impact of language, evidence of planning and reference to organisation.
- The current scale does not provide comprehensive and detailed diagnostic information about learners' writing abilities. Since it provides limited feedback in terms of learners' strong and weak points, it is not an effective tool for enhancing teaching and classroom practices.
- Some raters find the scale difficult to implement consistently, because the scale makes it difficult to reward learners who perform better in some features than in others related to a particular criterion. Raters may therefore equate performances that are not at the same level because each level is described in terms of a group of features. All features have to be considered, but raters cannot score the features individually.
- In practice, the current scale does not guide raters to discriminate effectively between performances. Raters tend not to use the full range of the scale, but rather to bunch marks around 50 – 55%. This may be owing to the fact that the scale level descriptors are not based on examples of learner performances, that raters have to account simultaneously for various features at each level, or because levels are not clearly distinguished.
- Although statistical measures are used in addition to moderators to control rater variability and make post-exam adjustments, it is unclear which statistical procedures are used for additional adjustments across the group to standardise results from year to year. Statistical procedures such as MFRM should be used to investigate and adjust for rater variability and bias, but it is not clear whether post-exam adjustment procedures include such investigation.

5.10 Conclusion

The focus of this chapter on scoring validity is central to the present project, which aims to develop an empirically derived and validated rating scale. Scoring validity concerns all elements that influence how scores are assigned, such as the type of rating scale and the criteria addressed, rater characteristics, rating procedure and situation, the training of raters, post-exam adjustments and grading and awarding.

This chapter supports the need for empirically validating a new scale for assessing the Writing paper of the NSC examination. It provides guidelines on how to develop a rating scale for the current project. The current rating scale threatens scoring validity, owing to the fact that it does not represent the construct and does not provide clear and distinct definitions of criteria or level descriptors. The discussion of these features serves to inform the development of the scale for the current project.

The following chapter describes the methodology followed to establish an empirically validated rating scale to assess writing in the South African context for the purpose of the FET examination.

CHAPTER 6

Method of Research

6.1 Introduction

This chapter provides a summary of the method of research followed in this study to develop and validate an empirically derived rating scale for assessing the writing ability of Grade 12 ESL learners. An overview of the empirical process, which consisted of four phases, is provided here. Each of the four phases is briefly summarised in terms of its aim, the participants involved, procedure followed, analyses made, and the final outcome of the particular phase.

6.2 Phase 1: Benchmarking exercise

6.2.1 Aim

The aim of Phase 1 was to benchmark examples of typical learner performances at each of the seven scale levels.

6.2.2 Participants

Fifteen participants were involved in the first phase along with the researcher, viz. fourteen expert and experienced raters from four provinces and a former Umalusi external moderator. The fourteen raters had an average of nineteen years experience as markers in the Grade 12 ESL examination, ranging from ten to thirty years. They included markers, deputy chief markers, chief markers as well as internal moderators. The former external moderator had 16 years' experience as moderator for Grade 12 ESL papers.

The panel were familiar with the current rating scale, had a thorough knowledge of the context in which assessment takes place, shared similar views on the construct that was assessed, and could be expected to allocate reliable and valid scores.

6.2.3 Procedure

In 2006, a random sample of 200 compositions written by Grade 12 ESL learners during the final matriculation examination was collected by the researcher with the help of an Umalusi external moderator. The essays had been sent in from each province for moderation. The necessary permission to use the essays for research purposes was obtained.

The compositions were typed on computer in their original form, with all errors retained. The typing was checked for accuracy. A sample of sixty-eight compositions was then selected from the 200 essays. (After analysis, the number was reduced to sixty-four. This is discussed in the next chapter.) They were selected according to the original mark allocated to them to represent all seven writing levels. They were sent to the fourteen raters, who scored the essays at home, using the current rating scale. Scoring took place in two intervals of four weeks each, viz. during October and November, 2006, and March and April, 2007. Each rater scored at least thirty two compositions, and each composition was scored by at least nine of the fourteen raters. Marks were assigned for the language and content criteria, and a combined score was also allocated. The external moderator moderated the marking to ensure that it was valid and reliable. The data were computerised and coded for statistical analysis.

6.2.4 Analysis

Data were analysed using the FACETS version of the multi-faceted Rasch program (Linacre, 2006b) to investigate the following:

- the scoring consistency among the fourteen raters;
- the degree to which the sample of essays represented the full range of abilities on the scale;
- item difficulty and rater bias towards either of the criteria;
- the accuracy of the levels at which essays were benchmarked by the raters.

As indicated in Chapter 5, multi-faceted measurement procedures such as Rasch may provide a more accurate indication of learner abilities than raw scores alone, because rater characteristics are taken into account (McNamara, 1996:118). According to Lumley and Brown (2005:830), "[p]erhaps the most significant achievement of Rasch analysis has been

conclusive documentation of the many ways in which rater behaviour can vary, as well as to identify some of the kinds of measures (such as training and multiple rating) that can be taken to assist in managing this variation".

McNamara (1996:9) notes that multi-faceted Rasch analysis is an increasingly valuable means of investigating performance-based assessment. It is useful because it accounts for rater variance and provides an accurate indication of learner abilities (cf. Chapter 5.7; Linacre, 2006a). It uses information on different facets in the data matrix to investigate and predict the interaction between these facets. If the facets are *learner ability*, *rater characteristics* and *item difficulty*, it can, for example, predict the likely score for a particular combination of these facets.

The accuracy of the prediction can also be evaluated. It is possible to describe learners' ability in terms of their chances of succeeding at a particular level on an item or task, given the information about item difficulty and rater severity supplied by Rasch (McNamara, 1996:134). McNamara (1996:133) explains that, to the extent that the data conform or "fit" the Rasch model, "it is possible to bring all the facets together into a single relationship, expressed in terms of the effect they are likely to have on a candidate's chance of success. That is, we can see precisely what sort of challenge the candidate was facing on that item with that rater, and are accordingly able to interpret the actual rating given".

The idea of probabilities of responses is central to Rasch procedures. The analysis extrapolates from a data set by considering the probabilities of patterns of responses in the data to arrive at general estimates of test takers' abilities and item difficulty. For example, if we know how a test taker performed on other items (ability estimate), and how other candidates performed on a particular item in question (difficulty estimate), it is possible to make a prediction about how likely it is that the test taker will perform in a certain way on a particular item. The probability that a learner with higher ability will answer a question correctly or perform better in a task is higher than for a test taker with lower ability. According to McNamara (1996:133), "the model states that the likelihood of a particular rating on an item for a particular rater for a particular candidate can be predicted mathematically from the ability of the candidate, the difficulty of the item and the severity of the rater".

McNamara (1996:161) points out that Rasch uses a mathematical procedure called maximum likelihood estimation. This estimation procedure or "calibration" is recursive and repeats until the desired level of accuracy of prediction is met. Once this level is reached, the calculation stops or "converges" and the ability and difficulty estimates or "measures" are reported (McNamara, 1996:161-162). FACETS (Linacre, 2006b) generates a number of useful reports, viz. a vertical ruler report, facet measurement reports providing so-called "fit" statistics, reliability indices and a report on unexpected results.

Vertical ruler reports are summary maps that provide information about the interaction of different facets (cf. for example Fig. 7.2 in Chapter 7). For example, scores can be seen as a result of the interaction between three facets, viz. learner ability, task difficulty and rater characteristics.

FACETS also provides measurement reports with detailed information on each individual facet. Table 6.1 provides an example of an extract from a facets measurement report. In this table, information is provided on the *essay* facet (cf. Chapter 7).

A measurement report such as the one in Table 6.1 identifies test takers (in the columns on the far right, in this case represented by essay numbers 1 to 64) and gives an estimate of their ability, expressed in logits.

Essay Measurement Report (arranged by fN).

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Nu	Essay
86	18	4.8	4.68	2.42	.28	3.00	4.1	3.03	4.1	-1.09	34	34
47	16	2.9	3.08	.01	.30	2.37	2.9	2.37	2.9	-.34	43	43
73	18	4.1	3.97	1.35	.29	2.23	2.8	2.23	2.8	-.19	24	24
95	18	5.3	5.18	3.14	.28	2.21	2.9	2.20	2.9	-.36	19	19
51	16	3.2	3.32	.38	.30	1.98	2.2	1.97	2.2	.13	50	50
40	16	2.5	2.64	-.62	.30	.25	-3.3	.25	-3.3	1.89	44	44
Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Nu	Essay
56.4	17.0	3.3	3.32	.32	.32	.94	-.3	.93	-.3			Mean (Count: 64)
25.2	1.0	1.5	1.49	2.32	.07	.54	1.5	.54	1.5			S.D. (Populn)
25.4	1.0	1.5	1.50	2.34	.07	.55	1.6	.55	1.5			S.D. (Sample)

Model, Populn: RMSE .33 Adj (True) S.D. 2.29 Separation 6.92 Reliability .98
Model, Sample: RMSE .33 Adj (True) S.D. 2.31 Separation 6.98 Reliability .98
Model, Fixed (all same) chi-square: 2883.5 d.f.: 63 significance (probability): .00
Model, Random (normal) chi-square: 61.5 d.f.: 62 significance (probability): .49

Table 6.1 Extract from the essay measurement report produced by FACETS

On a dichotomous test, for example, test takers placed above the 0 logit mark are more likely to answer a particular question correctly (i.e. higher ability), while those below the 0 mark are less likely (i.e. lower ability) (McNamara, 1996:136). Since ability estimates reported by Rasch are extrapolations of data, they are subject to error. Measurement reports therefore also indicate the likely error involved in each ability estimate, i.e. standard error measurement (Model S.E. in Table 6.1).

Facet measurement reports provide detailed information on the interaction between different facets in the form of "fit" statistics (Infit and Outfit mean square indicated in Table 6.1). Fit statistics indicate the degree to which the data fit the Rasch model. If the pattern of the data does not fit the probabilistic model, the data are identified as misfitting (i.e. overfit). McNamara (1996:132-133) provides a more technical explanation:

> We can attempt to model the data in the data matrix – to make it predictable. To the extent that this is possible, the data fit the model we are using; to the extent that it is difficult to do this for individual items, candidates or raters, this will also become clear; we have identified data-model mismatch or misfit. The model in this case is known as a multi-faceted Rasch model.

A particular interaction or result may occur more or less frequently than expected (e.g. more learners of lower ability get an item correct than expected) (McNamara, 1996:171). Fit values are expressed in terms of *t*-distribution, varying around a mean of 0 (McNamara, 1996:173). Depending on whether the observed values vary more or less than expected, *t*-values will be positive or negative. Both McNamara (1996:143, 173-174) and Bachman (2004:147) indicate that values greater than +2 and smaller than -2 signify particular bias or misfit. These guidelines were used in the present study.

Fit statistics can for example be used to identify unsuitable test items or assessment criteria, or to determine the degree to which criteria and/or salient features addressed in a scale measure the same construct (Lumley & Brown, 2005:830). Misfitting items may indicate that test items or scale criteria are poorly written or do not discriminate well between learners with different abilities. Alternatively, they could indicate that the item or criterion is good in itself, but that it does not measure the same ability or construct as the other items in the test/scale (McNamara, 1996:175). Items identified as misfitting can be examined and modified or deleted from the test or scale if necessary.

At the bottom of a measurement report table (Table 6.1) FACETS provides summarises of various estimates, including a reliability index. This index indicates the extent to which a test or scale, in this case, defines different levels of ability or its capacity to distinguish between performances (McNamara, 1996:138). McNamara (1996:138-139) emphasises that this reliability is not the same kind as that traditionally reported for performance assessment, which indicates the degree of agreement between raters. It is more like indices of reliability, such as Kudar-Richardson 21 or Cronbach's alpha, generally associated with tests with dichotomously scored items.

The Rasch reliability index is scaled from 0 to 1, with values closer to 1 indicating good reliability. According to McNamara (1996:140), high reliability indices indicate real differences between raters in the actual levels of scores assigned to them. This reliability index typically indicates substantial differences between raters, as can be expected since, despite training, raters are never exactly the same in their scoring. McNamara (1996:140) notes that eliminating all differences in rater severity is an unachievable goal, and perhaps an undesirable one (cf. Chapter 5).

6.2.5 Outcome

A set of benchmarked compositions that illustrate typical examples of writing at each of the seven scale levels was produced at the end of Phase 1. These essays were used during the development and validation of the rating scale to assess English FAL writing.

6.3 Phase 2: Drafting a rating scale

6.3.1 Aim

The aim of Phase 2 was to draft a new rating scale. In order to achieve this, salient features of writing (criteria) had to be identified and described in terms of band level descriptors. Results were calibrated to investigate the appropriateness of the selected features and the degree to which they represented the construct in question.

6.3.2 Participants

A panel of experts in the drafting of rating scales took part in this phase. The seven-member panel consisted of five experienced academics and teachers, a former Umalusi external moderator and the researcher. The participants were selected on the basis of their experience and expertise in the fields of ESL assessment, L2 writing and scale development. They came from three different L1 backgrounds and had an average of twenty five years' experience in teaching and eighteen years' experience in scoring English Second Language at matriculation and first-year university level.

6.3.3 Procedure

As in Phase 1, Phase 2 involved quantitative and qualitative methods. It involved two exercises, namely a workshop (qualitative procedure) and a calibration exercise (quantitative analysis).

A two-day workshop was held on 23 and 24 May 2008. The participants were provided with background reading as well as copies of the benchmarked essays before the workshop, and they were requested to study and analyse the essays.

During the workshop, the panel determined which type of scale would be most appropriate to assess the multi-faceted construct of writing for the purpose and in the context of the final FET examination. Once the type of scale was determined, the panel analysed the benchmarked compositions to determine which features were salient in learner writing and distinguished performances at different proficiency levels. After these analyses, the panel constructed a first draft of the scale and formulated descriptors to describe each salient feature addressed under each criterion. This draft scale was then revised and refined twice during the workshop.

A calibration exercise followed four weeks after the workshop to establish the accuracy and relevance of the criteria and level descriptors with regard to the construct. Five members of the panel blindly scored ten (unlabelled) benchmarked essays. The external moderator and the researcher were not involved in the marking.

6.3.4 Analysis

During the workshop, the panel analysed the benchmarked compositions. They evaluated the scripts, reporting and comparing individual findings and consolidating differences through discussion.

The calibration exercise served as statistical validation of the criteria and salient features as drafted in the scale by the panel. In addition to providing information on inter-rater consistency, scale calibration was conducted to determine whether the criteria and salient features included in the scale were relevant and accurately represented the construct of writing in the final FET examination. Linacre's (2006) FACETS Rasch program was used to analyse the scores assigned by the experts. The researcher calibrated the data to investigate the following:

- the scoring consistency among the five raters using the draft scale;
- the degree to which the sample of essays represented the full range of abilities on the scale;
- item difficulty and rater bias towards either of the criteria.
- the degree to which the fifteen features measure the same construct;
- the degree to which the features represent the construct of assessment.

6.3.5 Outcome

Phase 2 produced a draft scale in the form of a seven-point semantic differential multiple-trait scale totalling 100 marks. In addition, the panel compiled an explanatory scale guide or key in which they clearly defined and illustrated the criteria and salient features with examples from actual scripts.

Results from the quantitative validation in the form of the calibration were used to support the qualitative decisions regarding the content and format of the draft scale.

6.4 Phase 3: Refinement of the scale

6.4.1 Aim

The aim of the third phase was to trial the draft scale and to identify possible weaknesses in it.

6.4.2 Participants

A different panel was used to trial and refine the draft scale in Phase 3. The twelve-member panel comprised the researcher, the former external moderator and ten experienced Grade 12 ESL teachers. The teachers came from different language and cultural backgrounds and different education and teaching environments. They had between nineteen and thirty years' marking experience, with an average of twenty-one years. They were all qualified as ESL teachers. The group represented a range of schools from well-performing, privileged schools to previously disadvantaged and underprivileged schools.

6.4.3 Procedure

Phase 3 entailed both quantitative and qualitative procedures.

A two-day workshop was held on 31 September and 1 October 2008, during which the draft scale was trialled and refined. The workshop started with a blind-marking session during which the teachers scored two essays, using the draft scale. The results for the blind scoring were calibrated and the results compared to the calibration conducted at the end of Phase 2.

Following the blind scoring exercise, the group discussed the content of the scale and refined the salient features and the bi-polar descriptors describing these features.

Once the participants had reached consensus on all problematic aspects, they individually scored five essays using the revised draft. The researcher then conducted a second calibration exercise, using the revised draft scale.

Finally, verbal protocols and written feedback reports were collected on each individual's experience of rating with the draft scale. This was done directly after the final scoring session.

6.4.4 Analysis

A qualitative evaluation of the draft scale was conducted by the panel. They also considered the draft in terms of the degree to which it adhered to and addressed the specifications set out in the Assessment Standards in the NCS (2005:32-45) and in the SAG (2008).

Data in the form of scores were calibrated by the researcher, using Linacre's (2006) FACETS Rasch program. The aspects investigated in Phase 2 were again addressed (cf. 6.3.4). The results were compared to those of the previous calibration, viz. those at the end of Phase 2.

Finally, the written reports of the participants were analysed by the researcher to determine their opinions of the revised scale. The researcher compared these results with those of the calibration exercises as additional validation data.

6.4.5 Outcome

Phase 3 resulted in a revised draft scale. The format of the original draft scale was retained, but some features were slightly revised and the descriptors were refined.

6.5 Phase 4: Trialling of the scale

6.5.1 Aim

The aim of Phase 4 was to test the scale through an initial pilot implementation.

6.5.2 Participants

Twenty experienced teachers and the researcher were involved in the piloting of the revised draft scale. Six of the participants were familiar with the scale, having been instrumental in revising and refining the original draft scale in Phase 3. The additional participants included experienced National Senior Certificate examiners representative of the population of examiners. The participants' experience ranged from five to thirty years' teaching, with an average of seventeen years experience between them. All participants were qualified ESL teachers. They came from various L1 and socio-cultural backgrounds, and taught at schools ranging from previously disadvantaged to privileged ones.

6.5.3 Procedure

The scale was piloted during a two-day workshop conducted by the researcher on 17 and 18 April 2009. It involved trial standardising of raters through training and "certification" scoring. A combination of quantitative and qualitative methods was used.

All scripts used for training and scoring were benchmarked performances. Seven benchmarked compositions were selected to illustrate performances at each of the scale levels. The relevant performance level was indicated on each performance. These seven scripts were used by the researcher to illustrate typical performances across the range of the scale. The remaining benchmarked scripts were not labelled and were used during the scoring sessions following training. In total the raters each scored thirty-four compositions.

Trialling involved an introductory session, followed by four scoring iterations conducted on site. The first iteration comprised a blind scoring exercise, followed by standardisation in two iterations. After each iteration, results were compared, discussed and misunderstandings were clarified. The fourth and final iteration served as a trial examination scoring session (simulating FET scoring circumstances) to determine the degree of inter-rater and intra-rater consistency, using the proposed scale for assessing writing.

Scores for each of the four iterations were calibrated by the researcher.

In addition, qualitative data were collected in the form of retrospective written protocol reports and a trial examiner questionnaire (Appendix D). Written protocols were used as opposed to verbal protocols due to practical constraints. The questionnaire was based on one used by Shaw and Falvey (2008) to evaluate the attitudes of raters towards the IELTS scale for Writing Assessment. Their questionnaire was adapted for the purpose of the present study and used during Phase 4 to collect information on raters' experience and application of the scale, in addition to the individual written protocol reports.

6.5.4 Analysis

Analyses conducted individually for each of the four iterations included reliability estimates, viz. Pearson's correlation, Spearman's rank order correlation (similar to Fisher Z

transformation) and Kendall's concordance coefficient (similar to Cohen's Kappa) (cf. Cramer & Howitt, 2004 for a discussion of these procedures). Average inter-rater correlations (Pearson's correlation) with Cronbach alpha coefficient based on it, and Kendall's concordance coefficient, based on Spearman's rank correlation coefficient, were as calculated between all the raters (using STATISTICA, 2008) along with Cronbach's alpha coefficient. Then intra-class correlation coefficients were calculated to measure the reliability for an individual rater and generalisability coefficients (SAS, 2005) (based on intra-class coefficient) to measure the reliability of the panel as a whole.

Pearson's correlation matrix was used to calculate average inter-rater correlations and the Spearman's rank order was used to correct any distortion inherent in the Pearson's correlation (r) (Shaw, 2004b:5). Correlation coefficient is an index of the straight line (linear) relationship between two variables that can be ordered. Positive correlations indicate that high scores on one variable, such as for a rater or item, agree with high scores on another variable. Negative values indicate that, for example, one rater may have given a high score for a particular feature in a certain essay, while another rater gave a low score for that feature in the same essay. Values range between -1 to +1. Larger correlations – either positive or negative – indicate stronger linear relationships (or agreement) between the two variables, which are more likely to be statistically significant. According to Cramer and Howitt (2004:39), values of 0.80 or higher are usually accepted as large, strong or high. Correlations close to 0 indicate that there is little or no linear relationship between the variables. Small correlations may be statistically significant if the sample is big enough.

Extreme scores or outliers may affect Pearson's correlation coefficient strongly. Cramer and Howitt (2004:67) explain that the distribution of differences between the two correlation coefficients becomes skewed and unmanageable without transformation by means of calculations such as Spearman's rank order. Spearman's rank order can be used when data are not distributed normally, or when there are too few data to determine whether values are distributed normally. Spearman's rank order coefficient is a logarithmic transformation of Pearson's correlation coefficient. In other words, it is a Pearson's correlation conducted on data that have been rank ordered (Howitt & Cramer, 2000:116). It is used to compare the size of two correlation coefficients from unrelated samples. Data are ordered from the lowest value to the highest and assigned values accordingly, starting with 1 assigned to the lowest value.

If the data are not distributed normally, Pearson's correlation coefficient will lead to incorrect results. Spearman's rank correlation (or Spearman's rho) can be used to correct for such cases. It differs from Pearson's correlation in that the values are first converted to ranks and then the coefficient is calculated (Lohninger, 2009). Spearman's rank order coefficient (also called Spearman's rho) was calculated to correct for any bias inherent in Pearson's correlation coefficient calculation. This test uses a ranking system to assess the degree to which two sets of data correlate. The sets are placed in rank order next to each other to be compared statistically. Spearman's rho is a non-parametric measure of correlation.

The likelihood of obtaining a relationship between two variables is reported in terms of the probability that there is a relationship between the variables, that it is not due to chance, and that it may be significant. Howitt and Cramer (2000:129) explain how the probability estimate should be interpreted:

> If the chances of a relationship being found between two variables are five times or less out of a hundred, that relationship is unlikely to have occurred by chance. It suggests that a relationship actually exists in the population. This probability is usually expressed as being less that the proportion of 0.05 ... and is normally abbreviated "$p<0.05$ where *p* stands for 'probability', < for 'less than' and 0.05 for 'five times out of a hundred'.

Cronbach's alpha reliability test was used to calculate confidence levels for the reliability estimate calculations. It indicates the degree to which items that make up a scale are related. It should vary from 0 to 1, and measures with an alpha of 0.75 or higher are considered to be internally consistent. The more similar items are, the higher the alpha is likely to be (Cramer & Howitt, 2004:79). High coefficient values show that raters produced similar patterns of giving high/low scores for the same sample of performances (Stemler, 2004).

Kendall's W test was conducted to assess inter-rater reliability for each of the fifteen scripts scored in this iteration. Kendall's concordance coefficient expresses the relatedness of scores for different cases or, in this case, essays. In other words, it indicates the degree to which an essay received similar scores from different raters or the strength of agreement between raters. Kendall's concordance considers the essay, not the rater, as the main facet or the constant, with raters as variables.

Intra-class and generalisability coefficients were also calculated (by means of STATISTICA, 2008) to investigate the degree to which the scale could be generalised to other situations, i.e. may be applied consistently by the larger population of FET examiners and implemented for the purpose of assessing the FET Writing examination.

Rasch analyses were conducted for each iteration to determine the reliability of the rating procedure when applied by a single rater and by the group using the scale (Stemler 2004; Shaw, 2004b:5). Data from each iteration (in the form of scores) were used to investigate the following:

- whether the scoring consistency among the twenty raters increased with training across the four iterations using the draft scale;
- item difficulty and rater bias towards any of the fifteen features or the five criteria;.
- the degree to which the fifteen features measure the same construct;
- the degree to which the features represent the construct of assessment.

The qualitative data from the written protocol reports and the questionnaire were analysed to validate the results from the qualitative analyses. The feedback provided by the individual raters for each script were computerised and analysed by the researcher. Each rater's notes on each of the fifteen features scored, for all the essays, were compared. Questionnaire results were also computerised and responses from the different participants on each question were compared. The general tendencies were summarised and typical examples of raters' responses were selected as illustration. The researcher counted the number of positive and negative responses in each case and reported the general attitude of the raters. Contrasting comments were highlighted and reported.

6.5.5 Outcome

Phase 4 produced the final draft of an empirically validated rating scale to assess writing at Grade 12 level.

6.6 Conclusion

This chapter has provided an overview of the method of research followed in the present study to validate the proposed rating scale empirically. The following chapter describes the empirical process of developing the rating scale in detail.

CHAPTER 7

Development of the Rating Scale

7.1 Introduction

Shaw (2004b:3) points out that "(i)t is essential to the success of (a) revision project that considerable effort is devoted to the validation of the rating scale prior to widespread use". Following the overview of the method of research provided in Chapter 6, this chapter focuses in detail on the steps involved in each of the four phases of developing an empirically validated rating scale for assessing writing at the Grade 12 level. In the first place, a benchmarking exercise was undertaken, followed by the drafting of a new scale. After that, the new scale was revised and refined, and finally it was trialled. The results of the analyses during each phase are also reported and discussed.

7.2 Phase 1: Benchmarking exercise

As stated in the previous chapter, the aim of Phase 1 was to establish benchmarked examples of typical learner performances at each of the seven scale levels.

The sample compositions were sorted into seven groups, representing the seven scale levels, on the basis of the original scores assigned by examiners and recorded after the external moderation of the final examination. The groups for Levels 2 to 4 contained the most essays, while Level 7 contained the fewest. The external moderator and the researcher sifted through the seven groups and selected a smaller sample of sixty-eight essays to represent the full range of the seven-level scale. No particular criteria guided the selection, other than that the scripts had to illustrate "normal" performances. The selection did not include scripts containing obviously unique features that would be problematic in the benchmarking exercise. For example, scripts where learners merely copied the prompt a number of times were not included in the sample for benchmarking. These sixty-eight performances demonstrated abilities ranging from near-illiterate (little more than an unintelligible garble) to proficient (fluent, well-structured essays). These performances were numbered from 1 to 68. Fourteen experienced raters (cf. 6.2.2) scored the performances, using the current rating scale, individually and off-site over a period of two weeks. All raters were accustomed to scoring written performances for the FET examination using this scale. Each rater received a

printed set of the sample compositions, a copy of the current rating scale and a list of the topics to which test takers responded. They reported separate scores for each criterion (language and content) and recorded their motivation for assigning each score on a commentary page. Scores and commentary reports were returned to the researcher.

The researcher converted the raw scores to scores between one and seven on the basis of the levels in the current scale. The ranges of scores represented by each level of the current scale are not equally distributed across the different levels (cf. Chapter 5). The rationale for this unequal division is not clear, so the researcher revised the distribution of score points across the scale levels to be equally distributed across the levels. Figure 7.1 illustrates this conversion. After the conversion, each scale level represented a range of seven scores.

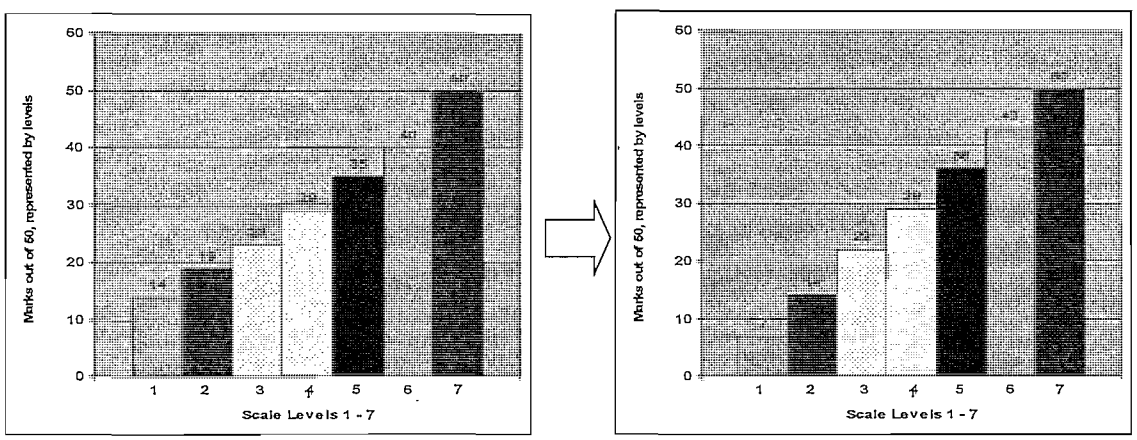

Figure 7.1 Conversion of unequally distributed score ranges per level to equal distribution of 50 marks across seven scale levels

Once the raters' scores had been converted to band level scores, the data were computerised and prepared for Rasch calibration. Three facets were specified for the calibration, namely *essay*, which represents the ability of the learner (1-68), *rater* (1-14), and *score* for items 1 (criterion 1: language) and 2 (criterion 2: content). For the purpose of this study, calibration entailed a statistical comparison of scores assigned by different raters to a set of performances using the same instrument. The researcher analysed the data using the FACETS Rasch program (Linacre, 2006b) to investigate the following (cf. 6.2.4):

- the scoring consistency among the fourteen raters;
- the degree to which the sample of essays represented the full range of abilities on the scale;

- item difficulty and rater bias towards either of the criteria;
- the accuracy of the levels at which essays were benchmarked by the raters.

Misfitting data points reported after each conversion of Rasch were deleted. The researcher repeated the analysis with the modified data until no unexpected results or misfitting items were identified. In total, three Rasch calibrations were conducted before no outliers were identified and the benchmark performances across the full range of the scale were established. During the three repetitions of the analysis, four essays were identified as particularly problematic and resulted in misfits. These essays were removed from the sample, leaving sixty-four examples of performances. These essays were established as benchmark performances, representative of all seven scale levels. Figure 7.2 reports the final results of the calibration in the vertical ruler report or summary map produced by the FACETS program (Linacre, 2006b) for the benchmark exercise calibration. It indicates the interaction between learner ability (*Essay*), rater severity (*Rater*) and item difficulty (*Criteria*).

The first column depicts the logit scale on which all facets are measured. The second column indicates learner ability or the distribution of learners' performances across the scale by placing the essay numbers on the logit scale. The higher the essay is placed on the scale, the higher the learner's ability.

The third column presents the distribution of raters, indicating the degree of inter-rater consistency as well as the degree of accuracy with which they scored. Raters placed higher on the scale were more lenient in their scoring than those placed lower on the scale. Ideally raters should be placed in line at the 0 logit, indicating perfect inter-rater agreement and accuracy. Such a spread is highly unlikely, however, since rater variance is never completely removed. A clustering of raters at the 0 logit mark is therefore acceptable.

The fourth column gives an indication of the difficulty level of the criteria or, in this case, whether one criterion was assessed more severely than the other. The scale levels are indicated in the last column on the right.

The distribution of the sixty-four performances across the range of levels (second column) shows that the performances represent all seven levels of the current scale. As expected, more

examples are placed around the middle, indicating lower to upper-average abilities, with fewer performances distinguished towards the ends of the scale.

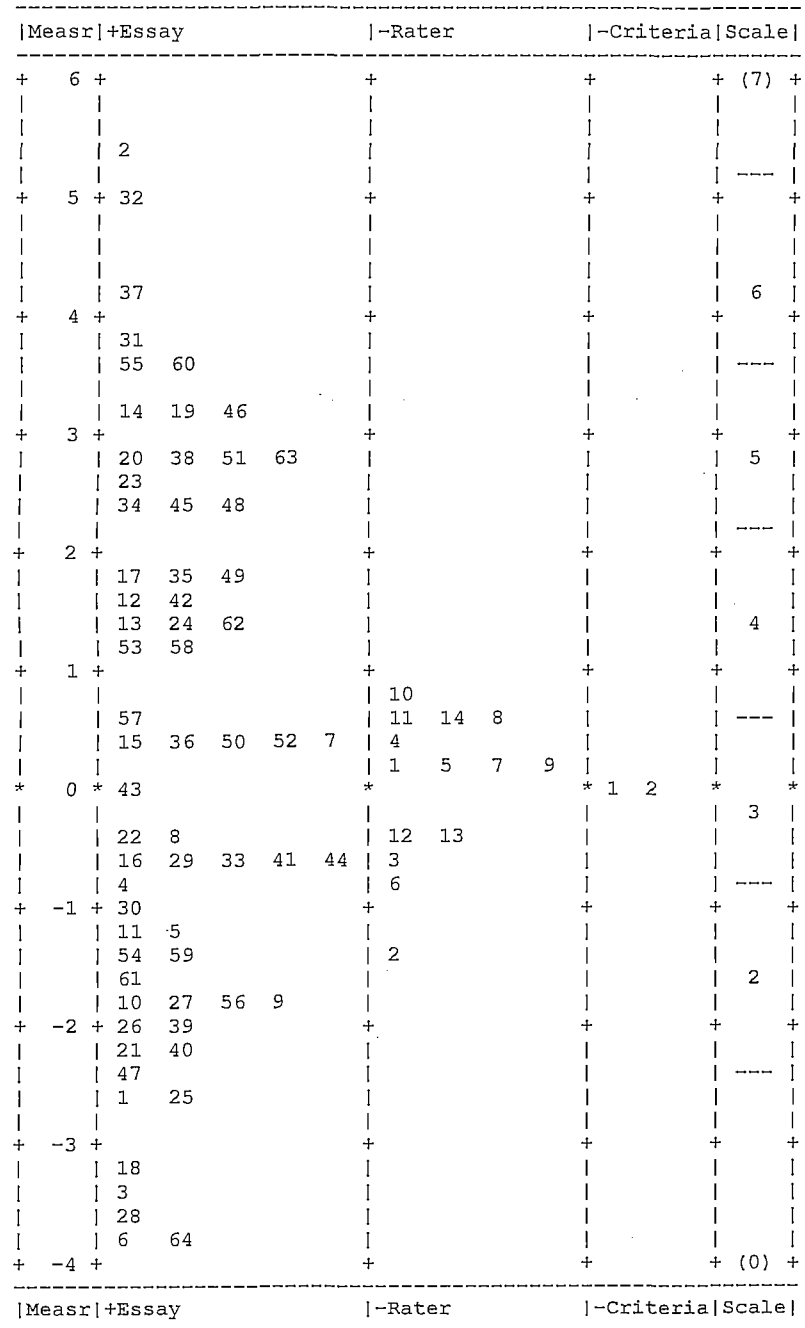

Figure 7.2 FACETS vertical ruler report for Phase 1 benchmarking exercise

Only one essay, number 2, was identified as a typical Level 7 performance. It would have been preferable to have had more examples of typical learner writing at Level 7 for the purpose of this study. A larger sample may have contained more examples of high-level

performances. However, the results in Figure 7.2 may be an indication that the current scale does not distinguish extreme level performances (e.g. Levels 6 and 7) clearly and effectively. The FACETS Vertical ruler report (Fig. 7.2) indicates the estimated true level of ability on the scale (as opposed to the observed values or true scores), since Rasch takes rater variance into account (cf. 5.7 and 6.2.4). Some of the compositions assigned a Level 7 score by the original examiners may therefore have received higher marks than warranted. If the sample is accepted as representative of Grade 12 English FAL learner performances and the results are generalised, the results in Figure 7.2 may indicate that few learners at Grade 12 level achieve high levels of writing proficiency (cf. Chapter 4).

The reliability index reported in the essay measurement report (cf. 6.2.4) indicates the degree to which different levels of ability were distinguished. FACETS reported a very high reliability index of 0.98, which indicates that the performances at different levels were clearly distinguished. As mentioned (6.2.4), the closer the reliability index value is to 1, the more accurate the distinction. The placement of essays illustrated in Figure 7.2 (second column) was used to identify essays at each performance level and assign benchmark level labels accordingly.

Column three shows all raters clustered together, indicating that they were in close agreement with each other. The raters are grouped around the 0 mark within 1 logit measure, which means that the raters were fairly accurate (not excessively harsh or lenient) in their scoring. Rater 2, placed slightly below the -1 logit mark, was the most severe rater and was consistently harsher in assigning scores (i.e. assigned lower scores) than the other raters.

The rater measurement report indicated fit values within the acceptable range for all fourteen raters involved in the benchmarking exercise. No significant inconsistencies were reported for any of the fourteen raters, so the data for all raters were included during the analyses. Despite the slight separation of Rater 2 in the summary map (Fig. 7.2), an infit value of 0.91 and outfit value of 0.87 was reported for this rater, indicating that it was not misfitting. Therefore, it was not necessary to exclude Rater 2 from the data set. In a few cases, a particular rater's score for one criterion was inconsistent with those of other raters. In these cases, the single data point was deleted. A very high reliability index of 0.94 was reported for the selection of raters, indicating distinct variance between the raters, as expected (cf. 6.2.4).

The fourth column may be interpreted as indicative of item difficulty, but in this instance it should be interpreted as indicating whether the first criterion (language) was marked more harshly or more leniently than the second (content). Items placed higher on the logit scale may be considered easier, or were marked more leniently by the raters, as opposed to those placed lower in the scale, which were marked more severely. Both criteria are placed at the 0 logit mark and no significant bias was reported for either criterion. The criteria measurement report did not identify either criterion as misfitting. A high reliability index of 0.80 was reported for the criteria.

In summary, the results of the benchmarking exercise were positive. A sufficiently representative sample was selected to establish typical examples of performances at different achievement levels. The panel of experts used to score the performances agreed sufficiently to establish benchmark scripts. Sixty-four scripts were established as benchmark performances across the seven performance levels. These typical examples of learner writing at each scale level were to be analysed in Phase 2 to determine the distinguishing features at each level and describe these features in a draft scale.

7.3 Phase 2: Drafting a rating scale

To achieve the aim of drafting a rating scale in Phase 2, the second panel of experts (cf. 6.3.2) analysed the benchmarked performances to identify and describe the distinguishing features at each of the seven performance levels.

The scale was drafted during a two-day workshop conducted by the researcher. Prior to the workshop, the participants involved in Phase 2 received copied sets of benchmarked performances labelled according to the performance level they represented. They each also received examples of ESL writing scales currently in use, including the current scale used for assessing the FET writing examination (Appendix A), as well as background information on the study, a literature overview related to validation and writing rating scale development, and the role of the participants in the second phase.

The participants familiarised themselves with the compositions and scales and evaluated both. The aim of this preparation exercise was to determine which features of writing individuals would identify as characteristic of each performance level and point out where the

current rating scale is deficient in describing such features. They studied the benchmarked performances, noting those features which they recognised as typical and prominent at each performance level. The experts also examined the examples of scales critically, focusing on different scale formats, features that were typically addressed, how criteria were generally labelled, and how descriptors were formulated. The aim of this critical evaluation was to identify the positive and negative aspects of these scales to inform decisions pertaining to the format and content of the new scale.

The workshop commenced with an orientation session conducted by the researcher, to inform the participants of the aim of the workshop and the role of the panel in the development of the rating scale.

After this initial session, the panel conferred to determine which type of scale would be most suitable for the purpose of the FET examination. With Section A of the National Senior Certificate Writing paper in mind, they determined the scope of the assessment, discussed the construct of writing and selected an appropriate scale type (cf. Chapter 5). The panel decided that a multiple trait rating scale would be appropriate to assess essay writing in the FET examination. Such a scale would offer specific guidelines to improve inter-rater consistency, as well as provide a comprehensive picture of learners' ability which could also enhance teaching and learning practices through feedback in a communicative context.

They then considered typical, established examples of multiple trait scales, such as Jacobs et al. (1981), the IELTS scale and the TEEP writing scale, and discussed the most appropriate format for addressing the construct in a scale. After evaluating a number of examples, the panel decided to draft the scale in the form of a 7-point semantic differential scale with extreme bi-polar descriptors. A 7-point scale adheres to the DoE's requirement regarding writing assessment, as pointed out in Chapter 5.

Semantic differential scales provide binary terms (such as "black" or "bad" as opposed to "white" or "excellent") at the ends of a continuum. Raters then have to evaluate the degree to which a performance accords with these extremes.

The proposed format would present raters with a number of distinct but related features under each criterion to be assessed individually on the 7-point Likert-type scale. A total score would be calculated by adding scores for individual features to a total out of 100.

The format of the current scale was therefore abandoned because the panel wanted to include more than two criteria. Other scales, such as Jacobs et al. (1981), provide a list of descriptors under each criterion to be considered simultaneously (similar to the current scale), requiring raters to assign one mark for various features. However, the panel felt that such a presentation may lead raters, consciously or not, to concentrate more on one micro-feature rather than consider all features related to a criterion equally. Raters may thus equate a criterion with one salient feature instead of a collection of related features, which is unfair towards test takers who may be more developed in another related feature.

After deciding on the most appropriate type of scale, each panel member reported his/her findings of the analysis of the benchmarked performances conducted in preparation for the workshop. Panel members generally agreed on the features they considered typical distinguishing characteristics of writing at different performance levels. Minor differences were resolved through discussion. During the discussion, aspects related to content, the function of the message, communication with the audience, structure and development (flowing ideas, paragraphing and coherence at sentence level), style and register, vocabulary, grammar, spelling and punctuation were identified as prominent aspects of communicative writing that needed to be addressed in the scale.

Once they had reached agreement on the features recognised as salient, the panel examined the scripts as a group to verify their findings. They first focussed on analysing the scripts to identify salient features at macro-level and establish criteria, and then at micro-level to phrase descriptors in the draft scale.

The macro-level analysis entailed grouping the features into prominent categories resembling different aspects of the construct and suggested criteria. All the features identified by the individual participants prior to the workshop were pooled and then divided into categories with similar or related features. The experts conferred over appropriate labels for the macro-level categories. These labels would constitute distinct criteria, representative of the various salient features of writing.

Prominent categories related to grammatical accuracy, organisation at paragraph and sentence level, development of ideas, vocabulary range and use of stylistic devises were identified. The criteria were evaluated in terms of the sample scripts and the pool of features to make sure that all categories and core features were exhausted. In addition, the Assessment Standards for Writing and Presenting as well as those for Language (NCS, 2005:38-43) were used as checklist to ensure all stated requirements were addressed by the criteria. After some discussion, five criteria were established, viz. content, structure and development, grammar, vocabulary and editing.

Once the assessment criteria had been established, the panel evaluated the scripts and criteria at micro-level in terms of the features addressed under each criterion. The original categorisation of the micro-level features under criteria was re-considered to ensure that the most salient features were accounted for and that all features were represented by the established criteria. Scripts were scrutinised again to ensure that all the most salient distinguishing features had been identified.

The panel also discussed typical problematic issues that occur during scoring, such as off-topic scripts and illegible handwriting, before phrasing band level descriptors and weighting criteria. The panel decided not to address off-topic scripts as a distinguishing feature in the scale, but rather as an additional comment to be considered before raters scored all the salient features that were typical of writing. They argued that, on the basis of the nature of the first three criteria (content, structure and development, and grammar), it would only be relevant to assess off-topic scripts for the remaining two criteria, vocabulary and editing. So, if a performance were off-topic, raters would only score that performance in terms of vocabulary and editing. Handwriting was considered as related to tidy presentation, which was addressed as one feature related to the editing criterion and for which only two marks were assigned.

Finally, the criteria and salient features were organised in a draft scale through a process of three iterations (cf. Appendix B for the three versions of the draft scale). After each iteration, the panel evaluated the draft scale critically, focussing on the organisation of micro-features to represent the criterion, as well as the short bi-polar descriptions of the ends of the continuum relevant to each feature.

The final draft (Appendix B3) comprised five criteria with fifteen salient features spread between the five main categories. Micro-features under each criterion were ordered to guide raters' focus during scoring from macro-level features to micro-level features.

Weighting of criteria was determined by the number of features related to each particular criterion. More features listed under a criterion indicated that that particular criterion warranted heavier weighting in assessing the construct than other criteria addressing fewer features. For example, category A, structure and development, comprised four sub-scales, thus carrying a weighting of twenty-eight out of one hundred marks. Category E, editing, comprised three sub-scales and was allocated twenty-one marks. In terms of reflecting the essentials of the writing skill, good structure and development is generally considered to be more important than editing. This system provides clear information on what learners are required to produce, and what they are rewarded for.

The following points summarise the rationale for using the proposed format for a new scale:

- The macro- and micro-categories included in the scale are based on the features identified in examples of actual learner writing at Grade 12 level.
- Raters need to consider features individually, thereby accounting for the overall mark they assign. Chances of teachers assigning impressionistic scores are reduced and the scoring validity therefore increases.
- Scoring validity should increase because final scores are calculated from a number of marks.
- The use of bi-polar descriptors, which describe only the extreme ends of the scale for each feature, reduces ambiguity in descriptors and should improve rater consistency. Vague terms such as "full" and "complete" used in the current scale to distinguish adjacent levels (cf. Appendix A) are avoided.
- The scale accommodates L2 learners who develop some features faster than others, and not necessarily at the pace as other learners at the same level of ability.
- Criteria are not weighted impressionistically, but based on the number of related features that need to be considered when assessing individual criteria.

- The order in which features are presented in the scale should guide raters to consider criteria first at macro-level or on the "surface" before "looking deeper" and considering the micro-level aspects related to the criterion.
- This scale provides explicit guidance on what raters should be looking for while assessing writing.
- Since the scale addresses specific features, it should be valuable in providing detailed feedback on learners' strong and weak points.
- Using a semantic differentiated scale offers the advantage that it is fully quantitative, making further statistical analyses easier.
- If a fully quantitative scale such as the proposed scale were implemented, it would be possible to sustain continuous *a posteriori* validation procedures (such as statistical analyses) to evaluate the consistency with which the scale is implemented. Such information may also be valuable in utilising training and retraining sessions optimally.

One concern regarding the format of the scale was that it may be time-consuming, owing to the additional administration of adding fifteen scores. As mentioned above, however, the proposed scale has the added value of providing specific diagnostic information on learners' abilities. In addition, it offers potentially more consistent and accurate scores. The panel argued that achieving high levels of accuracy and reliability should prevail over the convenience of raters and quickness of scoring. Weir (2005:49) also emphasises that practical convenience should not be the determining factor at the expense of a comprehensive representation of the construct in scale design (cf. Chapter 4). Practicality is not a necessary condition for validity and should therefore not be a determining factor in the scale design. Scale developers should strive to make an instrument as practical as possible, but not at the expense of accurate and fair scoring, i.e. scoring validity.

The panel further argued that scoring would become easier as raters became familiar with the scale. Assessing individual features should provide raters with a more systematic and focussed way of scoring, which may improve consistency. The clear focus may also make it easy for raters to internalise the scale.

In addition to the draft scale, the panel compiled an explanatory scale guide (Appendix B4) providing definitions and descriptions of each criterion and salient feature. Typical examples should accompany this guide to illustrate the standard of writing in terms of the features that could be expected at each level. The guide would need to be an integral part of training and re-training, as well as guide raters in their decisions while scoring.

After the workshop, the researcher conducted a calibration exercise to determine the following:

- the scoring consistency among the five raters using the draft scale.
- the degree to which the sample of essays represented the full range of abilities on the scale.
- item difficulty and rater bias towards either of the criteria.
- the degree to which the fifteen features measure the same construct.
- the degree to which the features represent the construct of assessment.

For this exercise, four members of the panel each scored thirty performances, using the draft scale, one month after the workshop. Performances were randomly selected from the scripts benchmarked in Phase 1 and labelled 1 to 30. Off-site scoring took place over a two-week period. Their scores were computerised and the data were analysed using the FACETS Rasch program (Linacre, 2006b). Results are illustrated in the vertical ruler report presented in Figure 7.3. This procedure verified the panel's analyses of salient features at macro- and micro-level, and the description of these features in the scale. It also indicated the degree to which the trial raters tended to apply the scale consistently.

These results indicate that the draft scale was successfully and consistently applied by the four raters involved in the calibration exercise. However, the placement of raters between -1 and -2 logits in Figure 7.3 (2nd column, "Rater") indicates that the raters tended to be somewhat harsh in their scoring. All four raters are placed within one logit measure of each (second column, Fig. 7.3), indicating a sufficient inter-rater consistency. If the raters had undergone standardisation training they may have scored more accurately, i.e. placed closer to the 0 logit mark.

None of the raters was identified as misfitting by the rater measurement report. A high reliability index of 0.96 was reported for the sample of raters. This indicates that the raters were clearly distinguished in the scores they assigned (i.e. there were differences between them), as expected.

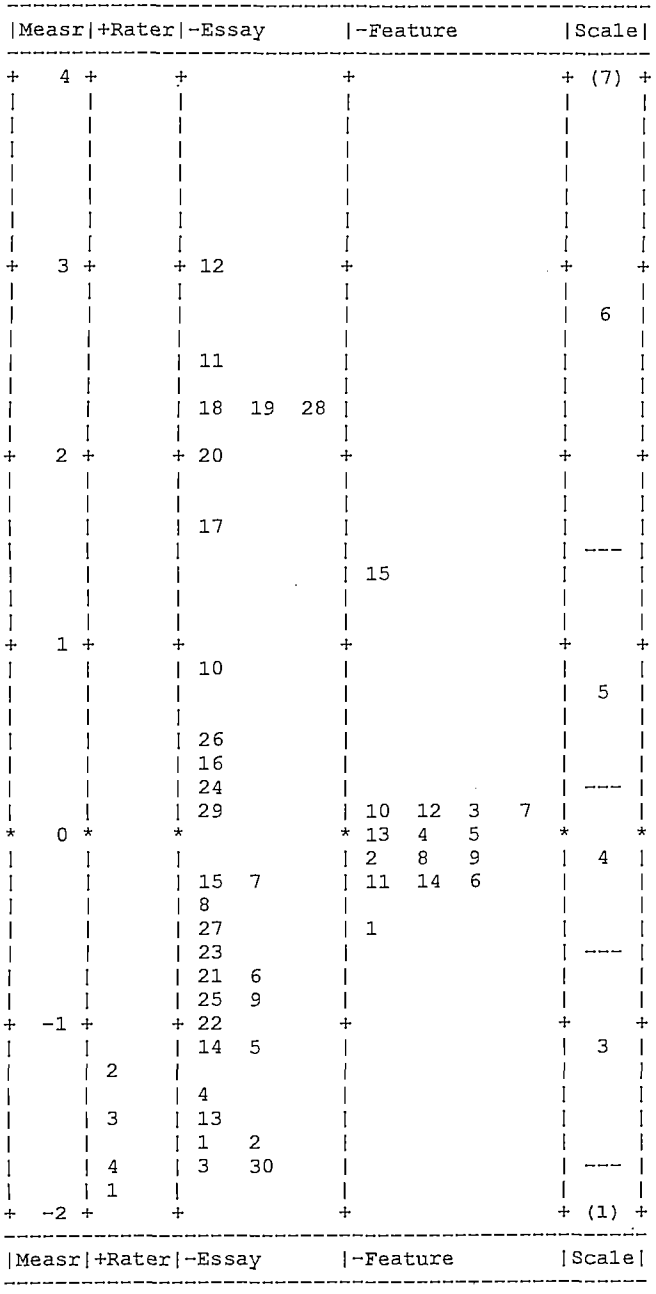

Figure 7.3 Vertical ruler report for the draft scale calibration

The fact that the sample of performances scored for this exercise did not represent the full range of the scale (3rd column, spread across levels 3 to 6) was not a reason for concern for the purpose of the calibration exercise.

Only one essay (number 18) was identified as an outlier with an infit value of 2.12 and outfit value of 2.11, indicating an overfit. This means that the essay received a higher score than expected, i.e. the raters were lenient in scoring this performance. Closer examination revealed that the raters differed considerably in the scores they assigned for this performance. The reason for this was unclear.

A Rasch reliability index of 0.99 was reported by FACETS in the essay measurement report for the calibration exercise. This indicates that there were distinct differences between essays (learners' abilities), which may imply that the raters were able to distinguish the level of performances clearly using the proposed scale. The index value is slightly higher than that achieved during the calibration in Phase 1, perhaps suggesting an improved scale that could guide raters to score more accurately.

Column 4 indicates the relative difficulty of the features addressed in the scale as well as the degree to which the features addressed in the scale are distinct (i.e. that a feature is not duplicated in the scale), while being relevant to the same construct. The placement of features in the fourth column at the 0 logit mark indicates that the content of the scale reflected the construct. Fourteen of the fifteen features were grouped closely together, indicating that these features were related but distinct. Furthermore, these features were scored consistently and could be used to distinguish different levels of ability in performances. Thus, the scale should assist raters in distinguishing candidates who perform better on some features than on others, as is typical in L2 learner writing.

Only feature 15, tidy presentation, seems misplaced in relation to the other features. From the results, this feature seems to have been considerably easier than the others and perhaps not directly related to the other features. This is not surprising, since this feature was assigned only two marks. The panel argued that tidy presentation was an aspect that examiners usually considered when scoring writing, but that it should not warrant seven marks. No extreme bias was reported for any of the features, including feature fifteen, for which with an infit of 0.72 and outfit of 1.48 were reported. A high reliability index of 0.96 was reported, supporting the

indication of the vertical ruler report that the features addressed in the rating scale are distinct from one another, and are all relevant to the same construct.

In summary, a scale was drafted in Phase 2, comprising five criteria with fifteen individual features specified across the criteria.

7.4 Phase 3: Revising and refining the scale

A second workshop was conducted by the researcher during which a third panel of experienced examiners revised and refined the draft scale. The workshop took place over two days. As mentioned in the previous chapter (cf. 6.4.2), the third panel comprised twelve members who evaluated the draft scale critically. They adjusted the scale, and the researcher conducted a Rasch calibration exercise to determine whether all potentially problematic aspects had been identified by the experts.

The participants were not given any material to study in preparation to the workshop. On arrival at the workshop, each one received a workshop folder containing materials for scoring, revising, training and trialling purposes. Each folder contained a copy of the draft scale and the scale guide constructed in Phase 2, accompanied by seven benchmarked performances labelled according to the levels illustrated by each. The folders also contained a number of unmarked (but benchmarked) scripts to be scored in a series of scoring sessions and examples of other writing rating scales currently used in international practice. Finally, participants received a blank feedback page for reporting their experience at the end of the second day.

The workshop commenced with a blind scoring session during which each participant marked two performances using the draft scale before the researcher formally gave a short presentation to introduce the scale. The basic structure of semantic differential scales was explained and illustrated by various examples. Raters were instructed to apply the proposed scale following a two-step decision making process to determine a suitable score in terms of each individual feature. First, they needed to decide whether the performance was average, or above or below average. Once this was established, raters could decide the extent to which the performance was average, or above or below average. Seven benchmarked examples were selected to illustrate the quality of performances expected at each of the seven scale levels.

The bi-polar adjectives used to describe each feature and problems related to choosing the most appropriate formulation were highlighted.

After the blind scoring and information sessions were concluded, the panel provided short verbal feedback reports on their first impression and experience of the scale. All participants were in general agreement that the draft scale was a great improvement on the current scale. Their feedback indicated that the draft scale was easy to understand and apply, and that it provided clear guidance as to what raters should look for in written performances. They also felt that it offered raters a means of scoring systematically with a clear focus.

One rater, however, noted that she found the five criteria somewhat overwhelming and that it was easier to concentrate on the two criteria distinguished in the current scale. This rater admitted that she was resistant to the idea of familiarising herself with a new scale. She did, however, find all five criteria (and all features addressed) appropriate, relevant to the construct, clear and easy to understand.

This rater also raised concern about the additional time it would take to arrive at an overall score, owing to the fifteen individual scores that had to be added up. Some of the other panel members acknowledged this, but were more willing to accept the additional effort of adding up scores if the scale assisted systematic and accurate scoring. In spite of her expressed apprehension about adopting a new scale, the rater acknowledged the potential benefits of such an explicit scale. The group agreed that a more precise scale and accurate scoring were more important considerations than accommodating examiners' preferences for fast scoring. After some discussion the apprehensive rater added that, once she had managed to internalise the scale, scoring would probably be easy and not excessively time-consuming. Further discussion revealed that raters could see themselves getting used to the scale quickly and internalise it easily. They rationalised that the proposed scale provided them with a recipe for step-by-step scoring of essays, which would make for a more systematic process.

The panel then evaluated the content of the scale critically to identify aspects that needed to be refined. They identified the number of criteria, their weighting in conjunction with the organisation of the salient features, and the formulation of the level descriptors as aspects that needed to be revised during the remainder of the workshop. These aspects were considered in

a panel discussion, after which the experts divided into smaller groups for closer evaluation and refining of the scale content.

As mentioned above, one of the raters indicated that she experienced the distinction of five criteria as overwhelming. The panel therefore discussed whether any of the criteria could perhaps be combined to distinguish four criteria, but they decided against this since the criteria were clearly distinct. They decided that all five criteria were relevant, but agreed that the weighting of the criteria and the descriptors needed to be refined.

The panel then focussed on the weighting of criteria and in particular that of the editing criterion. The editing criterion was assigned a weighting of sixteen marks, while that of the structure and development criterion was only fourteen (cf. Appendix B3). The raters considered this division unfair in terms of the purpose and construct of the FET assessment. In addition, the participants pointed out that the editing criterion did not address evidence of planning adequately, as required in the NCS (2005). This is also a shortcoming of the current scale and was therefore noted by the panel as an important aspect to be included in the draft. The group agreed to reconsider the criterion for editing and focus on finding a more appropriate way of incorporating evidence of planning, spelling and punctuation elsewhere in the scale.

Adjusting the weighting of criteria entailed adjusting the distribution of the criteria addressed under each of the five criteria. This resulted in a revision of all the micro-level features, which impacted on the labelling of the criteria.

In their evaluation of the micro-level features, the panel identified a number of aspects for reconsideration. Addressing the problem of explicitly incorporating planning in the scale proved problematic, since the panel differed on how they conceptualised evidence of planning. Some participants argued that planning was evident in a performance if the final product was coherent and organised. Others considered evidence of planning to be explicit rough notes such as mind maps. Both arguments may result in bias towards learners with different styles. Some learners may plan an essay in detail, but still only manage to present an average performance.

On the other hand, it would be difficult to grade planning depicted by mind maps. If marks were awarded for drawing a mind map, all learners who made an attempt at drawing a map would have to score seven points if it were included as a distinctive feature. Also, learners could easily construct a mind map after completing their performances, simply to comply with the requirements.

Furthermore, the panel questioned the term *editing* as the most appropriate one to refer to features such as planning and finally decided to incorporate features related to editing and planning as part of other criteria related to content and structure and development. The argument was that planning was not only evident in the form of mind maps, and that the presentation of a mind map does not necessarily indicate that learners had consciously planned their response. In addition, it would be difficult to train raters on how to score mind maps as proof of planning. No final decision was made at this point.

After reaching consensus on the five criteria and the weight assigned to each, the panel considered the formulation of the bi-polar descriptors. Their main concern was that the descriptors should be unambiguous and help raters to distinguish between different performance levels. In evaluating the bi-polar descriptors for the individual features under each criterion, the panel focussed on identifying formulations that were vague or terms that did not indicate the complete extremes of the scale. For example, terms such as "sufficient" were not considered to be indicative of top level performances, as these could also be used to describe performances at Levels 3, 4 or 5.

The panel agreed not to provide detailed level descriptors for each of the seven scale levels in order to avoid using ambiguous terms to distinguish adjacent levels. Some members, however, felt that describing only the extreme ends of the scale left too much to raters' personal interpretation. They suggested that one-word labels be used to demonstrate the general standard of performances expected at some or all of the levels. The rationale was that such labels would provide more explicit guidance, especially to less experienced raters, without confusing raters and avoid potential information overload in terms of descriptors.

After the panel discussion, the group divided into three smaller groups to evaluate the content of the scale, paying particular attention to the issues raised by the group. Each group evaluated the draft in more detail.

First, they focussed on addressing the problems related to the weighting of the criteria, which entailed revising the choice and distribution of the salient features to represent the different criteria. Then they refined the descriptors of the features and levels.

Each group presented a revised and refined version of the scale drafted in Phase 2. The panel compared and discussed these three revised versions of the scale. Each group stated the rationale for their suggestions and differences were resolved through discussion. The draft scale was adapted accordingly and copies were made for all the participants.

The weight for the content criterion was increased to thirty-five marks out of the total one hundred. It was adjusted to address planning and organisation of ideas, as well as sensitivity to the audience and appropriateness to the context.

Fourteen marks were assigned to the structure and development criterion. The panel agreed that the focus of this criterion was on organisational features on surface level, namely paragraphs and a clearly identifiable introduction, body and conclusion.

An additional feature was added under the grammar criterion, motivated by the NCS (2005) specification that learners at Grade 12 level should demonstrate their ability to use a variety of sentence types. Grammar was thus allocated twenty-eight marks.

The panel considered requirements for vocabulary and agreed that this criterion should address features such as range, appropriateness of diction, register and style, and the effective use of linking devices. These lexical aspects correspond with those indicated in the NCS (2005) as relevant at Grade 12 level.

During the comparison of the three revised versions of the draft scale, the panel finally reached consensus on the editing criterion. The editing criterion was replaced by one addressing length alone with a weight of two marks. The group considered feature 15, addressing tidy presentation, as irrelevant, arguing that learners generally tried their best to write legibly and present a tidy essay. Neat presentation was not a requirement for skilled writing, although it may have an impact on the reader's personal impression. In addition,

learners were usually constantly reminded to write neatly during the examination. Therefore, most learners would score the full two marks.

Instead the group argued that length needed to be addressed specifically, since it would reflect an aspect of learners' planning in responding to tasks. Weaker learners would be challenged to write more extensively, while learners who tended to write too much would need to rethink their writing.

Regarding the revision of the extreme bi-polar descriptors, the group found it sufficient to assign one-word descriptors to Levels 2, 4 and 6, with more detailed descriptors at the two extreme points (Levels 1 and 7) for each feature. These labels should guide raters in interpreting the levels and distinguishing between them accurately. The formulation of the extreme level descriptors were revised because the panel did not consider them to be descriptive of what an absolute performance would look like in terms of the particular feature. For example, the term 'adequate' was used to describe a Level 7 performance, but the panel argued that a performance could be much more than 'adequate'. The term only suggested that a piece of writing was 'good enough' – such as a Level 4 – but not 'excellent' or 'extraordinary'. After identifying this problem, the group quickly reached agreement on the descriptors. They concentrated on describing a Level 1 performance as one in which the feature is not present at all, and a Level 7 performance as one in which the feature is perfectly or completely present.

Like the original draft scale (Appendix B3), the final version of the revised scale (Appendix C1) consisted of five criteria and fifteen salient features. The features addressed were revised to ensure a fairer weighting of criteria and clearer level descriptors. This revised version of the draft scale would be piloted in the following phase.

The researcher conducted a calibration exercise to investigate the performance of the revised draft scale (cf. 6.3.4). At the end of the workshop, the raters scored five unmarked benchmark performances using the revised version of the scale. Directly after scoring, each rater provided written feedback in a one-page report, recording their scoring experience with the new draft scale, and particularly in comparison with the first draft scale. The result of the written feedback reports (reported below) were considered in conjunction with those of the statistical calibration.

Figure 7.5 presents the vertical ruler report for the calibration exercise in Phase 3. The five essays marked during this exercise were randomly selected from the benchmarked performances that were not used during the blind scoring session or during the information session. The results of the second calibration exercise indicated that the revised semantic differential scale may lead to more consistent marking.

In the vertical ruler report below, all but one of the raters (Rater 1) are grouped together between the 0 and -1 logit marks. This means that the raters again showed a general tendency to score a *little* harsher than expected, but did so consistently. Some were *still* harsher in their scores (raters placed closer to -1) than others (raters placed closer to 0).

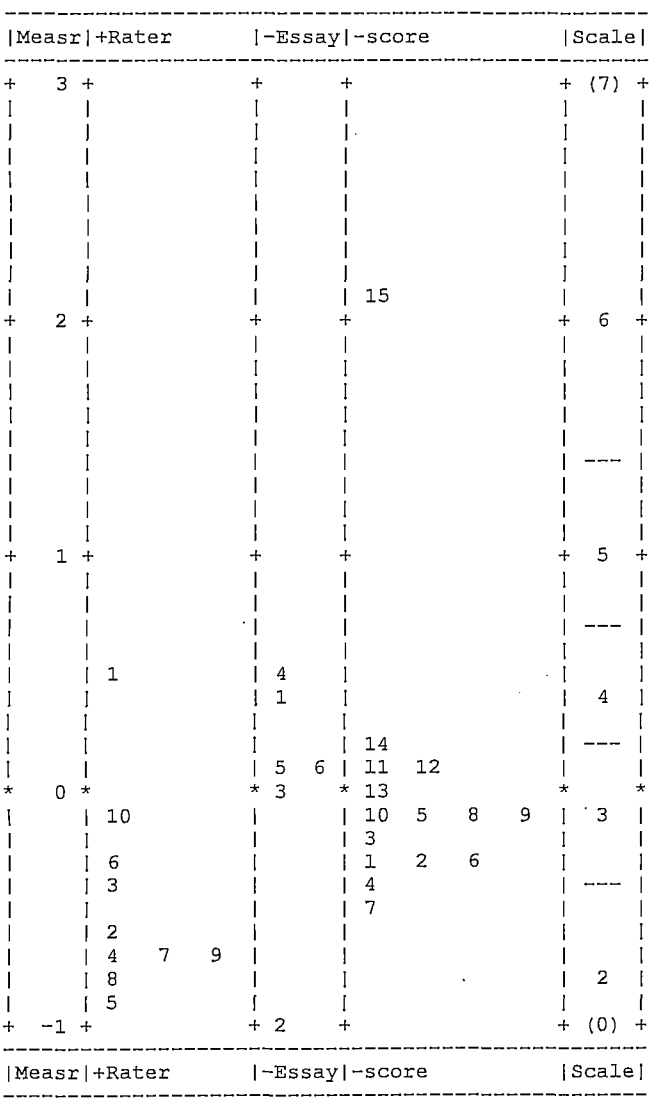

Figure 7.5 Vertical ruler report for revised draft scale calibration

Rater 2 was the only misfitting rater identified with an infit of 2.36 and outfit 2.51. FACETS reported some unexpected responses, which indicated that Rater 2 was inconsistent in scoring the first five features of the first essay. Training may be particularly valuable for such raters in helping them to overcome their uncertainty in applying the scale. A high reliability index of 0.95 was reported for the raters, indicating that there were distinct differences in the scores assigned by them, as was expected.

All five essays were identified as Levels 3 and 4 performances. None of the essays was identified as misfitting and a high reliability index of 0.98 was reported. This indicates that the performances were clearly distinguished from each other when assessed according to the revised draft scale.

Fourteen of the salient features are grouped fairly close to the 0 logit mark, indicating that these features are related to the same construct, while being distinct from one another. Feature fifteen is still an exception owing to the fact that it only contributes two marks out of one hundred. It could therefore be much easier for test takers to score full marks for this feature than for the other features. However, the raters' rationale for including length for two marks seems to justify its inclusion in the scale for the purpose of assessment. The other features are closely related and discriminate features of the construct sufficiently.

Unlike the results of the previous calibration exercises, feature fifteen was identified as misfitting with an infit value of 2.22 and outfit of 2.31. Fit values reported for all the other features were within the acceptable range. A high reliability index of 0.96 indicates that the features were sufficiently distinct, in other words each features addressed an individual aspect of the construct, and was therefore relevant as part of the content of the scale.

The result of Phase 3 was a refined rating scale (Appendix C1) and a revised explanatory guide (Appendix C2). The scale still consisted of five criteria, viz. content, structure and development, grammar, vocabulary and length. However, the distribution and the description of the salient features were revised and refined. This refined scale was then trialled in a pilot scoring session.

7.5 Phase 4: Trialling of the scale

As indicated in Chapter 6, the fourth phase entailed piloting the scale refined in Phase 3. A twenty-member panel of experienced examiners was trained to use the scale in a series of iterations, scoring a total of thirty-five performances. Data were collected in the form of scores, written protocols and questionnaire responses, which the researcher analysed to determine the effectiveness of the scale and whether it could be implemented in practice.

As indicated in 6.5.2, some of the members of the panel involved in Phase 4 were familiar with the proposed scale, having been instrumental in its refinement during Phase 3. The majority of the participants, however, were unfamiliar with it.

The trialling was conducted over a period of two days, taking the form of a workshop consisting of a series of training, scoring and discussion sessions. All scripts used for the trialling session were sample benchmarked scripts. The researcher divided the benchmarked scripts roughly to ensure that each set to be scored contained performances representing the full range of the scale. For the first iteration, seven scripts were selected, each one illustrating one of the seven levels. During the second iteration, the group scored performances illustrating distinctly different levels on the scale, but not representative of the full range of the scale. The third batch of scripts contained scripts representing the middle range of scale levels. This batch contained some scripts benchmarked as borderline between two levels. For the final iteration, the researcher randomly selected a number of essays to roughly represent the full range of the scale.

All sessions took place on site. The first day entailed an introduction to the project, discussion of the proposed scale and the initial standardisation training. On the second day, standardisation training continued, followed by a final session of uninterrupted scoring and written feedback.

The first day commenced with a short introductory presentation by the researcher to inform all the participants about the background of the project and the purpose of the workshop. The researcher introduced and illustrated the idea of assessing writing using a semantic differential rating scale, following a two-step decision-making process (cf. 7.4). The researcher also outlined the procedure that would be followed during the trialling workshop.

The first scoring iteration took place directly after the information session, without any discussion of the proposed scale. Seven scripts were scored blindly, without any discussion taking place during scoring. Although these scripts contained one example of writing representing a performance at each of the seven levels, the panel members were not informed that the set illustrated the full range of the scale. The researcher provided two example scripts, at Levels 2 and 5, to demonstrate the general standard of the performances which the examiners would be scoring. This was done because most of the raters were unfamiliar with the proposed scale.

Once the scripts had been scored, the group discussed their first impressions of the scale. They also compared their scores to those of other raters and with the benchmarked levels and discussed their differences.

All trial examiners reacted positively to the scale, but one examiner raised the issue of the time it took to assess an essay and adding the marks. The majority of the group, however, felt that scoring would become easier and faster once they were more familiar with the scale. They also considered it worthwhile to take a little more time scoring an essay if it meant arriving at a more accurate score. The following comments transcribed during this discussion reflect the most prominent attitudes in the group:

- It is a big improvement on the current rating scale.
- The scale is very comprehensive. It covers everything.
- It is precise and clear. The current rating scale is very general in its descriptions of the two criteria. With the current scale, raters are not always clear about what they are assessing.
- The proposed scale guides raters to address specific aspects; unlike the current scale, you have to consider features separately, so you are less likely to equate features either consciously or subconsciously.
- It is easy to understand the wording and content of the scale.
- Since features are considered separately, raters have a clear focus during scoring.
- The scale is flexible in that it allows for learners differences.
- It allows raters to reward learners for making an effort, even if their performances are not perfect. For example, a learner can be rewarded for demonstrating a logical flow of ideas, even if their paragraphing is inadequate.
- The proposed scale should improve the accuracy of scores.

- The layout of the scale is comfortable to work with. The current scale requires raters to rate in two directions (vertically and horizontally). It is sometimes difficult to reach a crossing point that resembles a fair mark on the current scale. The proposed scale makes this easier to achieve.
- With the current scale, it is very difficult to determine why a certain mark has been assigned to a performance. With the proposed scale, raters have to be specific in the marks they assign for various aspects. Differences in performance levels are therefore more meaningful than with the current scale division.

After discussing their initial reactions to the scale, the group compared the scores they assigned to each individual feature. This was done in order to identify any major differences in their initial interpretation and application of the scale.

The researcher provided each examiner with a clean set of the seven essays containing the benchmarked levels they represented for the purpose of comparison and to serve as reference points during the following iterations. The group discussed the content of the scale to determine whether any of the criteria or features were unclear.

Raters used the scale guide (Appendix C2) to clarify the meaning of features before scoring the second set of essays. One rater asked for further clarification about the feature addressing range in sentence style. This feature was clarified to the satisfaction of the examiner.

The trial examiners were standardised in terms of their interpretation and application of the scale during the second and third iterations. Standardisation entailed scoring sets of performances, followed by comparison and discussion of scores in order to identify any problematic aspects in the scale and discrepancies in level allocation.

During the second iteration, the group scored five essays, divided into two sets of two and three scripts respectively. They compared their results and discussed all discrepancies until they reached agreement. The purpose of the discussions was to determine if and why they had assigned different scores, i.e. which aspects of the scale they were misinterpreting or assessing inconsistently. Raters were asked to motivate why they assigned a particular score to a certain feature.

All discrepancies, including those within adjacent groups, were discussed until the trial examiners resolved their differences and agreed in their understanding of the particular feature. On most occasions the group was able to reach agreement fairly easily on the overall level scores. More discrepancies were revealed in their scoring of individual features. It

became apparent that some trial examiners were inconsistent in their interpretation of certain features in the sense that their judgement of one was influenced by another. Raters seemed to struggle distinguishing among four features in particular, viz. 1: insight into the topic, 2: presenting mature ideas, 3: developing the ideas (content criterion), and 7: paragraphing (structure and development).

Scores for these four features varied from a Level 2 to Level 5, while discrepancies were generally within adjacent levels for the remaining ones. They resolved their differences through discussion, which revealed that individual examiners seemed to disagree about the role that content played in considering these features. It became apparent that some of the raters tended to equate mature ideas with insight into the topic. They would penalise or credit a performance for both insight and maturity on both these features. The group finally agreed that some learners clearly understood the topic (showing insight) and provided relevant information on the topic, but that the information did not necessarily contain mature ideas. Also, mature ideas were not necessarily fully developed or structured into well-organised paragraphs.

The most problematic feature seemed to be feature 7, paragraphing. The most severe discrepancies were reported for this feature. One examiner in particular found it particularly problematic, and the panel discussed in detail how this feature should be interpreted. The main difference between feature 3 and feature 7 was the degree of focus on content. Whereas feature 3 explicitly focussed on content and whether ideas were organised in a logical order, feature 7 addressed content to the degree to which it supported the physical or surface structuring of paragraphs. Some of the raters argued that paragraphs served to distinguish new ideas. It would therefore not be practical to assess paragraphing purely in terms of surface structure (i.e. whether paragraphs were clearly visible), since some learners may indicate paragraphs, but are not able to separate ideas accordingly. Effective paragraphing would entail clear organisation of ideas into paragraphs. On the other hand, if, for example, ideas were presented in a logical order, but no paragraphs were indicated, the learner would be penalised on feature 7, but not on feature 3.

The second day of the workshop commenced with the third iteration. It followed the same procedure as the second iteration. Eight performances were scored in three sets containing three, three and two scripts respectively.

After scoring the first three scripts, the trial examiners considered their personal scoring tendencies in comparison with the rest of the group. They were asked whether they were able to recognise certain tendencies in their scoring that differed from that of the other trial examiners. Most raters were able to recognise whether they tended to be harsh or lenient in their scoring, and whether this was for particular features, for overall scores, or both.

One rater in particular reported consistently scoring particular features (such as spelling) harsher than the group, but other features (such as insight into the topic) more leniently. This rater received initial training in English as a first language at secondary school level, which may have influenced her perception of an acceptable standard. Since some features were scored harsher than others, this rater tended to agree with the group on the overall level score. Therefore, the researcher was not too concerned about these tendencies and the panel did not discuss them at length. However, the group more pertinently discussed individual raters' tendencies to consistently disagree on overall performance levels as a result of severe and continuous discrepancies on specific features. They were concerned that discrepant raters were allowing themselves to be influenced by certain aspects of performances other than the feature in question.

There was one rater who reported more lenient scores for most features and in particular for feature 7. The group therefore revised their comments on this feature from the previous day's discussion in order to reach consensus and standardise the particular rater's interpretation and application of the scale. This rater appeared to be influenced more heavily by either the content or the structure of the paragraphs, depending on the particular performance. He struggled to resolve his uncertainty in this regard by himself. In practice, such raters may need additional training.

A few raters reported difficulty in distinguishing between feature 1 and feature 2 (insight into a topic and presenting mature ideas on the topic). They tended to equate maturity of ideas with understanding of the topic and, as a result, penalised performances twice (on both features) when they found ideas to be immature. The features were discussed in order to clarify the focus of each feature, helping raters to achieve a crisper, more distinct understanding of the two. In assessing the first feature, the raters needed to concentrate specifically on insight, without considering maturity.

After scoring three more scripts, raters were in general agreement on overall scores with few and slight differences at feature level. They resolved their differences easily and felt confident about scoring accurately with the scale.

One potentially problematic aspect noted by the panel was that some raters differed by only a few marks from the rest of the group, but that was enough to place the performance at a higher or lower level than its benchmark. For example, they may have assigned a Level 5 to five features, where other raters assigned a Level 4. This difference would result in an overall difference of five marks, which may place a performance at a Level 5 while it is benchmarked at a Level 4. Close comparison of the scores assigned to individual features revealed where such cases had occurred. Generally, the group easily reached consensus on these cases, with raters admitting having been influenced by another aspect of the text, or having been "kind" in assigning a higher level score. Raters therefore seem to benefit from intensive training.

After the panel scored the remaining two scripts of the third batch, they did not report any new difficulties and seemed to agree in their interpretation and application of the scale. The trial examiners felt confident of scoring a larger batch of scripts without discussion, as they would under normal FET circumstances.

During the fourth and final iteration, the trial examiners scored a set of fifteen scripts without any interruption and reported their thought processes in writing directly afterwards. Before scoring commenced, the researcher briefed the group on think-aloud protocol procedures (cf. Green, 1998) and explained what would be expected of them after scoring the scripts. The markers were asked to record their opinions on each performance, their thoughts on the scores they had assigned, and whether and how the scale aided or complicated the scoring process. Owing to practical constraints, the participants provided written protocols as opposed to recorded verbal protocols, but the same procedure was followed as for verbal protocols. Trial examiners were asked to motivate their scores, explain any changes they made to scores, and indicate any uncertainty that they experienced in scoring any particular features. Following the scoring session, they were also asked to complete a questionnaire (Appendix D) without discussing it with other members of the group.

The researcher computerised the scores after each of the four iterations, labelling the data sets Batch 1 to Batch 4, for analysis. The written feedback protocols and questionnaire responses provided in the final iteration were also recorded electronically for analysis.

The data from each of the four batches were used to make calculations as described in Chapter 6 (cf. 6.5.4). Table 7.1 summarises the results for the average inter-rater correlation, Kendall's concordance coefficient and Cronbach's alpha based on it. Table 7.2 reports results for the inter-class correlations per individual rater, generalisability coefficients for the sum of all raters and Cronbach alpha coefficient based on it.

Reliability estimate calculations require complete data, i.e. scores reported by each rater for each feature on every script scored. Not all raters reported scores on every feature for all the scripts scored in the first three iterations. Batches 1, 2 and 3 did not render complete data sets. For the purpose of reliability estimate calculations, therefore, only those cases where all raters provided complete scores were used. Had incomplete data been used, lower reliability estimates would have been reported, and in some cases the calculations would not have been be feasible.

In the following tables (Tables 7.1 and 7.2), the number of observations, in terms of the number of essays (N) and the number of raters (K) used for the particular calculation, are indicated. The number of observations thus does not correspond with the total number of scripts scored during a particular iteration. For example, seven scripts were scored in the first iteration yet complete data were only reported by all raters for four scripts. Data for these four scripts were used for the particular calculation.

Data for Batch 4 were complete for all raters on twelve of the fifteen essays (Batch 4a). Two raters, however, proved inconsistent in their scoring, assigning very low scores when the majority of the group assigned higher scores. They were removed from the data set. Calculations were then repeated for the revised data set (Batch 4b), which contained complete data for all fifteen scripts and are reported alongside the original results for comparison.

Average inter-rater correlation coefficients reported in Table 7.1 demonstrate a decreasing trend for the first three batches (Batch 1 to 3), but an increasing trend for the last three (Batch 3 to 4b).

	Batch 1	Batch 2	Batch 3	Batch 4a	Batch 4b
N	4	4	7	12	15
K	16	17	16	19	17
Average inter-rater correlation	0.90	0.82	0.68	0.82	0.83
Cronbach's alpha	0.97	0.96	0.95	0.98	0.99
Kendall's concordance	0.64	0.63	0.51	0.81	0.90

Table 7.1 Results for reliabilities as calculated for each iteration

These results indicate that inter-rater agreement lessened somewhat during the initial stages of training, but increased steadily in the final phases. This may be a result of the selection of scripts scored in the second and third iteration. Scripts in Batch 2 more distinctly illustrated performances at different levels, making it perhaps easier to distinguish accurately between them than between the scripts in Batch 3, which illustrated adjacent performance levels (Levels 3 and 4). Closer examination of the scores assigned by raters for Batch 3 scripts reveals that raters generally differed within one adjacent level of each other on individual features. As discussed above, small discrepancies at feature level may result in a larger discrepancy in the overall level reported for a performance.

Increasing reliabilities (Cronbach alpha) are reported for Batches 3, 4a and 4b, indicating increasing reliability. For the final calculations, an almost perfect alpha value of 0.99 is reported.

Table 7.2 reports intra-class correlations indicating the degree to which each essay was awarded similar scores by different raters. It also reports generalisaibility estimates indicating the degree to which raters' performances using the scale may be interpreted as representative of raters in general. In other words it provides an answer to the question "Can the results be accepted as indicative of performance for the larger population of raters using the proposed scale?"

	Batch 1	Batch 2	Batch 3	Batch 4a	Batch4b
Intra-class coefficient for individual raters	0.37	0.58	0.30	0.74	0.82
Generalisability for the sum of all raters	0.90	0.97	0.91	0.98	0.99
Cronbach's alpha	0.95	0.97	0.92	0.98	0.99

Table 7.2 Results for inter-class correlation and generalisability coefficient as calculated for each iteration

Table 7.2 presents results for the intra-class coefficients and generalisability coefficients calculated for each batch individually. As noted above, two values are reported for Batch 4, viz. one for the original data (Batch 4a) and another (Batch 4b) after data for seemingly discrepant raters (raters 4 and 8) were removed from the set.

Intra-class coefficients for individual raters indicate the degree to which individual raters' scoring tendencies could be taken as indicative of other raters' tendencies. There is a general increasing tendency in the values, indicating that the individual raters became more standardised, i.e. more consistent in scoring as training progressed. Scores for Batch 3 scripts varied more than those for Batch 2 scripts, but for the final scoring in Batch 4 the high value of 0.82 indicates that raters were in close agreement in the scores they awarded.

High generalisability coefficients were reported for the sum of all raters on each of the four batches, meaning that the performance of the group of markers can be accepted as indicative of typical performances of examiners in general. Values decrease for the first three batches, but increase considerably for Batch 4 calculations. This may indicate that raters' application of the scale becomes more generalisable after training. Very high values are reported for both Batches 4a and 4b, with the coefficient for Batch 4b showing a slight increase as a result of the two inconsistent raters' data being excluded from the set. There is also a general increase in Cronbach alpha values.

The researcher calibrated the data sets for all four batches individually using FACETS Rasch (Linacre, 2006b). Results are reported for individual batches in Figure 7.6 to Figure 7.9. These results provide additional information on raters' scoring tendencies and their application of the rating scale.

Missing data are not problematic when conducting Rasch analyses, unlike for the calculations discussed above. Rasch takes into account and corrects for missing data points and can estimate true scores if some information (e.g. scores reported by some of the raters) is available on each facet. For example, seventeen out of twenty raters scored all fifteen features for all essays. The remaining three raters only scored ten of the fifteen essays, resulting in no data on five essays for these three raters. Rasch uses the available data to estimate how these three raters' were most likely to score the remaining five essays and fills the missing data

points accordingly. Thus, all scripts (N=7, 5, 8 and 15 for the batches respectively) and all raters (K=20) were included in the data used for calibration. The number of scripts indicated in Figures 7.6 to 7.9 therefore corresponds with the total number scored in each iteration.

Rasch calibration results illustrate the spread of raters, essays and features for each individual batch. These results support the results discussed above. In each of the following figures, the raters are placed in the second column, essays in the third column, and the individual features in the fourth column.

A comparison of Figures 7.6 to 7.9 illustrates an increase in rater consistency. As training progresses from the first iteration (Fig. 7.6) to the last iteration (Fig. 7.9), raters and individual features are placed closer together and to the 0 logit mark.

Figure 7.6 below illustrates the results of the calibration conducted for the first batch. The logit scale indicates a range of +2 to -2. This range is fairly small, indicating some level of consistency amongst the raters, which may be attributed to their experience as L2 markers. Raters are spread across the range of -2 to approximately +0.5 on the logit scale, which is an acceptable range for raters who are mostly unfamiliar with the scale and have not undergone standardisation training (training to ensure they apply the scale according to the same standard). However, the raters are not bunched together or placed at the 0 logit mark, which shows distinct variation in rater severity. Raters placed above the 0 logit mark were more lenient in scoring, while those below the 0 logit mark were more severe.

Essays scored during the first iteration were selected to represent each of the seven scale levels, but according to the results in Figure 7.6, the untrained raters did not apply the full range of the scale, in particular the higher extremes on the scale (Levels 6 and 7).

In examining the scores that raters assigned to individual features, the researcher found that some raters assigned Levels 6 and 7 to some features, but no performance as a whole achieved a score at either of these levels. The trial examiners may have been reluctant to assign extreme level scores because they had not yet been trained and did not want to seem too harsh or too lenient in their scoring.

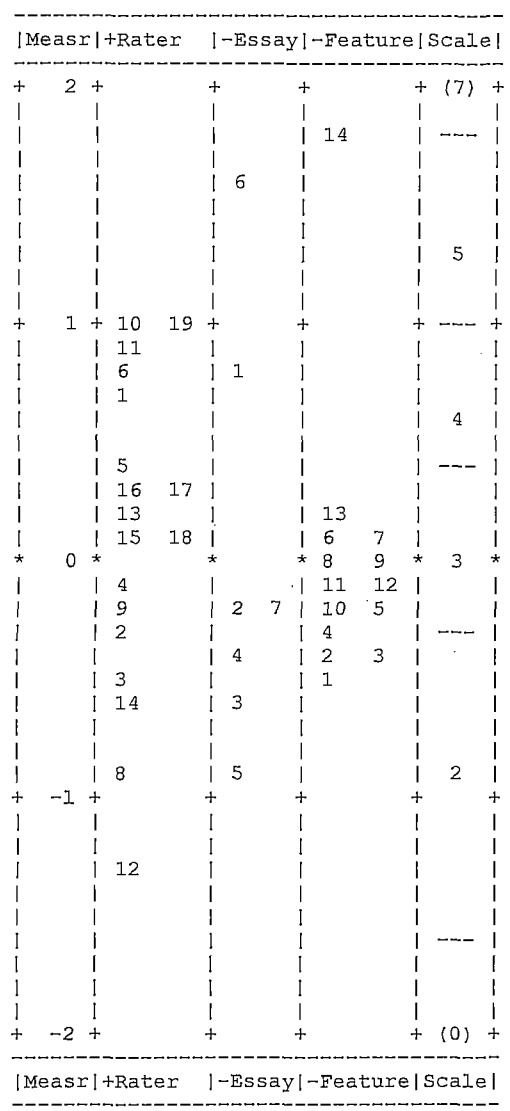

Figure 7.6 Vertical ruler report produced for Batch 1 data after blind scoring

Column four in Figure 7.6 (Feature) indicates raters' scoring tendencies in terms of their harshness or leniency in marking individual features. The placing of feature 14 may indicate that this feature, addressing the effective use of linking devices, may have been scored much more leniently than the other features. All other features are clustered together and close to the 0 logit mark. It may be that raters were not clear about the standard expected in terms of learners' use of linking devices, or that they shared a different understanding of the term than intended.

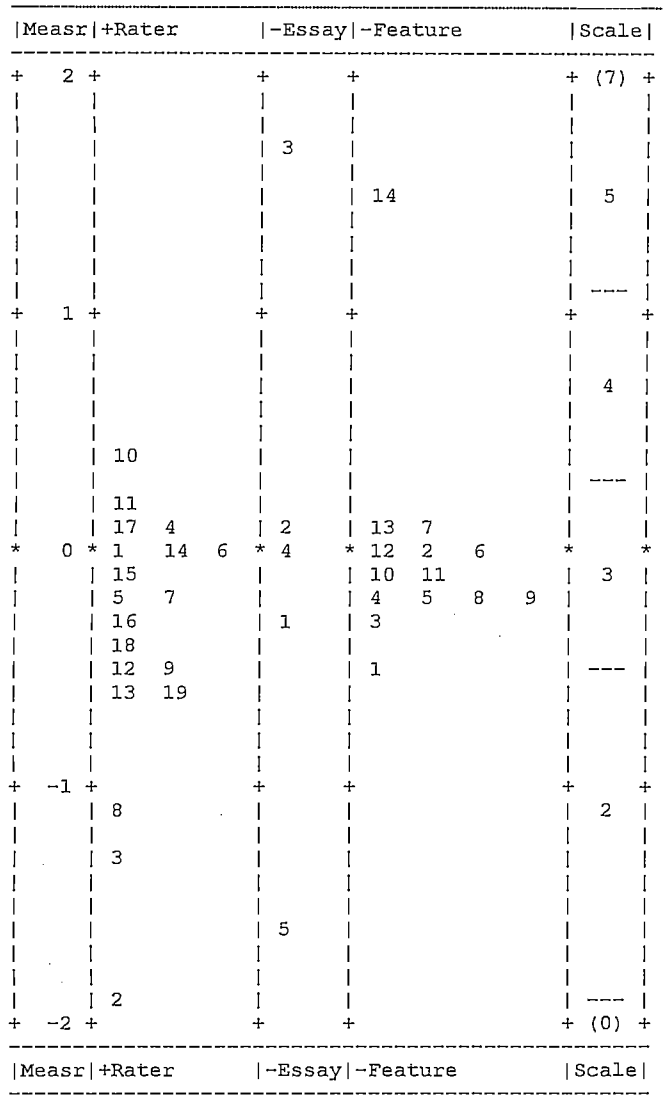

Figure 7.7 Vertical ruler report produced for Batch 2 data after the second iteration

Figure 7.7 reports the result for the calibration of data collected after the second iteration. It shows a much closer distribution of raters than Figure 7.6, indicating that after some discussion and clarification of the scale content, the raters were in closer agreement with each other and scored more consistently as a group. Three raters, viz. 2, 3 and 8, still rated harsher than the others, but fewer raters were notably more lenient in their scoring. The majority of raters form a closer cloud-like cluster around the 0 logit mark, indicating that they were in fair agreement in their scoring. The most severe raters were Raters 2 and 3 and the most lenient was Rater 10. In comparison to the results presented in Figure 7.6, Raters 2 and 3 were more severe scoring the second batch of scripts (placed closer to -1), whereas Rater 10's accuracy in scoring improved from the first iteration (placed closer to 0). The rater

measurement report identified Rater 2 as the only overfitting rater with infit 2.91 and outfit 4.66. All others were within the acceptable range. A high reliability index of 0.95 indicates that there were clear differences between the raters, as expected.

The distribution of the individual features closer to the 0 logit mark indicates that they were scored more consistently than in the previous iteration. Feature 14 remained displaced in relation to the other features, but was not identified as an outlier in the feature measurement report. Fit values for all features were within the acceptable range. A high reliability index of 0.95 was reported for the sample, indicating that the features were clearly distinguished by the trial examiners.

Figure 7.8 reports the results for the calibration of data after the third iteration, following more training and clarification of the scale content. It shows raters clustered together between the 0 and 1 logit marks, indicating closer agreement than that achieved during the first two iterations. However, all the raters are grouped above the 0 logit mark, indicating that they tended to be more lenient in the scores they assigned, and perhaps too much so. This may also provide an explanation for the decrease in the reliability estimate and rank order values from Batch 2 to Batch 3, reported in Tables 7.1 and 7.2.

No outliers were identified in the rater measurement report, all fit values being well within the acceptable range. A high reliability index of 0.87 was reported, but the value is distinctly lower than that reported after the previous two calibrations. In other words, there were less distinct differences between the raters, which may indicate that they were more like-minded in their scoring.

The features are not clustered as closely together as in the previous report, but there is a good balance in the spread of features close to the 0 logit. All fit measures are close to 1, indicating a good fit. The reliability index increased from 0.95, reported for Batch 2, to 0.97 for Batch 3, indicating that the features were distinguished more clearly as training progressed.

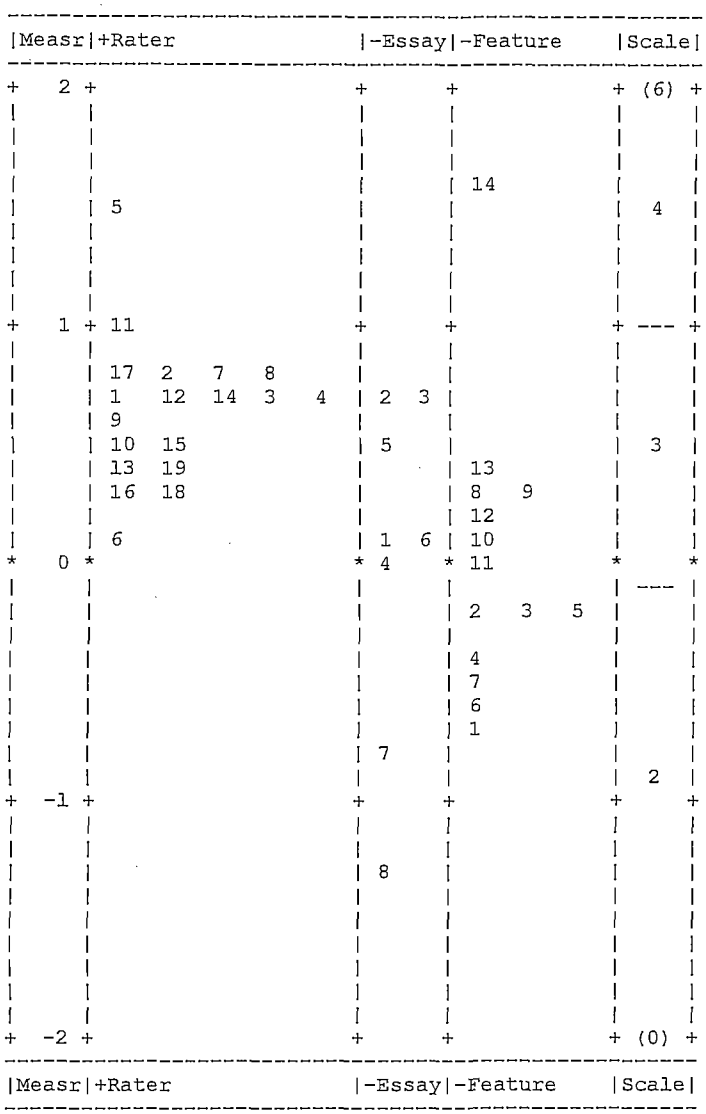

Figure 7.8 Vertical ruler report produced for Batch 3 data after the third iteration

Finally, Figure 7.9 below reports the calibration results of the data collected after the fourth iteration, which was the uninterrupted scoring session. These results indicate that raters were in closer agreement with each other, with the whole group showing a tendency to score more leniently than expected. There is also a more balanced distinction between individual features after training than the results reported above for the first three iterations.

All the raters are grouped closely together at the 0 logit mark, indicating close agreement and accurate scoring. High infit measures of 2.12 and 2.8 respectively for Raters 4 and 12, however, indicated an overfit. In other words, these raters were not consistent in their scoring

and may benefit from additional training. As indicated above, lower correlation coefficients were reported for Raters 4 and 8.

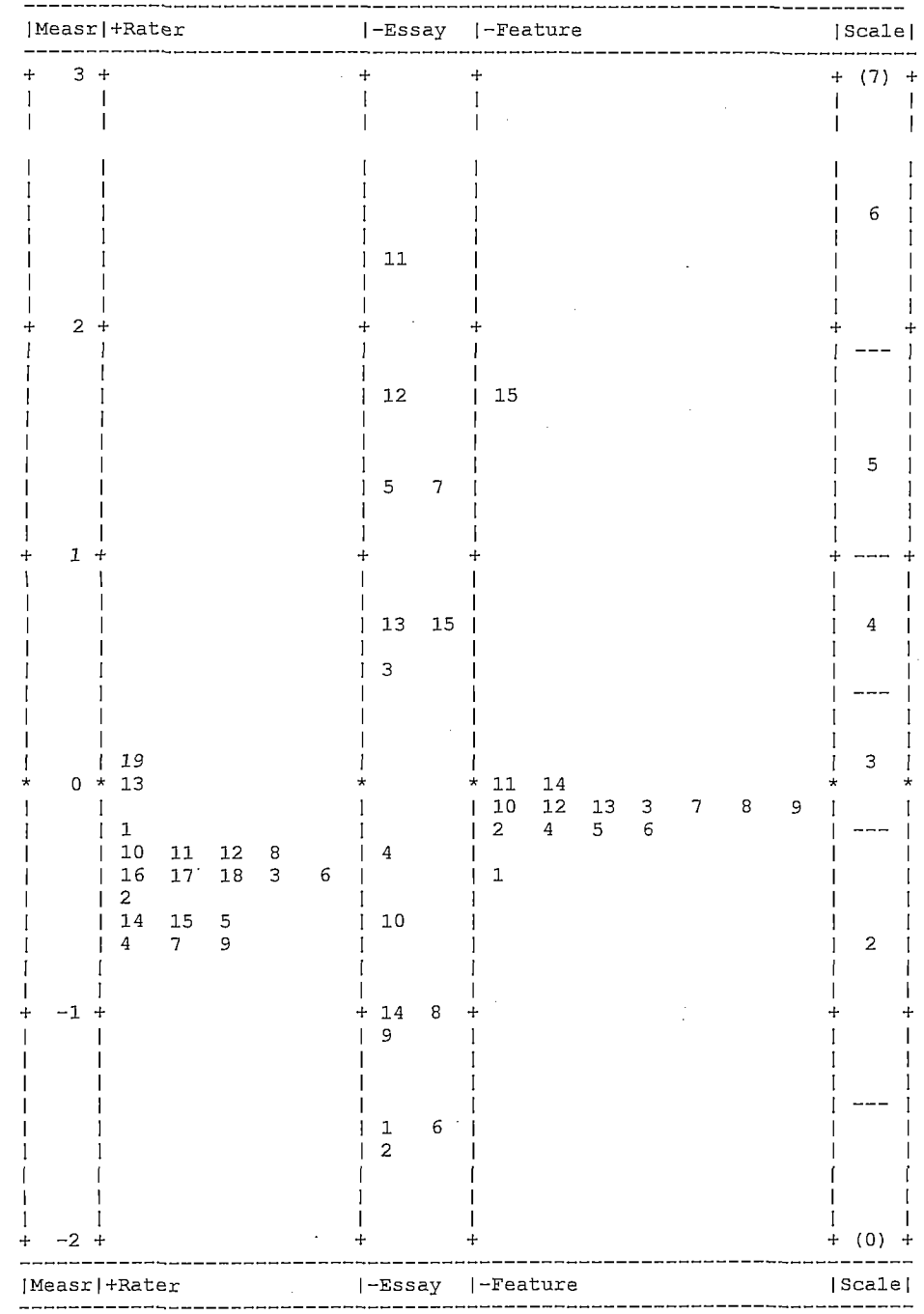

Figure 7.9 Vertical ruler report produced for Batch 4 data after the final iteration

Infit measures for Rater 8, however, are within the acceptable range. Closer examination of the data revealed that Rater 8 consistently assigned much lower scores than the other raters in

the group, but sustained this level of severity for all performances. The rater's self-consistency (albeit discrepant compared to other raters) may explain why the rater was not identified as an outlier in the Rasch results. A high reliability index of 0.96 was reported for the sample of raters, indicating that there were distinct differences between the raters' scoring.

All features, apart from F15, are placed in close proximity at the 0 logit mark. F14 no longer seems to be problematic. The results indicate no severe bias towards any of the features and show that the features are related to each other and to the same construct. The high reliability index of 0.86 indicates that the features were clearly distinguished.

The significance of the results reported for reliability estimates and Rasch analyses can be summarised as follows:

- After the first iteration, the raters, albeit untrained and unfamiliar with the scale, seemed to share some degree of like-mindedness in their initial interpretation and application of the scale. This may indicate that the scale presents criteria and features that are sufficiently distinct and clear.
- A comparison of the results across the four iterations reveals an increase in inter-rater agreement. After each session of standardisation training, raters were more consistent in their application of the scale. This may indicate that training helped to clarify certain features and standardise interpretation and application of the scale.
- Four iterations were sufficient to increase inter-rater agreement notably, but some raters may need further standardisation.
- After the fourth iteration, a high level of inter-rater agreement was reached, with raters being able to distinguish clearly between features and apply them accurately.
- The performance of this sample of examiners using the proposed scale can be generalised, i.e. interpreted as representative of the larger population of FET examiners.

After conducting the Rasch calibrations, the researcher analysed the results of the protocols recorded during scoring, as well as the examiner questionnaire retrospectively completed. The results are summarised below in Table 7.10.

In recording the protocols, the examiners generally reported their reactions to the different performances, revealing some information about their thoughts in applying the scale. Three main types of responses emerged during the analysis, and the researcher categorised the comments accordingly, viz. strengths and positive responses, weaknesses and suggestions, and comments evaluative of the scale features. Below are some of the (unedited) comments made by the raters:

Strengths and Positive Responses

- It became simpler and lighter to finalise the levels and score range to the essay as the marking was done and completed; focus was on specific aspects to pay attention to during marking process.
- This was very easy to score but I think somewhere a person is afraid of scoring thinking that you have become too lenient or too harsh
- The scoring is becoming much easier, even though one feels that some features like paragraphing are difficult to score.
- The scale to this [performance] helped a lot. It was easy to score even though I believe that I might have been too lenient on some features.
- I struggled to give scores as I think I might be too strict. The scale helped me a lot in this regard.
- No difficulties in giving a final score as guidance and directive in marking are outlined in the rating scale.
- The rating scale guided me in properly and correctly allocating a correct score range without unnecessarily being unjust and harsh to the candidate.
- Nothing was found to be appropriate and in accordance with the expectations of the rating scale, so the decision was an early and justifiable one as the score range resulted from what the rating scale clearly demonstrated to us.
- I found no difficulties when it came to the score range for the essay as this was easily informed and supplemented by aspects given under each level.
- I had a tough time marking the entire essay as everything is a mess. I had to try and make sense out of every sentence used, but because of the clearly outlined rating scale I used I didn't find it difficult or a problem to score under each level.
- It made the marking and allocation of marks justifiable as every aspect to be considered when marking is clearly tabled under each level. Marks are not just haphazardly allocated, hence the maintenance of correctness and standard in marking.
- Content: The learner seems not to have insight in the topic. No flow of ideas. The rating scale helped to recognise this. It helped a great deal.

There were thus many positive comments on the clarity of the scale and the clear guidance it provided during scoring. Some comments indicated raters' frustration during scoring, revealing aspects of the scale that could be adjusted, such as the score levels from 1 to 7,

without offering an option to score 0. The following comments reveal, in particular, raters' frustration in trying to score poor performances:

Weaknesses and Suggestions

- It is difficult to score [this essay] as it is poorly presented. This makes one not to understand really what is discussed. If on the rating scale a 0 mark was given, there are some features I strongly believe that 0 would have been appropriate. I have given 1 to those.
- Please review marks on Length. The essay looks long though I am not sure. I should have scored a 1 for length if accommodated by the scale, so I scored a 2.
- This essay, its content does not feature in this rating scale therefore it was a little bit tough to allocate this learner marks under Features 1-5. I could realise that there is no way in which a learner cannot be allocated a mark under this rating scale because the essay was very horrible and very hard to understand.
- Really there was no sense at all when marking this type of an essay because one could not clearly know and think of what was it all about. Therefore I've decided to allocate most of the marks for each feature under poor performance. But with this rating scale learners get marks even when there is nothing at all that they have written.
- It [the essay] was totally Greek as it was a difficult also for me to allocate a mark above 2 for each feature according to the scale. This evenly proved to me that according to the rating scale there is no learner who can get a mark below seven because the features rangers from 1-14 including the length therefore learners are advantaged by the scale.
- This [essay] is horrible without the topic it is difficult to know if he/she is relevant. The words have no meaning as such no proper sentences were constructed. The syntax is worse, no vocabulary, you could say the learner does not know what to write. This score is just a generous gift. I think this is where the scoring scale is disadvantaged, because it is just the same as written nothing however the learner should get something, because he/she wrote something.
- Length - more flexible? Level descriptors could help, e.g. how many errors can still be very good? Linking devices - what about no use at all?
- Level descriptors could help, e.g. how many errors can still be very good.
- ...[s]light use of linking devices. Grammar - can't comprehend his intended meaning. Only just scraping by into Level 2 - possibly undeservedly. Too short.
- What is this person saying? Is this English? Level 2, only because I feel sorry! Rather Level 1. Almost impossible to give a mark. The meaning of this essay is lost. Even reading it out loud wouldn't elucidate it! No planning, no concept of sentence structure. Strictly speaking, many of the criteria should (could) have gotten a 0 but since the lowest mark is 1, okay then 1. Criteria 1 & 2 may be marked with 2 if one really searches for sense. Level 2, but Level 1 is more honest.

Raters particularly found the score ranges of the scale from 1 to 7 limiting because they could not assign a Level 0 to very poor performances. The scale therefore advantaged very poor learners unfairly in this regard; even an unintelligible garble could not be scored less than fourteen marks out of one hundred. This could result in some poor performances being allocated a higher level than warranted; for example, being awarded a Level 2 when Level 1 is more appropriate. They also indicated their need for a middle range Level 1 to score length.

This would allow them to score fairly good performances that are slightly too long or too short. Since these suggestions were also evident during training, the proposed scale was adjusted accordingly (cf. Appendix E).

Two of the raters indicated a need for labelled band levels to indicate the standard at each of the seven performance levels, but this need was not raised prominently in the group. This raised a complicated problem. Descriptors were not provided for all adjacent levels in an attempt to avoid possible confusion as a result of ambiguous terms. Some raters may be uncertain when making judgements regarding the "in-between" levels. However, the majority of the trial examiners used in this exercise (eighteen out of twenty) did not indicate a need for descriptors at each of the seven levels. During training, the raters were also reminded of the two-step decision making approach they were encouraged to follow in using the scale. Most raters seemed to find this approach useful and implemented it in their own decision-making processes. A pertinent focus on this approach to scoring during training may help less experienced raters to become more confident in distinguishing unlabelled levels. Considering that only two of the twenty raters suggested additional descriptors as a possible adjustment to be considered, the scale was not adjusted in this way.

Finally, many comments pertained to raters' thoughts on essay performances, reflecting their evaluation of individual features. Some excerpts are provided below:

Evaluative Thoughts

- [This essay is] Difficult to rate - is there any meaning?
- Rater changed his/her mind on feature 11 from 2 to 1; comment: ideas are immature but original.
- This essay is worth a 6, but individual feature paragraphing gave me a problem as paragraphs are not visible. I ended up giving a 3 even though the whole essay is good. The language is not up to standard, the spelling and sentence structures are poorly used.
- Feature 5: I wanted to give 2 but I scored 3 because the general meaning is outlined that women are better drivers than men, though the learner struggled to construct sentences. Feature 6: I scored 6 because there's an introduction but I could not make meaning from reading the last paragraph so no conclusion. Feature 11: I settled for 1 because there are lots of spelling errors & punctuation.
- Content: The learner seems not to have insight in the topic. No flow of ideas. The rating scale helped to recognise this. It helped a great deal. Structure & style: no sense could be made on the essay because of poor language usage.
- Changed mark for feature 12 from 3 to 2. I don't know if the student had insight into the topic because the topic wasn't specified and I can't deduct the topic from the essay. Spelling is not too bad - punctuation lacking. Slight use of linking devices. Grammar - can't comprehend his intended meaning. Only just scraping by into Level 2 - possibly undeservedly. Too short.

- The essay was in accordance with the requirements of the rating scale and the learner wrote the arguments showing a flow of ideas coherently. The introduction, body and conclusion were contained main ideas supported by sentences in the paragraphs. The learner's essay deserves to be rated a Level 6.

- The paragraphing is very bad, a single sentence is not enough to make a paragraph. The learner uses full matured sentences with well constructed ideas. There are a number of spelling mistakes and punctuation. However in general this is a good essay because it is thought provoking, though it's narrative, you can be taken on imagery and he successfully does that. He used a variety of range and tone.

- I feel that the content was well presented as the learner showed good insight into the topic and was able to sustain it to the end. The essay was engaging and points were relevant to the topic. Structure & Development: The intro, body & conclusion were not well presented - unclear. Though what seems like paragraphs contain different ideas, paragraphs are not clear. Grammar: The use of syntax, tense and concord etc. are above adequate. Vocabulary: satisfactory and above adequate. Candidate adhered to required length.

Such comments are valuable because they provide information on raters' perceptions of performances in terms of the criteria and give insight into their decision-making. These comments reveal that the trial examiners generally followed the scale systematically and assessed performances in terms of the different aspects addressed in the scale, rather than using the scale to justify their own impressionistic judgements. The protocols also reveal that raters seem to recognise differences in learners' development pertaining to individual features, i.e. a learner may perform better in some areas than in others. The proposed scale therefore seems to help raters consider this typical aspect of L2 learner writing.

The responses provided on the examiner questionnaire support the findings from the statistical analyses and the protocol analysis. Overall, the participants reported positive experiences using the scale and considered it to be an improvement on the current scale. They generally considered the content and wording of the scale to be clear, easy to understand, specific and user-friendly, although they indicated that some features may need a little clarification during training. The scale guided them to consider each script in detail, assess comprehensively and with a specific focus. The responses also indicated that the group felt the scale made scoring more objective, guiding them to assess each performance in detail.

After completing the final iteration, each trial examiner also completed a questionnaire (Appendix D) on their experience using the scale (cf. 6.5.4).

The explanatory scale guide proved valuable as supporting material. It served as a key to the scale, clarifying and guiding the trial examiners in interpreting and applying the scale

accurately and consistently. It should be integrated as standard support material to be used with the scale during scoring.

Table 7.4 below reports the general responses of the trial examiners to each question and provides some unedited examples of typical responses provided. For each question, excerpts from typical comments provided by individual raters are provided to illustrate their general attitude on that particular aspect.

1. What is your general reaction to the new approach to scoring texts?

In general the examiners indicated a positive attitude and experiences using the proposed scale.

- Positive, marker is assisted in rating, guided as to what to consider.
- The rating scale effects correctness and maintenance of standard with regard to the marking of essays
- I believe it excellence because it covers all the aspects in marking an essay and gives you a better understanding of rating learners accordingly.
- Fairly well structured, a bit difficult to use at the beginning but I believe with time and constant reference one can get used to it. I in general believe it is user friendly
- I found the scale a great help in narrowing down several issues compared to the present grid used in schools, especially the Grammar (F8-11). The scale also helps to separate the different aspects like content development so that one doesn't penalise the learner for the same error over and over. The issue of planning is also taken up into the scale with the Structure and Style & Length. If the learner hasn't planned, these marks will be low. Suggestion: Change Length 0-1-2 to differentiate between a good essay, but a little too long or somewhat short.
- After marking approximately 15 scripts it got easier to use the rating scale. It is more flexible overall and covers a wide range of criteria. This could perhaps benefit learners more than the older rubrics.
- Although it initially takes getting used to, and it seems time consuming, the overall result is much better! Criteria are clearly set out and by dividing content and language into "smaller parts" it makes the allocation of a mark much easier.

Question	Indication of agreement or disagreement	General perception and listed examples of typical comments provided by the respondents (unedited)
2. Are you confident about rating accurately using the proposed scale?	Most agreed	Some uncertainty remains, but generally raters felt comfortable that they would get use to the scale with practise. Other: • Still worry if I'm being too lenient or too strict
3. Do you find the revised scale more difficult to use than the current scale?	All disagreed	Although it needs getting used to, the scale is simple and straight forward. • It is very specific and simple • As soon as you get used to the sequence it flows easily. • The scale is not difficult as long as you are familiar with the features • The previous one [current scale] only focussed on two [general] features i.e. the content and language but now the new one is diverse. • I've used other scales and I can't imagine they're as

		comprehensive as this one.
4. How well do you understand the criteria? (are they clear?)	Most agreed	• Fairly clear, practice will make perfect • They are clear, we just have to get used to it, everything is clearly pointed out. • The criteria is very clear and easily understood • The criteria were clear enough but I think markers would need a guide as to what warrants a 1 or 2 or 3 etc. • Very well- they are very clear and also serve as guidelines • They are better than the methods we use and it gives an understanding of rating a letter correctly • Good, paragraphing posed a few problems but it is clear now. • The explanatory piece and the discussion made everything easier to understand.
5. Do you find the criteria to be clearer in comparison to the current scale?	All agreed	• It is specific. • It allows certain aspects to be dealt with which were not dealt with the current one. • Each category is clearly explained.
6. Do you think the features in the proposed scale are clear?	All agreed	Additional clarification is needed. • Some might need more clarification, e.g. F10 & F14.
7. Do you think the features under each criterion address relevant aspects?	Most agreed	• In a rubric some of the things are difficult to solve, but this scale will make one's work easier than in a rubric. • For example, under content, difficult features are clearly outlined and it is easy for an educator to score. • I think it's a good breakdown and students can score in places where otherwise they have been penalised. Other: • [One rater indicated no, but did not provide any commentary.]
8. Do you find it easy to mark the five criteria?	Most agreed	Criteria are clear and distinct. • [E]very feature is clearly explained and you'll know what or how can you score. • The 5 criteria are easy to mark • It was a little difficult to start, not letting the one criteria affect the other but eventually it gets easy enough • They are more like guidelines • It is fairer towards the candidate but restrictions might be fewer • Very clear • Most have always been marked in general before Other: • No, because when there is no cohesion an coherence it is going to be difficult for one to allocate marks for the learners.
9. How adequate is the new scale wording?	Most considered it adequate	Generally no problems, but additional level descriptors were suggested. • Well worded. • The wording is simple and easy to understand. • Very clear and to the point. • Simple and easy to read. Other: • Some more level descriptors as guide.
10. Does the revised scale capture the essential qualities of the written performance?	All agreed	Comprehensive • Learners were disadvantages with the current one and teachers confused in terms of how to reach out for a mark looking at the language and content at the same time.

		• Every feature is explained well, it gives you a clear understanding of what and how it going to be assessed. • All the aspects to be considered are clearly given.
11. Can you distinguish all the band levels clearly and interpret them consistently?	Most agreed	Some raters need to distinguish between the seven levels and the fifteen features. • They are clearly explained. • ...with timeous reference to the scale guide when marking the essay. • The levels seem to reflect the overall impression the essay left on the marker. • ...though sometimes consistency is subjective. Other: • Levels 4 and 5, Level 13 and 14 seem to address one thing, maybe [it would be better] if they were grouped together.
12. Can you use all seven levels distinguished on the scale? (Do you think that seven levels are enough or too many?)	Most agreed	The seven levels are enough • They are enough for the maintenance of consistency and standard in marking. • They are enough because they address all the requirements of essay writing. • They are enough because we are able to exhaust all that is needed. • Because they emphasise the rating levels used nationally • I don't think it should be any less. • Seven levels are enough. The wording/descriptor of each level could perhaps change, e.g. Use ones on current scale - average, below average, meritorious. Other: • I think they are too many and require one to be too much focussed.
13. Do you think the scale increases the reliability of subjectively judged ratings?	All agreed	• Because the scale makes you not guess the score but to know what and how much to give for [a performance]. • An educator should be subjective when marking but give scores reliably so the scale helps because [it] specifies all the categories and addresses them individually. • Definitely. • It helps standardise marking.
14. Can you interpret effectively any "relative" language terms, e.g. 'adequate', 'limited'. If not, are there any words or phrases you cannot interpret?	Most agreed	Generally raters find it easy, but some questions remain. • The mentioned words/ terms 'adequate'; 'limited' are obvious to be interpreted. • the language used in the scale is descriptive and easy to interpret • The wording is simple and familiar. Other: • How good is "very good" [e.g. to assign a Level 5 or 6]? • The wording/descriptor of each level could perhaps change, e.g. use ones on current scale - average, below average, meritorious. • No such as poor - how poor is a thing?
15. Do you always confine yourself exclusively to the context of the scale? If not, what else influences your judgements?	Most agreed	Raters generally seem to base decisions on the scale, rather than use the scale to support their own judgements, but sometimes they are influenced by other factors. • The scale is detailed enough not to need anything else to make a judgement • I try as hard as possible because it is more accurate • I'm convinced and impressed by everything in the scale • Take note not to be influenced by e.g. spelling when insight is judged - not always easy

		Other: • The general impression when reading the learners task. • The way in which the learners write sometimes - they write Greek and this influences my judgement. • The instructions can influence one's judgement • Your gut feeling? Experience?
16. Do you concentrate on one criterion at a time and read the response specifically for that criterion?	All indicated usually or sometimes	Raters find that they can consider different criteria or features simultaneously while reading through the essay once or twice. 1. I read the essay as a whole and then re-read it (scanning) when considering each criterion.
17. Do you read the whole essay before you start scoring individual criteria?	All indicated always or usually	Most raters seem to read scripts more than once when scoring. 2. All the way through any item perceived 1 score
18. Do you consider all the criteria simultaneously while scoring?	All indicated usually or sometimes	Cf. Q16 3. One after the other.
19. How does the application of the revised scale affect the time it takes to rate a full script?	Mostly positive	It takes longer, but gets faster with practice; the additional time is worth the effort. • It takes a bit longer. • It takes more time but it's worth it. • Good, once you get used to it, it becomes easy and quicker and less time-consuming. • I'd think it was faster than the current scale, because there are specific things to consider- is not so much left to your own personal opinion. Other: • It takes too much time because it is also too long.
20. Do you consider the current and proposed scales to be of the same standard?	Most disagreed	Proposed scale is generally considered to be more detailed and specific, thus more advanced. • The current scale is more advanced and score range is very good/ fair • The one we have been using for research purposes [proposed scale] is user friendly and justifiable with regard to scoring in particular • The proposed scale is much better than the current one because it is clear and simple • The previous scale did not have enough and was too restricted, the proposed one is better. Other: • Depends on the commitment of the marker. • Yes the aim is the same as the benefit is given to a learner.
21. Do you believe the candidates would get the same score on the current and the proposed scale?	Half agreed and half disagreed	Various factors may influence learners' scores. The general attitude indicates that raters think learners will receive fairer scores on the proposed scale than on the current scale. • [T]he new scale [should] be introduced to schools. • The candidates would not get the same score, the proposed scale is very good as it is specific. • There is no way in which learners are going to fail if ever we as educators also show then the scale for them to know when writing that they are scored in this fashion • The current one did not give learners a lot of marks and the proposed one will up the level of the learners • I think they have a chance of scoring a bit better. • They might get a better more justifiable mark. • No. The categories are marked separately and allocation of marks is different.

		• Currently candidates are scored on content & language holistically - therefore poor spelling might influence the language mark so that syntax, tense & concord etc. are not fairly assessed.
22. Do you think that the proposed scale helps you to evaluate performances more critically?	All agreed	Raters indicated that the scale provides guidance and aids focussed and systematic scoring. • One is obliged to read a learners essay with understanding [and scoring]. • ...because the scale encompasses and covers all the critical aspects to be borne in mind when evaluating and marking essays. • It helps because more [aspects] is involved and considered • Definitely- and more objectively.
23. Does the proposed scale make you aware of possible personal bias that you may have when scoring?	Most agreed	Raters believed that the scale helped prevent personal bias. Raters may be influenced by the personal content of performances. • With the old/ current scale one will just give a general mark without even thinking. • ...because everything is clearly outlined so as to cap off biasness when evaluation is done. • There is no way in which one can be biased when scoring using the scale. • [I]t clearly indicates how you should score... • and helps you to eliminate them • [I]t allows me to go through each script and score learners according to their deserving mark. Other: • [I am influenced] Especially in cases I find very offending or against my principles.

24. What do you feel is the main advantage of / what do you think is the best asset of the proposed scale?

Raters indicated that the scale made for more accurate and fair marking because learners were rewarded for their attempts. The format of the scale was also considered an improvement on that of the current scale.

- Credit and penalty can be given where due, not too much.
- It is advantageous to the learners [because it credits them fairly].
- The way in which it is structured, unlike the current one.
- It's detailed and separating the criteria and marking them independently is definite a huge advantage and a fair advantage at that for the learners
- It is more accurate, it identifies the aspects individually and allocates score for each aspect
- It identifies specific areas to focus on, whereas the previous scale presented a general vague description
- It is easier to pinpoint problem areas and one allocates a more justifiable mark.
- Learners will benefit more. Educators will be forced to look at all the issues involved.
- Learners have a fair chance of getting a just mark for their efforts.

25. What do you think is the main disadvantage of or what is your main concern regarding the proposed scale?

The most disadvantages noted were that performances could not be assigned a 0 mark

and that the length feature could not be scored a middle range score, i.e. 1. The additional time it took to score was also noted, but not as frequently as originally expected by the researcher.

- The disadvantage is that even learners who did not write any concrete essay are given marks
- There are no marks allocated for the length even if the learner has written quite a bit he/ she is not recognised in terms of this
- My main concern is awarding a learner marks even though he/she is irrelevant but meet the requirements, does not look fair
- It takes time but very helpful
- Probably that you can't score a zero
- Marker may lose focus on specific area/
- It's a bit time- consuming
- It may take longer to mark- more aspects- but not as long as the ancient scale where errors (minor and major) had to be counted! 2.
- Restrictions will be scarcer for this grid since [it indicates insufficiencies] which the other didn't.
- Inexperienced markers might struggle at first. [More] Good and clear descriptors should accompany the scale.

26. If you could change anything in the proposed scale, would you? If so, what would you change?

Raters mostly felt that they did not want to change much, but requested that the scoring assigned for length be revised, and that additional descriptors be indicates for the Levels 3 and 5.

- More info on length, perhaps more clarity ion physical paragraphing
- I would change the scores in length only
- I feel there is nothing to change for now, except that it should be given chance and the results will inform what can be done
- Perhaps only the general comment insert between poor reasonable adequate good... excellent.

Figure 7.4 Trial examiner questionnaire results summarised

7.6 Conclusion

Following the overview of the method of research provided in Chapter 6, the aim of this chapter was to describe the process followed to develop an empirically validated rating scale for assessing English FAL writing at Grade 12 level. It comprised four phases, each with a particular aim contributing to the overall aim of the study. Each phase was discussed in terms of the methods used to collect data and the analyses conducted. Both quantitative and qualitative methods were used during the phases as appropriate for the purpose of each phase. Results were reported and discussed accordingly.

The final product was a multiple trait rating scale comprising five criteria, addressing fifteen individual features. The results indicated that the proposed scale could be implemented successfully as a valid instrument for measuring English FAL writing at Grade 12 level. The

statistical procedures indicated high levels of inter-rater and intra-rater consistency in applying the scale, as well as high generalisability. Raters proved positive about the scale, indicating that they considered it a more accurate and fairer instrument than the current rating scale.

CHAPTER 8

Conclusion

8.1 Introduction

As a conclusion to the study, this chapter synthesises the various aspects discussed in the preceding chapters. This study aimed to develop an empirically validated rating scale for assessing second language writing at Grade 12 level in the final Further Education and Training examination. This chapter reviews and evaluates the outcome of the study, discusses its limitations and makes recommendations for further research.

8.2 The development of the rating scale

Validation is a continuous process of arguing for or against a claim that a particular instrument is valid within a particular context and for a specific purpose. It involves collecting and evaluating evidence to support – or counter – this argument. This study therefore amounts to a validation argument that the proposed scale is viable for its intended purpose in its particular context. Assessment is faced with increasing demands for accountability, and in a high-stakes examination such as that in the final Grade 12 one, the crucial question is whether we can depend on the scores which result from the examination.

The literature review in this study discussed the concept of validity and the issue of developing and validating a rating scale for measuring writing ability of second language speakers. Modern interpretations of validity regard it as a unified concept comprising different types of validity. Construct validity is generally accepted as an overarching term, and regarded as a function of the interaction between various types of validities.

Shaw and Weir's (2007) socio-cognitive framework was adopted in this study, as it focuses specifically on the validation of writing assessment instruments. The interactionalist nature of the framework lies in the interpretation of the construct as residing in the interaction between cognitive ability and the context of use (Shaw & Weir, 2007:3). Shaw and Weir (2007) regard construct validity as a function of the interaction between cognitive validity, context validity and scoring validity. A valid rating scale contributes to scoring validity, since it

reflects the construct being assessed (directly or indirectly) and influences the reliability of scores. Scoring validity was therefore the central concern of this study.

A validation argument comprises a claim (or claims) accompanied by evidence to support it. A variety of methods can be used to investigate claims and provide evidence either in support of or against claims that an instrument renders accurate scores that can be interpreted validly. As many means as possible should be used to collect different types of evidence to strengthen the validation argument. This study adopted an empirical approach to the development of the scale, and both quantitative and qualitative data were collected and analysed to ensure a sound argument and provide evidence to support the claim that the scale is valid for its intended purpose.

A combination of expert judgements and statistical analyses was used to establish the degree of validity at each phase of the development of the scale and provide supporting evidence for the validation argument. The empirical validation procedure entailed collecting and benchmarking samples of learner writing, analysing them and constructing a scale that addresses the salient features of learner writing identified as typical of each of the seven performance levels.

Once the benchmarks had been established, a new scale was drafted in a series of validation steps. The content of the draft scale was then subjected to a number of evaluations for verification of the criteria and sub-features addressed in the scale. Assessment Standards stipulated in the National Curriculum Statement (2005) served as checklist for the features identified by the panel of experts. In the following phase, the scale was refined to ensure that the micro-level features addressed under each criterion were relevant, self-explanatory and that the descriptors were formulated clearly.

In order to estimate the extent to which the scale could be implemented successfully in practice, the draft scale was piloted. Following training, a panel of experienced FET examiners conducted a trial scoring exercise using the scale. Quantitative and qualitative data were collected to provide potential evidence for the validity of the proposed scale. The various analyses indicated that the trial examiners succeeded in applying the scale consistently. The results also emphasised the importance of training in the use of a rating scale. Feedback indicated that the examiners experienced the scale positively and believed it

to be an accurate instrument which would promote fair assessment. Their evaluation of the content of the scale also indicated that the scale was likely to carry validity at face value. The explanatory scale guide developed to define criteria and explain the features addressed in the scale proved useful as support material for training and to guide raters during scoring.

Results from the series of quantitative and qualitative analyses conducted in this study suggest that the proposed rating scale offers an accurate means of scoring second language essays. The evidence collected in this study therefore supports the argument that the proposed scale is valid and can be used to assess writing at Grade 12 level.

Thus, in developing and justifying, through a process of validation, a validated multiple trait rating scale for assessing second language writing at Grade 12 level, this study has had a positive outcome.

8.3 Limitations of the study

For the purpose of the present study, the proposed scale was piloted on a small scale. The results indicated generalisability, but further piloting of the scale may be necessary with larger groups of examiners, in order to determine whether this generalisability indeed holds, and, if so, to what degree.

This study focussed on developing a rating scale for assessing learners' writing abilities in essay-type responses as assessed in Section A of the FET Writing examination. The scope of the project was therefore limited and did not include assessing transactional writing assessed in Sections B and C.

The scale was developed and validated for assessment at Grade 12 level in the South African FET examination context. Its relevance for assessment purposes in other examination contexts, viz. grades other than Grade 12 or other national or international examinations, may therefore be limited, but not unimportant.

Finally, although the scale was piloted and results indicated generalisability, the full impact that it may have in the FET examination context on the accuracy, consistency and reliability of scores can only be investigated once the scale has been implemented.

8.4 Recommendations for further study

A large-scale piloting project can be undertaken to investigate the potential success of the proposed scale further before it is fully implemented.

Further research could be conducted in order to establish a valid rating scale for scoring responses to Sections B and C of the FET Writing examination. The proposed rating scale may serve as starting point for such a study.

Ideally, learners' writing ability should be assessed in a uniform manner from lower grade levels to provide a profile of their development from one grade level to the next. A scale of similar format and content (if relevant) would be useful in providing continuous diagnostic information on the development of learners' writing ability across grade levels. Studies could therefore be undertaken to adapt and validate the proposed scale for assessing writing ability at other grade levels, or to develop similar scales for assessment at these levels.

The proposed scale may also lend itself to application at first-year tertiary level. Additional study would be necessary in order to determine whether the scale would be valid in such a context. Validation studies similar to the present study could be undertaken to develop rating scales relevant to an academic writing construct.

If the scale were fully implemented, impact and washback studies could be conducted to investigate the consequential validity of the proposed scale.

8.5 Conclusion

Test scores serve as indication of an individual's ability to perform in a certain area or complete a task, or the degree to which he or she possess a particular skill. Stakeholders use scores to make various decisions that affect the individual directly and could affect other stakeholders. It is therefore the responsibility of those who develop testing instruments to ensure that the instrument is valid for the purpose of the assessment in a particular context. The instrument should render fair and accurate scores to provide an accurate indication of a person's ability.

The current rating scale used to score Grade 12 FET Writing examination responses has not been empirically validated and may therefore provide stakeholders with a skewed indication of learners' writing abilities. This study entailed an extensive empirical validation of a rating scale for measuring English FAL at Grade 12 level. The proposed rating scale offers a valid alternative. It is therefore proposed that the scale be implemented for assessing essay writing at Grade 12 level.

BIBLIOGRAPHY

AERA, APA and NCME. 1974. See American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

AERA, APA and NCME. 1985. See American Educational Research Association, American Psychological Association, National Council on Measurement in Education

AERA, APA and NCME. 1999. See American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

ALDERSON, J.C. 1983. The use of cloze procedure and proficiency in English as a foreign language. (*In* Oller, J.W. Jr. *ed.* Issues in language testing research. Rowley, MA: Newbury).

ALDERSON, J.C. 1991. Bands and scores. (*In* Alderson, J.C. and North, B. *eds*. Language testing in the 1990s. London: Macmillan).

ALDERSON, J.C. 2000. Assessing reading. Cambridge: Cambridge University Press.

ALDERSON, J.C. and Windeatt, S. 1991. Lancaster University Computer-Based Assessment System (LUCAS). Lancaster: Department of Linguistics and Modern English Language, Bowland College, Lancaster University.

ALDERSON, J.C. and Banerjee, J. 2002. Language testing and assessment (Part 2). *Language Teaching,* 35: 79-113.

ALDERSON, J.C. and BUCK, G. 1993. Standards in testing: a study of the practice of UK Examination Boards in EFL/ESL testing. *Language Testing,* 10(2): 1-26.

ALDERSON, J.C., CLAPHAM, C. and WALL, D. 1995. Language test construction and evaluation. Cambridge: Cambridge University Press.

AMERICAN PSYCHOLOGICAL ASSOCIATION. 1954. Standard for Educational and Psychological Testing. Washington, DC: AERA.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYHOLOGCIAL ASSOCIATION, NATIONAL COUNCIL on MEASUREMENT in EDUCATION. 1974. Standard for Educational and Psychological Testing. AERA, Washington, DC: American Psychological Association.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYHOLOGCIAL ASSOCIATION, NATIONAL COUNCIL on MEASUREMENT in EDUCATION. 1985. Standard for Educational and Psychological Testing. Washington, DC: AERA.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYHOLOGCIAL ASSOCIATION, NATIONAL COUNCIL on MEASUREMENT in EDUCATION. 1999. Standard for Educational and Psychological Testing. Washington, DC: AERA.

ANASTASI, A. 1976. Psychological testing. 4th ed. New York: Macmillan.

ANDERSON, N.J., BACHMAN, L., PERKINS, K. and COHEN, A. 1991. An exploratory study into construct validity of a reading comprehension test: triangulation of data sources. *Language Testing,* 8:41-66.

ANDERSON, T. 2003. Modes of interaction in distance education: Recent developments and research questions. (*In* M. Moore., *ed.* Handbook of Distance Education. Mahwah, NJ.: Erlbaum).

APA. 1954. See American Psychological Association.

BACHMAN, L.F. 1990. Fundamental considerations in language testing. Oxford: Oxford University Press.

BACHMAN, L.F. 2002. Some reflections on task-based language performance assessment. *Language Testing*, 19(4):453-476.

BACHMAN, L.F. 2004. Statistical analyses for language assessment. Cambridge: Cambridge University Press.

BACHMAN, L.F. 2005. Building and supporting a case for test use. *Language Assessment Quarterly,* 2(1):1-34.

BACHMAN, L.F. and PALMER, A.S. 1979. Convergent and discriminant validation of oral language proficiency tests. (*In* Silverstein, B., *ed.* Proceedings of the Third International Conference of Frontiers in Language Proficiency and Dominance Testing. Carbondale, Ill: Department of Linguistics, Southern Illinois University).

BACHMAN, L.F. and PALMER, A.S. 1981. A multi-trait method investigation into the construct validity of six test of speaking and reading. *(In* Palmer, A.S., Groot, A.S.P.J. and Trosper, F.A., *eds.* The construct validity of communicative competence. Washington, DC: TESOL).

BACHMAN, L.F. and PALMER, A.S. 1982. The construct validation of some components of communicative proficiency. *TESOL Quarterly,* 16(4):449-465.

BACHMAN, L.F. and PALMER, A.S. 1996. Language testing in practice. Oxford: Oxford University Press.

BARKHUIZEN, G. 2005. Social influences on language learning. (*In* Davies, A. and Elder, C., *eds*. The handbook of applied linguistics. Malden, MA: Blackwell Publishing).

BARKAOUI, K. 2007. Rating scale impact on EFL essay marking: a mixed method. *Assessing Writing,* 12(2):86-107.

BECK, S.W. and JEFFERY, J.V. 2007. Genres of high-stakes writing assessments and the construct of writing competence. *Assessing Writing,* 12(1):60-79.

BEREITER, C. and SCARDAMALIA, M. 1987. The psychology of written composition. Hillside, NJ: Lawrence Erlbaum.

BERNSTEIN, B. 1990. Class, codes and control 4: the structuring of pedagogic discourse. London and New York: Routledge.

BLOMMAERT, J., MUYLLAERT, N., HUYSMANS, M. and DYERS, C. 2005. Peripheral normativity: the production of locality in a South African township school. *Linguistics and Education,* 16:378-403.

BORSBOOM, D., MELLENBERGH, G.H. and VAN HEERDEN, J. 2004. The concept of validity. *Psychological Review,* 111(4):1061-1071.

BORSBOOM, D., VAN HEERDEN, J. and MELLENBERGH, G.J. 2003. Validity and truth. [Available online:] http://users.fmg.uva.nl/dborsboom/BorsboomTruth2003.pdf Date accessed: 15 November 2006.

BRENNAN, R.L. 1998. Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice,* 17(1):5-9.

BRIGGS, D.C. 2004. Comment: making an argument for design validity before interpretive validity. *Measurement: Interdisciplinary Research and Perspectives,* 2(3):171-174.

BRINDLEY, G. 1998. Outcomes-based assessment and reporting in language learning programmes: a review of the issue. *Language Testing,* 15(1):45-85.

BROWN, G.T.L., GLASWELL, K. and HARLAND, D. 2004. Accuracy in the scoring of writing: studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing,* 9(2):105-121.

BROWN, J.D. and HUDSON, T. 2002. Criterion-referenced language testing. Cambridge: Cambridge University Press.

BRUALDI, A. 1999. Traditional and modern concepts of validity. *ERIC Digests* [Available online:] http://www.ericdigests.org/2000-3/validity.htm Date of access: 22 November 2006.

CANALE, M. and SWAIN, M. 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics,* 12(1):1-47.

CARR, N. 2000. A comparison of the effects of analytic and holistic composition in the context of composition tests. *Issues in Applied Linguistics,* 11(2):207-241.

CARROLL, J.B. 1968. The psychology of language testing. (*In* Davies, A., *ed.* Language Testing Symposium: a psycholinguistic approach. London: Oxford University Press).

CEF. 2001. See Council of Europe.

CHALHOUB-DEVILLE, M. 2003. Second language interaction: current perspectives and future trends. *Language Testing,* 20(4):369-383.

CHAPELLE, C. 1998. Construct definition and validity inquiry in SLA research. (*In* Bachman, L.F and Cohen, A.D., *eds.* Interfaces between second language acquisition and language testing research. Cambridge: Cambridge University Press).

CHAPELLE, C.A. 1999. Validity in language assessment. *Annual Review of Applied Linguistics,* 19:254-272

CHERRYHOLMES, C.H. 1988. Power and criticism: poststructural investigation in education. New York: Teachers College Press.

CLARK, J.L.D. 1972. Foreign language testing: theory and practice. Philadelphia: The Centre for Curriculum Development.

COHEN, J. 1960. A coefficient for agreement for nominal scales. *Educational and Psychological Measurement,* 20(1):37-47.

COOPER, M. 1986. The ecology of writing. *College English,* 4:364-375.

COOPER, C.R. and ODELL, L. 1977. Evaluating writing: describing, measuring, judging. National Council of Teachers of English. Urbana, IL: State University of New York at Buffalo.

COUNCIL OF EUROPE. 2001. Common European framework of reference for languages: learning, teaching, assessment. Cambridge: Cambridge University Press.

CRAMER, D. and HOWITT, D. 2004. The SAGE dictionary of statistics. London: SAGE. [Available online:] http://books.google.co.za/books?id=8BwXzD-pKC4C&printsec=frontcover Date accessed: 28 April 2009.

CROCKER, L. 2003. Teaching for the test: validity, fairness, and moral action. *Educational Measurement: Issues and Practice*, 22(3):5-11.

CRONBACH, L.J. 1971. Test validation. (*In* Thorndike, R.L., *ed.* Educational Measurement. 2nd ed. Washington DC: American Council of Education).

CRONBACH, L.J. 1980a. Selection theory for a political world. *Public Personnel Management,* 9:37-50.

CRONBACH, L.J. 1980b. Validity on parole: how can we go straight? New directions for testing and measurement: Measuring achievement over a decade. Proceedings of the 1979 ETSI invitational Conference. San Francisco: Jossey-Bass.

CRONBACH, L.J. 1988. Five perspectives on validity argument. (*In* Wainer, H. & Braun, H., *eds.* Test Validity. Hillside, NJ: Lawrence Erlbaum Associates, Inc.).

CRONBACH, L.J. and MEEHL, P.E. 1955. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281-302.

CUMMING, A. 1990. Expertise in evaluating second-language compositions. *Language Testing,* 7(1):31-51.

CUMMING, A., KANTOR, R. and POWERS, D. 2001. Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: an investigation into raters' decision making, and development of a preliminary analytic framework. TOEFL Monograph Series. Princeton, NJ: Educational Testing Service.

CUMMING, A., KANTOR, R. and POWERS, D. 2002. Decision making while assessing ESL/EFL writing: A descriptive framework. *Modern Language Journal,* 86(1):67-96.

CUMMINS, J. 2000. Language, power and pedagogy. Clevedon: Multilingual Matters.

CURETON, E. E. 1951. Validity. (*In* Linquist., *ed.* Educational measurement. Washington, DC: American Council of Education).

DAVIDSON, F. 1991. Statistical support for training in ESL composition rating. (*In* Hamp-Lyons, L., *ed.* Assessing second language writing in academic contexts. Norwood, NJ: Ablex).

DAVIES, A. 1990. Principles of language testing. Oxford: Basil Blackwell.

DAVIES, A., BROWN, A., ELDER, C., LUMLEY, K. and McNAMARA, T. 1999. Dictionary of language testing. Studies in Language Testing Series, 7. Cambridge: Cambridge University Press/ UCLES.

DAVIES, A., and ELDER, C. 2005. Validity and validation in language testing. (*In* Hinkel, E. *ed.* Handbook of research in second language teaching and learning. Lawrence Erlbaum, Mahwah, NJ).

DEPARTMENT OF EDUCATION (DoE) Examination Guidelines. 2009. See South Africa.

DYERS, C. 2004. Ten years of democracy: attitudes and identity among some South African school children. *Per Linguam,* 20(1):22-35.

DYERS, C. 2008. Truncated multilingualism or language shift? An examination of language use in intimate domains in a new non-racial working class township in South Africa. *Journal of Multilingual and Multicultural Development,* 29(2):110-126.

DOUGLAS, D. 2000. Assessing language for specific purposes. Cambridge: Cambridge University Press.

DOUGLAS, D. 2001. Language for specific purposes assessment criteria: where do they come from? *Language Testing,* 18(2):171-185.

EBEL, R. 1961. Must all tests be valid? *American Psychologist,* 16:640-647.

EBEL, R. 1983. The practical validation of tests of ability. *Educational Measurement: Issues and Practice,* 2(2):7-10.

EBEL, R.L. and FRISBIE, D.A. 1991. Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall.

ECKES, T. 2005. Examining rater effects in TestDaF writing and speaking performance assessment: a many-facet Rasch analysis. *Language Assessment Quarterly,* 2(3):197-221.

ELDER, C. 1993. How do subject specialists construe classroom language proficiency? *Language Testing,* 10(3):235-254.

ELDER, C. 2005. Individual feedback to enhance rater training: does it work? *Language Assessment Quarterly,* 2(3):175-196.

ELLIS, R. 1994. The study of second language acquisition. Oxford: Oxford University Press.

EMBRETSON, W.S. 1983. Construct validity: construct representation versus nomothetic span. *Psychological Bulletin,* 93(8):179-197.

EMBRETSON, W.S. 2007. Construct validity: a universal validity system or just another test evaluation procedure? *Educational Researcher,* 36(8):449-455.

ENGELHARD, G. 1992. The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education,* 5(3):171-191.

ENGELHARD, G. 1994. Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement,* 31(2):93-122.

ERDOSY, M.U. 2004. Exploring variability in judging writing ability in a second language: a study of four raters of ESL compositions. TOEFL Research Report RR-03-17. Princeton, NJ: Educational Testing Service.

FALVEY, P. & SHAW, S. 2006. IELTS writing: revising assessment criteria and scales (Phase 5). *Research Notes,* 23:7-12.

FIELD, J. 2004. Psycholinguistics: the key concepts. London: Routledge.

FREMER, J. 2000. Promoting high standards and the "problem" with construct validation. *NCME Newsletter,* 8(3):1.

FULCHER, G. 1987. Test of oral performance: the need for data-based criteria. *English Language Teaching Journal,* 41(4):287-291.

FULCHER, G. 1993. The construction and validation of rating scales for oral tests in English as a Foreign Language. Lancaster: University of Lancaster (thesis - Ph.D.).

FULCHER, G. 1996. Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing,* 13(2):208-238.

FULCHER, G. 1999. The communicative legacy in Language Testing. *Applied Linguistics,* 28(4):483-497.

FULCHER, G. 2000. The 'communicative' legacy in language testing. *System,* 28(4):483-497.

FULCHER, G. 2003. Testing second language speaking. Harlow: Pearson.

FULCHER, G. and DAVIDSON, F. 2007. Language testing and assessment: an advanced resource book. Abingdon, Oxon: Routledge.

GARSON, G. D. 2006. Validity. *Statnotes: topics in multivariate analysis.* [Available online:] http://www2.chass.ncsu.edu/garson/pa765/validity.htm Date accessed: 18 May 2008.

GOODWIN, L.D. and LEECH, N.L. 2003. The meaning of validity in the new standards for educational and psychological testing: implications for measurement courses. *Measurement and Evaluation in Counselling and Development,* 36:181-191. [Available online:] http://web.ebscohost.com/ehost/pdf?vid=2&hid=13&sid=22c39f1a-5d2b-4b72-bbfb-a67293685588%40sessionmgr102 Date accessed: 12 May 2008.

GRABE, W. and KAPLAN, R.B. 1996. Theory and practice of writing: an Applied Linguistics perspective. New York: Longman.

GREEN, A. 1998. Verbal protocol analysis in language testing research: a handbook. Cambridge: Cambridge University Press.

GORIN, J.S. 2007. Reconsidering issues in validity theory. *Educational Researcher,* 36(8):456-462.

GUILFORD, J.P. 1946. New standards for test evaluation. *Educational and Psychological Measurement,* 6(5):427-439.

GUION, R.M. 1980. On trinitarian doctrines of validity. *Professional Psychology,* 11(3):385-398.

GÜLES, N. 2005. The struggle for English. *Sunday Times: insight & opinion,* 9 January:3.

GUMPERZ, J.J. 1992. Contextualisation and understanding. (*In* Duranti, A. and Goodwin, C., *eds*. Rethinking context: language as an interactive phenomenon. Cambridge: Cambridge University Press).

HAMBLETON, R.K., SWAMINATHAN, H., ALGINA, J. and COULSON, D.B. 1978. Criterion-referenced testing and measurement: a review of technical issues and developments. *Review of Educational Research,* 48(1):1-47.

HAMP-LYONS, L. 1990. Second language writing: assessment issues. (*In* Kroll, B., *ed.* Second language writing: research insights for the classroom. Cambridge: Cambridge University Press.

HAMP-LYONS, L. 1991a. The writer's knowledge and our knowledge of the writer. (*In* Hamp-Lyons, L., *ed.* Assessing second language writing in academic contexts. Norwood, NJ: Ablex).

HAMP-LYONS, L. 1991b. Scoring procedures for ESL contexts. (*In* Hamp-Lyons, L., *ed.* Assessing second language writing in academic contexts. Norwood, NJ: Ablex).

HAMP-LYONS, L. 1995. Rating nonnative writers – the trouble with holistic scoring. *TESOL Quarterly,* 29(4):750-762.

HAMP-LYONS, L. 2007. Worrying About Rating. *Assessing Writing,* 12(1):1-9.

HARRISON, A. 1983. A language testing handbook. London: Macmillan.

HAWKEY, R. and BARKER, F. 2004. Developing a common scale for the assessment of writing. *Assessing Writing,* 9(2):122-159.

HAWKEY, R. 2006. Impact theory and practice: studies of the IELTS Test and Progetto Lingue 2000. Cambridge: Cambridge University Press.

HAYES, J.R. 1996. A new framework for understanding cognition and affect in writing. (*In* Levy, C.M. and Ransdell, S., *eds*. The science of writing. NJ: Lawrence Erlbaum).

HAYES, J.R. and FLOWER, L.S. 1980. Identifying the organization of writing processes. (*In* Gregg, L.W. and Steinberg, E.R., *eds.* Cognitive Processes in Writing. Hillsdale, NJ: Lawrence Erlbaum).

HAERTEL, E.H. 1999. Validity arguments for high-stakes testing: in search of the evidence. *Education Measurement: Issues and Practice*, 18(4):5-9.

HAERTEL, E.H. 2004. Interpretive argument and validity argument for certification testing: can we escape the need for psychological theory? *Measurement: Interdisciplinary Research and Perspectives*, 2(3):175-178.

HENNING, G. 1987. A guide to language testing. Cambridge, MA: Newbury House.

HENNING, G. 1991. Issues in evaluating and maintaining an ESL writing assessment program. (*In* Hamp-Lyons, L., *ed.* Assessing second language writing in academic contexts. Norwood, NJ: Ablex).

HOUSE, E.R. 1980. Evaluating with validity. Beverly Hills, CA: Sage.

HUBLY, A.M. and ZUMBO, B.D. 1996. A dialect on validity: where have we been and where are we going? *Journal of General Psychology*, 123. [Available online:] http://www.questia.com/googleScholar.qst?docId=76927046 Date accessed: 12 May 2008.

HUDSON, T. 2005. Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics.* 25:205-227.

HUGHES, A. 1989. Testing for language teachers. Cambridge University Press: Cambridge.

HUGHES, G. F. 1996. The need for clear purposes and new approaches to the evaluation

of writing-across-the-curriculum programs. (*In* White, E. M., Lutz, W.D, and Kamusikiri, S., *eds*. Assessment of writing: politics, policies, practices. New York: Modern Language Association of America).

HUGHES, A. 2003. Testing for language teachers. Cambridge: Cambridge University Press.

HUGHES, A., PORTER, D. and WEIR, C.J. 1988. Validating the ELTS test: a critical review. Cambridge: The British Council and the University of Cambridge Local Examination Syndicate.

HUOT, B. 1990. Reliability, validity and holistic scoring: what we know and what we need to know. *College Composition and Communication,* 41(2):201-213.

HUOT, B. 1996. Toward a new theory of writing assessment. *College Composition and Communication,* 47(4):549-566.

HYLAND, K. 2002. Teaching and researching writing. London: Longman.

HYMES, D.H. 1972. On communicative competence. (*In* Pride, J.B. and Holmes, J., *eds*. Sociolinguistics. Harmondsworth: Penguin).

IELTS. 2007. IELTS Handbook. University of Cambridge: ESOL examinations. [Available online:] http://www.cambridgeesol.org/assets/pdf/resources/IELTS_Handbook.pdf
Date accessed: 14 March 2009.

INOUE, A.B. 2005. Community-based assessment pedagogy. *Assessing Writing: An International Journal* 3(9):208-238.

IVANIC, R. 2004. Discourses of writing and learning to write. *Language and Education,* 18(3): 220-245.

JACOBS, H., ZINGRAF, S.A., WORMUTH, D.R., HARTFIELD, V.F. and HUGHEY, J.B. 1981. Testing ESL compositions: a practical approach. Rowley, MA: Newbury House.

JONES, N. 1979. Performance testing and second language proficiency. (*In* Briere, E.J. & Hinofotis, F.B., *eds.* Concepts in language testing: some recent studies. Washington, DC: TESOL).

JONES, N. 2001. Reliability as one aspect of test quality. *Research Notes,* 4:2-5.

KANE, M. 1990. An argument-based approach to validation. *ACT Research Report Series,* 90(13):1-44.

KANE, M. 1992. An argument-based approach to validity. *Psychological Bulletin,* 112:527-535.

KANE, M. 2002. Validating high-stakes testing programs. *Educational Measurement,* 38(4):319-342.

KANE, M. 2004. Certification testing as an illustration of argument-based validity. *Measurement,* 2(3):135-170.

KANE, M. 2005. Validation. (*In* Brennan, *ed.* Educational measurement. 4th ed. New York: American Council on Education and Praeger).

KANE, M., CROOKS, T.J. and COHEN, A.S. 1999. Validating measures of performance, *Educational Measurement: Issues and Practice*, 18(2):5-7.

KAPLAN, R.B. 1972. The Anatomy of Rhetoric: prolegomena to a functional theory of rhetoric. Philadelphia: Centre for Curriculum Development.

KAPP, R. 2004. 'Reading on the line': an analysis of literacy practices in ESL classes in South African Township schools. *Language and Education*, 18(3):246-263.

KOBAJASHI, H. and RINNERT, C. 1996. Factors effecting composition evaluation in an EFL context: cultural rhetorical pattern and reader's background. *Language Learning,* 46(3):397-437.

KONDO-BROWN, K. 2002. A FACET-analysis of rater bias in measuring Japanese L2 writing Performance. *Language Testing,* 19:3-31.

KNOCH, U., READ, J. and VAN RANDOW, J. 2007. Re-training Writing Raters Online: How does it compare with face-to-face training? *Assessing Writing,* 12(2):26-43.

KRAMSCH, C. 2005. Language, thought and culture. (*In* Davies, A. and Elder, C., *eds.* The handbook of Applied Linguistics. Malden, MA: Blackwell).

LADO, R. 1961. Language testing: the construction and use of foreign language tests. New York: McGraw Hill.

LADO, R. 1957. Linguistics across cultures: applied linguistics for language teachers. Ann Arbor: University of Michigan Press.

LANDY, F.J. 1986. Stamp collecting versus science: validation as hypothesis testing. *American Psychologist,* 41:1183-1192.

LANE, S. 1999. Validity evidence for assessments. [Available online:] www.nciea.org/publications/ValidityEvidence_Lane99.pdf Date accessed: 23 October 2006.

LANGUAGE PROGRAMME GUIDELINES. 2008. See South Africa.

LEE, Y. 2005. Demystifying validity issues in language assessment. *ALAK Newsletter* October. [Available online:] http://www.alak..or.kr/2_piblic/2005_oct/article3.asp Date accessed: 17 March 2008.

LINACRE, J.M. 1998. Many-faceted Rasch Measurement. Chicago: MESA Press.

LINACRE, J. M. 2006a. A User's Guide to FACETS Rasch-Model Computer Programs. [Software User Manual]. Facets for Windows version 3.61.0.

LINACRE, J.M. 2006b. FACETS Rasch Measurement Computer Program Facets for Windows version 3.61.0. Chicago: Winsteps.com.

LISSITZ, R.W. and SAMUELSEN, K. 2007. A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher,* 36(8):463-469.

LOEVINGER, J. 1957. Objective tests as instruments of psychological theory. *Psychological Reports Monograph,* 3(9):635-694.

LOHNIGER, H. 2009. Correlation. *Fundamentals of statistics.* [Available online:] http://www.statistics4u.info/fundstat_eng/wrapnt_correlation170.html. Date accessed: 4 May 2009.

LUMLEY, T. 2000. The process of the assessment of writing performance: the rater's perspective. Department of Linguistics and Applied Linguistics. The University of Melbourne (thesis - Ph.D.).

LUMLEY, T. 2002. Assessment criteria in a large-scale writing-test: what do they really mean to the raters? *Language Testing,* 19(3):246-276.

LUMLEY, T. and BROWN, A. 2005. Research methods in language testing. (*In* Hinkel, E., *ed.* Handbook of research in second language teaching and learning. Mahwah, NJ: Lawrence Erlbaum).

LUMLEY, T. and McNAMARA, T. 1995. Rater characteristics and rater bias: implications for training. *Language Testing,* 12(1):54-71.

LUMLEY, T., Lynch, B.K. and McNAMARA, T. 1994. A new approach to standard setting in language assessment. *Melbourne Papers in Language Testing,* 3(2):19-40.

LUOMA, S. 2004. Assessing Speaking. Cambridge: Cambridge University Press.

LYNCH, B.K. 2003. Language Assessment and Programme Evaluation. Edinburgh: Edinburgh University Press.

MAFISA, P.J. 2008. Improving the teaching and learning of ESL in under performing schools of the North-West province. Potchefstroom: NWU (thesis - Ph.D.).

MARCOULIDES, G.A. 2004. Conceptual debate in evaluating measurement procedures. *Measurement: Interdisciplinary Research and Perspectives*, 2(3):182-184.

McCABE, R. V. 2007. Development and application of evaluation criteria for tertiary in-house EAP materials. Potchefstroom: NWU (thesis - Ph.D.).

McCOLLY W. 1970. What does educational research say about judging of writing? *Journal of Educational Research,* 64(4):148-156.

McKAY, P. 2000. On ESL standards for school-aged learners. *Language Testing,* 17(2):185-214.

McNAMARA, T. 1990. Item Response Theory and the validation of an ESP test for health professionals. *Language Testing*, 7(1):52-75.

McNAMARA, T. 1996. Measuring second language performance. New York: Longman.

McNAMARA, T. 2000. Language testing. Oxford: Oxford University Press.

McNAMARA, T. 2003. Book review: fundamental considerations in language testing. Oxford: Oxford University Press. Language testing in practice: designing and developing useful language tests. *Language Testing,* 20(4):466-473.

McNAMARA, T. and Roever, C. 2006. Language testing: the social dimension. (*In* Young, R., *ed.* Language learning monograph series. Oxford: Blackwell).

MESSICK, S. 1975. The standard problem: meanings and values in measurement and evaluation *American Psychologist,* 30(10):955-966.

MESSICK, S. 1980. Test validity and the ethics of assessment. *American Psychologist,* 35(11):1012-1027.

MESSICK, S. 1981. Evidence and ethics in the evaluation of tests. *Educational Research,* 10(9):9-20.

MESSICK, S. 1983. The psychology of educational measurement. *Journal of Educational Measurement*, 21(3):215-237.

MESSICK, S. 1988. The once and future issues of validity: assessing the meaning and consequences of measurement. (*In* Wainer, H. & Braun, H., *eds.* Test validity. Hillsdale, NJ: Lawrence Erlbaum).

MESSICK, S. 1989. Validity. (*In* Linn, R., *ed.* Educational Measurement. New York: Macmillan).

MESSICK, S. 1992. Validity of test interpretation and use. (*In* Alkin, M.C., *ed.* Encyclopaedia of Educational Research. New York: Macmillan).

MESSICK, S. 1996. Validity and washback in language testing. *Language Testing,* 13(4):241-257.

MILANOVIC, M., *ed.* 1998. ALTE Multilingual glossary of language testing terms. *Studies in Language Testing*, 6. Cambridge: UCLES/Cambridge University Press.

MILANOVIC, M and WEIR, C.J. 2007. Series Editor's Note (*In* Shaw, S.D. & Weir, C.J. Examining writing: research and practice in assessing second language writing. Cambridge: Cambridge University Press).

MISLEVY, R.J. 1996. Test theory reconceived. *Journal of Educational Measurement,* 33(4):379-416.

MIYATA-BODDY, N. & LANGHAM, C.S. 2000. Communicative language testing – an attainable goal? The British Council, Tokyo. [Available online:] www.kasei.ac.jp/library/kiyou/2000/5.LANGHAM.pdf Date accessed: 28 August 2008.

MOLLER, A.D. 1982. A study in the validation of proficiency of tests of English as a Foreign Language. University of Edinburgh (thesis - Ph.D.).

MOORE, H. 1996. Telling what is real: competing views in assessing English as a second language. *Linguistics and Education,* 8(2):189-228.

MOSS, P. 1992. Shifting conceptions of validity in educational measurement: implications for performance assessment. *Review of Educational Research,* 62(3):229-258.

MOSS, P. 2007. Reconstructing validity. *Educational Researcher,* 36(8):470-476.

MULLER, J. 2005. The challenge of cognitive demand. Umalusi and Centre for Higher Education Transformation Seminar: 23 June. *Matric: What should be done?* [Available online:] http://www.umalusi.org.za/ur/research/ Date accessed: 19 February 2009.

MULQUEEN, C., and BAKER, D.P. 2002. Pilot instructor rater training: the utility of the multifacet Item Response Theory model. *The International Journal of Aviation Psychology,* 12(3):287-303.

MYFORD, C.M. and WOLFE, E.W. 2003. Detecting and measuring rater effects using many-faceted Rasch measurement: Part I. *Journal of Applied Measurement,* 4(4):386-422.

NATIONAL RESEARCH COUNCIL (NRC). 2002. See Shavelson, R. and Towne, L., *eds.*

NATIONAL CURRICULUM STATEMENT (NCS). 2005. See South Africa.

NDABA, S. 2005. Halos and horns: reliving constructions of matric performance in the South African education system. *Matric: What should be done?* Umalusi and Centre for Higher Education Transformation Seminar: 23 June. [Available online:] http://www.umalusi.org.za/ur/research/ndaba.pdf Date accessed: 19 February 2009.

NORTH, B. 2000. The Development of a Common Framework of language proficiency. New York: Peter Lang.

NORTH, B. and SCHNEIDER, G. 1998. Scaling descriptors for language proficiency scales. *Language Testing,* 15(2):217-263.

OECD. 2008. See Organisation for economic co-operation and development.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. 2008. Reviews of national policies for education: South Africa. Paris: OECD Publishing. [Available online:] http://books.google.com/books?id=WlqOgc2YwMsC&printsec=frontcover&source=gbsViewAPI#v=onepage&q=&f=false Date accessed: 16 May 2009.

OLLER, J.W. 1979. Language tests at school: a pragmatic approach. Longman: Longman.

O'SULLIVAN, B. 2000. Toward a model of performance in oral testing. University of Reading (thesis - Ph.D.).

O'SULLIVAN, B. 2005. Evidence-based validation: what makes a test 'specific'. Specificity Workshop Handout - ALTE Cardiff (November). London: Roehampton University.

O'SULLIVAN, B. 2006. Issues in business English testing: the BEC revision project. Cambridge: Cambridge University Press.

OSMAN, R., COCKCROFT, K. and KAJEE, A. 2008. English second language (ESL) students as new members of a community of practice: some thoughts for learning and assessment. *Per Linguam,* 24(1):1-10.

PHAKITI, A. 2003. A closer look at the relationship between cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing,* 20(1):26056.

PHAKITI, A. 2008. Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. *Language Testing,* 25(2):237-272.

POLLIT, A. 1991. Giving students a sporting chance: assessment by counting and judging. (*In* Alderson, J.C. & North, B., *eds*. Language testing in the 1990s. London: Macmillan).

POPHAM, W.J. 1981. Modern educational measurement. Englewood Cliffs: Prentice Hall.

PURPURA J.E. 1999. Learner strategy use and performance on language tests: a structural equation modelling approach. Cambridge: Cambridge University Press.

PURPURA, J.E. 2004. Assessing grammar. Cambridge: Cambridge University Press.

READ, J. 1990. Providing relevant content in an AEP Writing Test. *English for Specific Purposes,* 9(2):109-121.

RIMMER, W. 2006. Measuring grammatical complexity: the Gordian knot. *Language Testing*, 23(4):497-519.

RIMMER, W. 2008. Book review. Examining writing: research and practice in assessing second language writing (Studies in Language Testing 26). *TESL Electronic Journal,* 11(4): March. [Available online:] http://tesl-ej.org/ej44/r3.pdf Date accessed: 4 May 2009.

RYAN, J.M. 2002. Issues, strategies and procedures for applying standards when multiple measures are employed. (*In* Tindall, G. and Haladyna, T.M., *eds*. Large-scale assessment programs for all students: validity, technical adequacy and implementation. Mahwah, New Jersey and London: Lawrence Erlbaum).

ROZEBOOM, W.W. 1966. Foundations of the theory of prediction. Homewood, IL: Dorsey.

SUBJECT ASSESSMENT GUIDELINES. 2008. See South Africa.

SAS. 2005. SAS Institute Inc., SAS OnlineDoc®, Version 9.1, Cary, NC.

SASAKI, M. and HIROSE, K. 2004. Development of an analytic rating scale for Japanese L1 writing. *Language Testing,* 16(4):457-478.

SAVILLE, N. 2001. Test development and revision. *Research Notes,* 4:5-8.

SAVILLE, N. 2002. IELTS The test development process for CELS. *Research Notes,* 9:8-10.

SAVILLE, N. 2003. The process of test development and revision within UCLES EFL. (*In* Weir, C, and Milanovic, M., *eds.* Continuity and innovation: revising the Cambridge Proficiency in English Examination 1913-2002. Cambridge: Cambridge University Press).

SAVILLE, N. 2004. The ESOL test development and validation strategy. Internal discussion paper. Cambridge: Cambridge University Press and Cambridge ESOL.

SCHARTON, M. 1996. The politics of validity. (*In* White, E. M., Lutz, W. D. and Kamusikiri, S., *eds.* Assessment of writing: Politics, policies, practices. New York: Modern Language Association of America).

SCHAEFER, E. 2008. Rater bias patterns in an EFL writing assessment. *Language Testing,* 25(4):465-493.

SHAVELSON, R. and TOWNE, L. *eds.* 2002. NATIONAL RESEARCH COUNCIL (NRC). Scientific research in education. Committee on Scientific Principles for Education Research. Washington, DC: National Academy Press

SCHILLING, S.G. 2004. Conceptualising the validity argument: an alternative approach. *Measurement: Interdisciplinary Research and Perspectives*, 2(3):178-182.

SCHMITT, N. 2005. Lexical resources in main suite examinations. Cambridge UCLES internal report. Cambridge: Cambridge University Press.

SCHMITT, N., SCHMITT, D. and CLAPHAM, C. 2001. Developing and exploring behaviour of two new versions of the Vocabulary Levels Test. *Language Testing,* 18(1):55-88.

SCHOONEN, R. 2005. Generalisability of writing scores: an application of structural equation modelling. *Language Testing,* 22(1):1-30.

SHAVELSON, R.J. and WEBB, N.M. 1991. Generalisability theory: a primer. Newbury Park, CA: Sage.

SHAW, S.D. 2002a. The effect of training and standardization on rater judgement and inter-rater reliability. *Research Notes,* 8:13-17.

SHAW, S.D. 2002b. IELTS writing: revising assessment criteria and scales (Phase 2). *Research Notes,* 10:10-13.

SHAW, S.D. 2004a. IELTS Writing: revised assessment criteria and scales (Concluding Phase 2). *Research Notes,* 15:9-11.

SHAW, S. 2004b. IELTS writing: revising assessment criteria and scale (Phase 3). *Research Notes,* 16:3-6.

SHAW, S.D. 2006. IELTS Writing: revising assessment criteria and scales (Conclusion). *Research Notes,* 24:19-22.

SHAW, S.D. and FALVEY, P. 2008. The IELTS writing assessment revision project: towards a revised rating scale. *Research Report,* 1: January. Cambridge: University of Cambridge ESOL Examinations.

SHAW, S.D. and JORDAN, S. 2002. CELS Writing: test development and validation activity. *Research Notes,* 9:10-13.

SHAW, S.D. and WEIR, C. 2007. Examining writing: research and practice in assessing second language writing. Studies in Language Testing, 26. Cambridge: Cambridge University Press.

SHEPARD, L. 1993. Evaluating test validity. *Review of Research in Education,* 19(1):405-450.

SHOHAMY, E. 1984. Does the testing method make a difference? The case of reading comprehension. *Language Testing,* 1(2):147-170.

SHOHAMY, E. 2001. The Power of tests: a critical perspective on the uses and consequences of language tests. Harlow: Longman/Pearson.

SMITH, A.G. and LIEBENBERG, L. 2003. Understanding the dynamics of parent involvement in schooling within the poverty context. *South African Journal of Education,* 23(1):1-5.

SOUTH AFRICA. Department of Education. 2005. National Curriculum Statement (Grades 10-12). Pretoria: Department of Education.

SOUTH AFRICA. Department of Education. 2008. Language Programme Guidelines. Pretoria: Department of Education.

SOUTH AFRICA. Department of Education. 2008. Subject Assessment Guidelines. Pretoria: Department of Education.

SOUTH AFRICA. 2009. Department of Education Examination Guidelines. Pretoria: Department of Education.

SPOLSKY, B. 1995. Measured words. Oxford: Oxford University Press.

SPSS. 2002. SPSS for Windows Release 11.5.0. Chicago: SPSS.

STATISTICA. 2008. StatSoft, Inc., 1984-2008. STATISTICA (data analysis software system), version 8.0. [Available online:] www.statsoft.com.

STEMLER, S.E. 2004. A Comparison of consensus, consistency, and measuring approaches to estimating interrater reliability. *Practical Assessment, Research and Evaluation,* 9(4). [Available online:] http://PAREonline.net/getvn.asp?v=9&n=4 Date accessed: 25 October 2007.

STEVENSON, D.K. 1985. Authenticity, validity and a tea-party. *Language Testing,* 2(1):41-47.

SUBJECT ASSESSMENT GUIDELINES. 2008. See South Africa.

TAYLOR, N. and PRINSLOO, C. 2005. The quality learning project lessons for high school improvement in South Africa. Johannesburg: Joint Education Trust.

TAYLOR, L. 2000. Approaches to rating scale revision. *Research Notes*, 3:14-16.

TAYLOR, L. 2002. IELTS writing test revision steering group discussion paper. Internal report. Cambridge: UCLES.

TAYLOR, N and VINJEVOLD, P. 1999. Getting learning right. Johannesburg: Joint Education Trust.

TOULMIN, S.E. 2003. The use of argument. 2nd ed. Cambridge: Cambridge University Press.

TURNER, C.E. 2000. Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *Canadian Modern Language Review*, 56(4). [Available online:] http://www.utpjournals.com/product/cmlr/564/564-Turner.html Date accessed: 31 August 2006.

TURNER, C.E. 2002. Listening to the voices of rating scale developers: identifying salient features for second language performance assessment. *Canadian Modern Language Review*,

56 (4). Available online: http://www.utpjpurnals.com/product/cmlr/564/564-Turner.html. Date of access: 31 August, 2006.

UCLES. 1998. See Milanovic, M.

UNDERHILL, N. 1987. Testing spoken language. Cambridge: Cambridge University Press.

UPSHUR, J.A. and TURNER, C.E. 1995. Constructing rating scales for second language tests. *ELT Journal,* 49:3-12.

URQUHART, A.H. and WEIR, C.J. 1998. Reading in a second language: process, product and practice. Harlow: Longman.

VAN DER WALT, J.L. 2009. Personal communication. Potchefstroom (notes in possession of researcher).

VAN DER WALT, J.L and STEYN, H.S. (jr). 2007. Pragmatic validation of a test of academic literacy. *Ensovoort,* 11(2), 138-153.

VAN DER WESTHUIZEN, P.C., MENTZ, P.J.J., MOSOGE, M.J., NIEWOUDT, H.D., STEYN, H.J., LEGOTLE, M.W., MAAGA, M.P. and SEBEGO, 1999. A quantitative analysis of the poor performance of Grade 12 students in 1997. *South African Journal of Education,* 19(4):315-319.

WALL, D., CLAPHAM, C. and ALDERSON, J.C. 1991. Validating tests in difficult circumstances. (*In* Hamp-Lyons, L. (ed.) Assessing second language writing in academic contexts. Norwood, NJ: Ablex).

WEIDEMAN, A. 2009. Constitutive and regulative conditions for the assessment of academic literacy. Forthcoming in *South African Linguistics and Applied Linguistics Studies (SALALS),* 27(3).

WEIGLE, S.C. 1994. Effects of training on raters of ESL compositions. *Language Testing,* 11:197-223.

WEIGLE, S.C. 1999. Investigating rater/prompt interactions in writing assessment: quantitative and qualitative Approaches. *Assessing Writing,* 6(2):145-178.

WEIGLE, S.C. 2002. Assessing writing. Cambridge: Cambridge University Press.

WEIR, C.J. 1983. Identifying the language problems of overseas students in tertiary education in the United Kingdom. University of London (thesis - Ph.D.).

WEIR, C.J. 1990. Communicative language testing. New York: Prentice Hall.

WEIR, C.J. 1993. Understanding and developing language tests. London: Prentice Hall.

WEIR, C.J. 2005. Language testing and validation: an evidence-based approach. Oxford: Palgrave Macmillan.

WEIR, C.J. and SHAW, S.D. 2005. Establishing the validity of Cambridge ESOL Writing tests: towards the implementation of a socio-cognitive model for test validation. *Research Notes,* 21:10-14.

WEST, M. 2008. The mystery of Zipf. *Plus Magazine – Living Mathematics,* 21 August. [Available online:] http://plus.maths.org/blog/2008/08/mystery-of-zipf.html Date accessed: 13 January 2009.

WHITE, E.M. 1985. Teaching and assessing writing. San Francisco, CA: Jossey-Bass.

WIGGLESWORTH, G. 1993. Exploring bias analysis as a tool for improving rater consistency in assessing oral Interaction. *Language Testing,* 10(3):305-335.

WOLF, A. 1995. Competence based assessment. Milton Keynes: Open University Press.

WOLFE, E., KAO, C., and RANNEY, M. 1998. Cognitive differences in proficient and non-proficient essay scorers. *Written Communication,* 15(4):469-482.

ZIPF, G.K. 1949. Human behaviour and the principle of least-effort. Addison-Wesley.

ZIPF, G.K. 2006. Bibliography on Zipf's law. [Available online:] http://www.nslij-genetics.org/wli/zipf/2006.html Date accessed: 13 November 2008.

ZIPF'S LAW. 2008. See West, M.

APPENDICES

Appendix A: Current scale used for assessing Grade 12 English FAL at Grade 12 level for the FET examination

		Outstanding	Meritorious	Substantial	Adequate	Moderate	Elementary	Not achieved
ENGLISH FIRST ADDITIONAL LANGUAGE RUBRIC NSC SECTION A: ESSAY 50 MARKS	LANGUAGE	- Language, punctuation effectively used. Uses figurative language appropriately. - Choice of words highly appropriate. - Sentences, paragraphs coherently constructed. - Style, tone, register highly suited to topic. - Text virtually error-free following proof-reading, editing. - Length in accordance with requirements of topic.	- Language, punctuation correct, and able to include figurative language correctly. - Choice of words varied and Correctly used. - Sentences, paragraphs logical, varied. - Style, tone, register appropriately suited to topic. - Text largely error-free following proof-reading, editing. - Length correct.	- Language and punctuation mostly correct. - Choice of words suited to text. - Sentences, paragraphs well constructed. - Style, tone, register suited to topic in most of the essay. - Text by and large error-free following proof-reading, editing. - Length correct.	- Language simplistic, punctuation adequate. - Choice of words adequate. - Sentences, paragraphing might be faulty in places but essay still makes sense. - Style, tone, register generally consistent with topic requirements. - Text still contains errors following proof-reading, editing. -Length correct.	- Language ordinary and punctuation often inaccurately used. - Choice of words basic. - Sentences, paragraphs, faulty but ideas can be understood. - Style, tone, register lacking in coherence. - Text contains several errors following proof-reading, editing. - Length – too long / short.	- Language and punctuation flawed. - Choice of words limited. - Sentences, paragraphs constructed at an elementary level. - Style, tone, register inappropriate. - Text error-ridden despite proof-reading, editing. - Length – too long / short	- Language and punctuation seriously flawed. - Choice of words inappropriate. - Sentences, paragraphs muddled, inconsistent. - Style, tone, register flawed in all aspects. - Text error-ridden and confused following proof-reading, editing. - Length – far too long / short
CONTENT		Code 7: 80 – 100%	Code 6: 70 – 79%	Code 5: 60 – 69%	Code 4: 50 – 59%	Code 3: 40 – 49%	Code 2: 30 – 39%	Code 1: 00 – 29%
Outstanding - Content shows impressive insight into topic. - Ideas: thought-provoking, mature. - Coherent development of topic. Vivid detail. - Critical awareness of impact of language. - Evidence of planning and/or drafting has produced virtually flawless, presentable essay.	Code 7 80-100%	40 - 50	38 – 42	35 – 39				
Meritorious - Content shows thorough interpretation of topic. - Ideas: imaginative, interesting. - Logical development of details. Coherent. - Critical awareness of impact of language . - Evidence of planning and/or drafting has produced a well crafted, presentable essay.	Code 6 70-79%	38 – 42	35 – 39	33 – 37	30 – 34			

Substantial - Content shows a sound interpretation of topic. - Ideas: interesting, convincing. - Several relevant details developed. - Critical awareness of language evident. - Evidence of planning and/or drafting has produced a presentable and very good essay.	Code 5 60-69%	35 – 39	33 – 37	30 – 34	28 – 32	25 - 29		
Adequate - Content: an adequate interpretation of topic. - Ideas: ordinary, lacking depth. - Some points, necessary details developed. - Some awareness of impact of language. - Evidence of planning and/or drafting has produced a satisfactorily presented essay.	Code 4 50-59%		30 – 34	28 – 32	25 – 29	23 – 27	20 – 24	
Moderate - Content: ordinary. Gaps in coherence. - Ideas: mostly relevant. Repetitive. - Some necessary points evident. - Limited critical language awareness. - Evidence of planning and/or drafting that has produced a moderately presentable and coherent essay.	Code 3 40-49%			25 – 29	23 – 27	20 – 24	18 – 22	15 – 19
Elementary - Content not always clear, lacks coherence. - Ideas: few ideas, often repetitive, - Sometimes off topic. General line of thought difficult to follow. - Inadequate evidence of planning/drafting. Essay not well presented.	Code 2 30-39%				20 – 24	18 – 22	15 – 19	03 – 17
Not Achieved - Content irrelevant. No coherence. - Ideas: repetitive, off topic. - Non-existent planning/drafting. Poorly presented essay.	Code 1 00-29%					15 – 19	03 – 17	00 – 14

Appendix B1: Phase 2 First Draft Scale

A. CONTENT								
1. Too short.	1	2	3	4	5	6	7	Appropriate length.
2. Main idea unclear.	1	2	3	4	5	6	7	Main meaning is clear.
3. Lacking originality and//or interest.	1	2	3	4	5	6	7	Demonstrates originality and//or interest.
B. GRAMMAR								
4. Inadequate introduction, body and conclusion.	1	2	3	4	5	6	7	Adequate introduction, body and conclusion.
5. Inadequate flow of ideas through essay.	1	2	3	4	5	6	7	Inadequate flow of ideas through essay.
6. Inadequate paragraphing.	1	2	3	4	5	6	7	Adequate paragraphing.
C GRAMMAR								
7. Incorrect syntax.	1	2	3	4	5	6	7	Correct syntax.
8. Incorrect use of tense and/or concord.	1	2	3	4	5	6	7	Correct use of tense and/or concord.
D. VOCABULARY								
9. Inadequate range.	1	2	3	4	5	6	7	Adequate range.
10. Wording inappropriate to text type.	1	2	3	4	5	6	7	Wording appropriate to text type.
11. Inappropriate use of idiomatic expression.	1	2	3	4	5	6	7	Appropriate use of idiomatic expression.
12. Inappropriate use of linking words and phrases.								Adequate use of linking words and phrases.
E. EDITING								
13. Incorrect spelling and/or capitalisation.	1	2	3	4	5	6	7	Correct spelling and/or capitalisation.
14. Incorrect punctuation.	1	2	3	4	5	6	7	Correct punctuation.
15. Untidy presentation.	0			2				Tidy presentation.
TOTAL	100/2							

Appendix B2: Phase 2 Second Draft Scale

A. CONTENT								
1. Demonstrating a lack of insight into and understanding of the topic.	1	2	3	4	5	6	7	Demonstrating insight into and understanding of the topic.
2. Lacking originality and//or interest.	1	2	3	4	5	6	7	Demonstrates originality and//or interest.
B. STRUCTURE AND DEVELOPMENT								
3. Inadequate introduction, body and conclusion.	1	2	3	4	5	6	7	Adequate introduction, body and conclusion.
4. Inadequate flow of ideas through essay.	1	2	3	4	5	6	7	Inadequate flow of ideas through essay.
5. Inadequate paragraphing.	1	2	3	4	5	6	7	Adequate paragraphing.
6. Structure inappropriate to type.	1	2	3	4	5	6	7	Structure appropriate to type.
C. GRAMMAR								
7. Incorrect syntax.	1	2	3	4	5	6	7	Correct syntax.
8. Incorrect use of tense and/or concord.	1	2	3	4	5	6	7	Correct use of tense and/or concord.
9. Inadequate range of sentence types.	1	2	3	4	5	6	7	Adequate range of sentence types.
D. VOCABULARY								
10. Inadequate range.	1	2	3	4	5	6	7	Adequate range.
11. Language inappropriate to text type.	1	2	3	4	5	6	7	Language appropriate to text type.
12. Inappropriate use of linking words and phrases.	1	2	3	4	5	6	7	Appropriate use of linking words and phrases.
E. EDITING								
13. Incorrect spelling and/or capitalisation.	1	2	3	4	5	6	7	Correct spelling and/or capitalisation.
14. Incorrect punctuation.	1	2	3	4	5	6	7	Correct punctuation.
15. Untidy presentation.	0			2				Tidy presentation.
TOTAL	100/2							

Appendix B3: Phase 2: Third and Final Draft Scale

A. CONTENT								
1. Demonstrating a lack of insight into and understanding of the topic.	1	2	3	4	5	6	7	Demonstrating insight into and understanding of the topic.
2. Lacking originality and//or interest.	1	2	3	4	5	6	7	Demonstrates originality and//or interest.
B. STRUCTURE AND DEVELOPMENT								
3. Inadequate introduction, body and conclusion.	1	2	3	4	5	6	7	Adequate introduction, body and conclusion.
4. Inadequate flow of ideas through essay.	1	2	3	4	5	6	7	Inadequate flow of ideas through essay.
5. Inadequate paragraphing.	1	2	3	4	5	6	7	Adequate paragraphing.
6. Structure inappropriate to type.	1	2	3	4	5	6	7	Structure appropriate to type.
C. GRAMMAR								
7. Incorrect syntax.	1	2	3	4	5	6	7	Correct syntax.
8. Incorrect use of tense and/or concord.	1	2	3	4	5	6	7	Correct use of tense and/or concord.
9. Inadequate range of sentence types.	1	2	3	4	5	6	7	Adequate range of sentence types.
D. VOCABULARY								
10. Inadequate range.	1	2	3	4	5	6	7	Adequate range.
11. Language inappropriate to text type.	1	2	3	4	5	6	7	Language appropriate to text type.
12. Inappropriate use of linking words and phrases.	1	2	3	4	5	6	7	Appropriate use of linking words and phrases.
E. EDITING								
13. Incorrect spelling and/or capitalisation.	1	2	3	4	5	6	7	Correct spelling and/or capitalisation.
14. Incorrect punctuation.	1	2	3	4	5	6	7	Correct punctuation.
15. Untidy presentation.	0			2				Tidy presentation.
TOTAL	100/2							

Appendix B4: Explanatory scale guide based on the third draft

GRID EXPLANATION & INSTRUCTION SHEET

Instructions: How to choose an appropriate level for each of the 15 features.

1) Decide whether the performance is above or below average (level 4) for that feature.
2) Decide to what extent the performance is above or below average for that feature.

	Criteria	Instructions and Explanation
	A. CONTENT	Read through the essay and award marks for 1 and 2
1	Insight into and understanding of the topic	• Assess whether candidate has addressed, developed and sustained the topic
2	Originality and/or interest	• Assess the extent to which the essay engages the reader. • Give credit for any response that provides fresh/creative perspective on the topic
	B. STRUCTURE AND DEVELOPMENT	This category refers to the way information is organised in the essay in accordance to the essay type (argumentative, narrative, descriptive, comparison and contrast, cause and effect).
3	Essay structure: introduction, body and conclusion	• These three sections have to be present, appropriate to the text type and functional (introducing, developing and rounding off the topic)
4	Appropriateness of structure to text type	• •he main sections of the essay must follow the conventions of the essay type (argumentative, narrative, descriptive, comparison and contrast, cause and effect).
5	Flow of ideas through the essay	• The essay must show natural/ logical progression of ideas/ events/ facts from the introduction to the conclusion.
6	Paragraphing	• The essay must be divided into paragraphs • Each paragraph must have a main idea (usually a topic sentence) • The main idea should be developed further by the supporting sentences in the paragraph
	C. GRAMMAR	• This section deals with the accurate use of grammatical structures
7	Syntax	• As a rule, sentences must be complete (subject & main verb), and contain correct word order. Exceptions used for creative effect should not be penalised if appropriate.
8	Use of tense and concord	• Tense and concord must used be correctly and appropriately
9	Range of sentence types	• The essay must demonstrate a variety of sentence types and sentences of different lengths and structures accurately and effectively
	D. VOCABULARY	This section assesses the extent, accuracy and appropriateness of a candidate's vocabulary
10	Range of vocabulary	• Candidates have to demonstrate the they have a sufficient extent of vocabulary to express their ideas • Credit must be given for sophistication in words and expressions
11	Appropriateness of vocabulary	• Words must be used correctly and appropriately
		• Assess the candidate's ability to use style appropriately, such as formal and informal, narrative, descriptive and argumentative

12	Use of linking words and phrases	• The candidate demonstrates the ability to use conjunctions, pronouns, adverbs and other devices to link parts of sentences, sentences and paragraphs.
	E. EDITING	**This section assesses the product in terms of technical aspects such as spelling, punctuation and legibility.**
13	Spelling and/or capitalisation	• Spelling must be accurate (this includes the use of the apostrophe) • Capital letters must be used appropriately • N the entire essay is written in capital letters, award a maximum of 3 for category E13.
14	Punctuation	• Punctuation (e.g. full stops, commas, colons, dashes and inverted commas) must be used appropriately and correctly.
15	Presentation	• A legible and neatly presented essay is credited with 2 marks

Appendix C1: Phase 3 Revised Draft Scale

FINAL PROPOSED DRAFT RATING SCALE								
		Poor		Adequate		Very good		
A. CONTENT								
1. No insight into and understanding of topic.	1	2	3	4	5	6	7	Outstanding insight into and comprehensive understanding of topic.
2. Hardly any originality and/or little interest/ mundane.	1	2	3	4	5	6	7	Highly original/ Fresh perspective/ original/ engaging creativity.
3. Irrelevant and immature ideas	1	2	3	4	5	6	7	Mature and thought provoking ideas.
4. Does not follow the conventions of essay type.	1	2	3	4	5	6	7	Ideally follows conventions of essay type.
5. Incoherent flow of ideas.	1	2	3	4	5	6	7	Highly coherent flow of ideas.
B. STRUCTURE AND STYLE								
6. No division into introduction, body, conclusion.	1	2	3	4	5	6	7	Effective division into introduction, body and conclusion.
7. No paragraphing.	1	2	3	4	5	6	7	Effective paragraphing.
C. GRAMMAR								
8. Incorrect syntax.	1	2	3	4	5	6	7	Correct syntax.
9. Incorrect tense & concord.	1	2	3	4	5	6	7	Correct tense & concord.
10. No variety in range of sentence types.	1	2	3	4	5	6	7	Wide variety in range of sentence type.
11. Multiple errors in spelling & punctuation.	1	2	3	4	5	6	7	Error-free spelling & punctuation.
D. VOCABULARY								
12. Limited range.	1	2	3	4	5	6	7	Extended range.
13. Inappropriate style, diction & register.	1	2	3	4	5	6	7	Highly appropriate style, diction & register.
14. Ineffective use of linking devices (words & phrases).	1	2	3	4	5	6	7	Sophisticated use of linking devices (words & phrases).
E. LENGTH								
Deviation from requirement.	0					2		Adheres to requirement.
TOTAL	100/2							

Appendix C2: Revised Scale Guide

SCORING GUIDE		
	Criteria & Features	INSTRUCTIONS AND EXPLANANTIONS
	A. CONTENT	**Read through the essay and award marks for 1 and 2**
1	Insight into and understanding of the topic	• Assess whether candidate has addressed, developed and sustained the topic
2	Originality and/or interest	• Assess the extent to which the essay engages the reader. • Give credit for any response that provides fresh/creative perspective on the topic
3	Relevance and maturity of ideas	• The essay must be clearly relevant to the topic. • Ideas should be thought through and contribute to the main topic.
4	Appropriateness of structure to text type	• The main sections of the essay must follow the conventions of the essay type (argumentative, narrative, descriptive, comparison and contrast, cause and effect).
5	Flow of ideas through the essay	The essay must show natural/ logical progression of ideas/ events/ facts from the introduction to the conclusion and between paragraphs.
	B. STRUCTURE AND DEVELOPMENT	**This category refers to the way information is organised in the essay in accordance to the essay type (argumentative, narrative, descriptive, comparison and contrast, cause and effect).**
6	Introduction, body and conclusion	The essay must contain a clear introduction, body and conclusion.
7	Paragraphing	• The essay must be divided into paragraphs • Each paragraph must have a main idea (usually a topic sentence) • The main idea should be developed further by the supporting sentences in the paragraph
	C. GRAMMAR	**This section deals with the accurate use of grammatical structures.**
8	Syntax	• As a rule, sentences must be complete (subject & main verb), and contain correct word order. Exceptions used for creative effect should not be penalised if appropriate.
9	Use of tense and concord	• Tense and concord must be used correctly and appropriately.
10	Range of sentence types	• The essay must demonstrate a variety of sentence types and sentences of different lengths and structures accurately and effectively.
11	Spelling and/or Capitalisation and Punctuation	• Spelling must be accurate (this includes the use of the apostrophe) • Capital letters must be used appropriately • If the entire essay is written in capital letters, award a maximum of 3 for category C11. • Punctuation (e.g. full stops, commas, colons, dashes and inverted commas) must be used appropriately and correctly.
	D. VOCABULARY	**This section assesses the extent, accuracy and appropriateness of a candidate's vocabulary**
12	Range of vocabulary	• Candidates have to demonstrate the they have a sufficient extent of vocabulary to express their ideas • Credit must be given for sophistication in words and expressions
13	Appropriateness of vocabulary	• Words must be used correctly and appropriately • Assess the candidate's ability to use style appropriately, such as formal and informal, narrative, descriptive and argumentative
14	Use of linking words and phrases	• The candidate demonstrates the ability to use conjunctions, pronouns, adverbs and other devices to link parts of sentences, sentences and paragraphs
	E. LENGTH	**The candidate must adhere to the length limitation as specified on the examination question paper**

Appendix D: Phase 4 Trial Examiner Questionnaire

Phase 4: Trial Examiner Questionnaire

Please note: your responses are confidential. You do not need to reveal your name

Instructions: For Section A, please base your responses on your overall experience as an examiner in the light of the revised scale.

A. EXAMINER BACKGROUND INFORMATION

1. Where are the centres you examine?
2. How many years experience do you have as an examiner of L2 writing?
3. How many years experience do you have as an English L2 teacher?

B. USING THE PROPOSED SCALE – GENERAL RATING ISSUES

00 What is your general reaction to the new approach in relation to rating a new text?

_____

_____

01 Are you confident about rating accurately?	Yes No
02 Do you find the revised scale more difficult to use than the current scale?	Yes No Comment:
03 How well do you understand the criteria? (are they clear?)	Comment:
04 Do you find the criteria to be clearer in comparison to the current scale?	Yes No Comment:
05 Do you think the features in the proposed scale are clear?	Yes No
06 Do you think the features under each criterion address relevant aspects?	Yes No Comment:
07 Do you find it easy to mark the five criteria?	Yes No Comment:
08 How adequate is the new scale wording?	Adequate Inadequate Comment:
09 Does the revised scale capture the essential qualities of the written performance?	Yes No, because..... Comment:
10 Can you distinguish all the band levels clearly and interpret them consistently?	Yes No, because..... Comment:
11 Can you use all seven levels distinguished on the scale? (Do you think that seven levels are enough or too many?)	Yes No Comment:
12 Do you think the scale increases the reliability of subjectively judged ratings?	Yes No Comment:
13 Can you interpret effectively any	Yes No....such as

"relative" language terms, e.g. 'adequate', 'limited'. If not, are there any words or phrases you cannot interpret?	Comment:
14 Do you always confine yourself exclusively to the context of the scale? If not, what else influences your judgements?	Yes No..... Comment:
15 Do you concentrate on one criterion at a time and read the response specifically for that criterion?	Always Usually Sometimes Never
16 Do you read the whole essay before you start scoring individual criteria?	Always Usually Sometimes Never
17 Do you consider all the criteria simultaneously while scoring?	Always Usually Sometimes Never
18 Does the application of the revised scale affect the time it takes to rate a full script?	Yes No Comment:
19 Do you consider the current and proposed scales to be of the same standard?	Yes No Comment:
20 Do you believe the candidates would get the same score on the current and the proposed scale?	Yes No, because..... Comment:
Do you think that the proposed scale helps you to evaluate performances more critically?	Yes No
21 Does the proposed scale make you aware of possible personal bias that you may have when scoring?	Yes No Comment:

22 What do you feel is the main advantage of / what do you think is the best asset of the proposed scale?

_____

_____

23 What do you think is the main disadvantage of / what is your main concern regarding the proposed scale?

_____

_____

24 If you could change anything in the proposed scale, would you? If so, what would you change?

_____

_____

Appendix E: Final Outcome: the Proposed Scale

FINAL PROPOSED DRAFT RATING SCALE									
		Poor		Adequate	Very good				
A. CONTENT									
1. No insight into and understanding of topic.	0	1	2	3	4	5	6	7	Outstanding insight into and comprehensive understanding of topic.
2. Hardly any originality and/or little interest/ mundane.	0	1	2	3	4	5	6	7	Highly original/ Fresh perspective/ original/ engaging creativity.
3. Irrelevant and immature ideas	0	1	2	3	4	5	6	7	Mature and thought provoking ideas.
4. Does not follow the conventions of essay type.	0	1	2	3	4	5	6	7	Ideally follows conventions of essay type.
5. Incoherent flow of ideas.	0	1	2	3	4	5	6	7	Highly coherent flow of ideas.
B. STRUCTURE AND STYLE									
6. No division into introduction, body, conclusion.	0	1	2	3	4	5	6	7	Effective division into introduction, body and conclusion.
7. No paragraphing.	0	1	2	3	4	5	6	7	Effective paragraphing.
C. GRAMMAR									
8. Incorrect syntax.	0	1	2	3	4	5	6	7	Correct syntax.
9. Incorrect tense & concord.	0	1	2	3	4	5	6	7	Correct tense & concord.
10. No variety in range of sentence types.	0	1	2	3	4	5	6	7	Wide variety in range of sentence type.
11. Multiple errors in spelling & punctuation.	0	1	2	3	4	5	6	7	Error-free spelling & punctuation.
D. VOCABULARY									
12. Limited range.	0	1	2	3	4	5	6	7	Extended range.
13. Inappropriate style, diction & register.	0	1	2	3	4	5	6	7	Highly appropriate style, diction & register.
14. Ineffective use of linking devices (words & phrases).	0	1	2	3	4	5	6	7	Sophisticated use of linking devices (words & phrases).
E. LENGTH									
Deviation from requirement.	0			1			2		Adheres to requirement.
TOTAL	100/2								