

**The incremental validity of three tests of academic literacy in the context of a South African university of technology**

**Kabelo Wilson Sebolai**

Submitted in fulfilment of the requirements for the degree  
D. Phil. Applied Linguistics  
In the Faculty of Humanities,  
Department of Linguistics and Language Practice at the University of  
the Free State

Supervisor: Prof. A. J. Weideman  
Co-supervisor: Prof. T. Van Dyk

July 2016

# Declaration

(i) “I, Sebolai Kabelo Wilson, declare that the Doctoral Degree research thesis or interrelated, publishable manuscripts / published articles, or coursework Doctoral Degree mini-thesis that I herewith submit for the Doctoral Degree qualification D. Phil. (Language Practice) at the University of the Free State is my independent work, and that I have not previously submitted it for a qualification at another institution of higher education.”

(ii) “I, Sebolai Kabelo Wilson, hereby declare that I am aware that the copyright is vested in the University of the Free State.”

(iii) “I, Sebolai Kabelo Wilson, hereby declare that all royalties as regards intellectual property that was developed during the course of and/or in connection with the study at the University of the Free State will accrue to the University.”

In the event of a written agreement between the University and the student, the written agreement must be submitted in lieu of the declaration by the student.

**Signature**

---

**Date**

---

# Acknowledgments

I would like to express my gratitude to the following:

- My supervisors Professors Albert Weideman and Tobie Van Dyk – for their prompt feedback on all my submissions and the expert guidance they provided from the beginning to the end of this thesis. I am particularly grateful to them for believing in me and unlocking the potential I needed to complete the thesis.
- My colleague Janine Dunlop – for her tireless technical assistance with the layout of the thesis.
- Anna Weideman, Professor Weideman’s wife – for her assistance with the layout and complete presentation of the bibliography.
- My colleague and friend Natalie Le Roux – for her unwavering moral support.
- Professor Robert Schall – for statistically analysing the data used in the study.
- My former colleague Anneli Hardy – for the extra statistical support she provided.
- My former colleague, Magauta Kenke, for assisting with the collection of the data for this study.
- My partner Nketsi Matasane – for her love and support from the beginning to the end of this thesis.
- The Vice Chancellor of the Central University of Technology Professor Thandwa Mthembu – for believing in me and relentlessly encouraging me to obtain a PhD.
- God Almighty – for giving me the opportunity and strength to complete this study.

## Table of Contents

Acknowledgments	ii
List of Tables	vii
List of Figures	x
List of acronyms	xi
Appendices	xiii
Abstract	xiv
Opsomming	xvi
Chapter 1	1
1.1 Background to the problem	3
1.2 A definition of academic literacy in the South African higher education context	7
1.3 Constructs of the four assessments	11
1.3.1 The construct of the National Benchmark Test in Academic Literacy	11
1.3.2 The construct of the Test of Academic Literacy Levels	15
1.3.3 The construct of the English HL and FAL examinations	18
1.3.4 The construct of the Placement Test English Second Language Advanced Level	20
1.4 Problem statement	21
1.5 Aim of the study	25
1.6 Hypothesis of the study	25
1.7 Chapter outline	25
1.8 Conclusion	28
Chapter 2	29
2.1 Introduction	29
2.2 Validity	29
2.2.1 Content validity	35
2.2.2 Construct validity	36
2.2.3 Criterion-related validity	36
2.3. Hypothesis of the study	39
2.3.1 TALL is a reliable test of academic literacy	40
2.3.2 TALL has construct validity	44
2.3.3 TALL possesses acceptable item difficulty and discrimination levels	45
2.3.4 Factor analysis has attested to the construct validity of TALL	48
2.3.5 Decision Theory has been used to identify misclassification in TALL	54
2.3.6 TALL items function similarly for all test takers	57
2.3.7 There is an acceptable degree of internal correlations between the TALL test tasks	60

2.3.8	Studies have been conducted to obtain feedback from TALL test-takers	61
2.4	Conclusion	63
Chapter 3		65
3.1	Introduction	65
3.2	Data collection	65
3.3	Sampling	66
3.4	Ethical considerations	69
3.5	Procedure for defining academic success	73
3.6	Procedure for data analysis	74
3.7	Test Specifications	79
3.7.1	The National Benchmark Test of Academic Literacy	79
3.7.2	Proficiency Test English Second Language Advanced Level	82
3.7.3	Test of Academic Literacy Levels	83
3.7.4	The Grade 12 English Home and First Language examinations	85
3.8	Conclusion	88
Chapter 4: Data analysis		90
4.1	Introduction	90
4.2	Results of the study	90
4.3	NBT AL, PTESLAL, Grade 12 English and 2012 average	93
4.3.1	Descriptive statistics	93
4.3.2	Intercorrelations: NBT AL, PTESLAL, Grade 12 English and 2012 average	94
4.3.3	Linear regression analyses: NBT AL, PTESLAL, Grade 12 English and 2012 average	96
4.3.4	Multiple regression analysis: NBT AL, PTESLAL, Grade 12 English and 2012 Average	106
4.4	TALL, PTESLAL, Grade 12 English and 2014 average	113
4.4.1	Descriptive statistics	113
4.4.2	Intercorrelations: TALL, PTESLAL, Grade 12 English and 2014 average	115
4.4.3	Linear regression analyses: TALL, PTESLAL, Grade 12 English and 2014 average	117
4.4.4	Multiple regression analyses: TALL, PTESLAL, Grade 12 English and 2014 average	127
4.5	TALL, Grade 12 English and 2014 average	134
4.5.1	Descriptive statistics	134
4.5.2	Intercorrelations: TALL, Grade 12 English and 2014 average	135
4.5.3	Linear regression analysis: TALL, Grade 12 English and 2014 average	136
4.6	Conclusion	144
Chapter 5		145

5.1	Introduction	145
5.2	The results in relation to Grade 12 English	146
5.2.1	Construct definition with regard to Grade 12 language assessments	148
5.2.2	The issue of the validity of the Grade 12 English examination	150
5.2.3	The issue of the reliability of Grade 12 English assessment	151
5.3	The results in relation to PTESLAL	153
5.4	The results in relation to NBT AL	154
5.5	The results in relation to TALL	159
5.6	Recent studies on the predictive validity of academic literacy tests and Grade 12 results	162
5.7	Conclusion	166
Chapter 6		168
6.1	Introduction	168
6.2	Overview of the current investigation	168
6.3	Recommendations	170
6.4	Limitations of the study	184
6.5	Suggestions for further research	187
6.6	The low graduation output by South African universities	189
6.7	Conclusion	193
Chapter 7		194
7.1	Introduction	194
7.2	Theories of test validity	194
7.2.1	The traditional view of test validity	194
7.2.2	Messick's view of test validity	196
7.3	Implications of the results of the study for theories of test validity	200
7.3.1	Weideman's framework for applied linguistic designs	209
7.4	Reinterpretations of Messick's view of validity	212
7.4.1	McNamara and Roever's reinterpretation of Messick's view of test validity	213
7.4.2	Weideman's reinterpretation of Messick's view of test validity	214
7.5	Implications of the literature for the validity of tests of academic literacy	221
7.6	Implications of the results of the study for course validity	229
7.7	Conclusion	233

## List of Tables

Table 1 Bachman and Palmer's areas of language knowledge	12
Table 2 Bachman and Palmer's areas of strategic competence	13
Table 3 Selected properties of the academic literacy test (2005-2008) (standard deviations in italics)	42
Table 4 Two perspectives on language ability	53
Table 5 Potential misclassifications on the English version of the academic literacy test	56
Table 6 T-values of differences between mean scores on TALL of first year students who have an African language, English, or Afrikaans as their first language	59
Table 7 The subdomains of NBT AL	79
Table 8 Levels of cognitive challenge for the NBT AL	80
Table 9 Test specifications for PTESLAL	82
Table 10 The subdomains and task types proposed for TALL	84
Table 11 Task types, number of items and mark allocation for items in TALL	85
Table 12 Specifications for Grade 12 English HL and FAL final assessments	87
Table 13 Levels of cognitive challenge for Grade 12 English HL and FAL assessments	88
Table 14 The means and standard deviations for the scores on NBT AL, PTESLAL, Grade 12 English and 2012 average	93
Table 15 Intercorrelations for NBT AL, PTESLAL, Grade 12 English and 2012 average	95
Table 16 The F statistic from the linear regression of 2012 average on NBT AL (n=309)	97
Table 17 The R-Square for the model with NBT AL as the specified predictor of 2012 average performance (n=309)	98
Table 18 The t statistic for NBT AL as a predictor of 2012 average (n=309)	99
Table 19 The F statistic for the model involving PTESLAL as the specified predictor of 2012 average (n=303)	101
Table 20 The R-Square for the model including PTESLAL as the specified predictor of 2012 average (n=303)	101
Table 21 The t statistic for PTESLAL as a predictor of 2012 average (n=303)	102
Table 22 The F statistic for the model including Grade 12 English as the specified predictor of 2012 average (n=223)	104
Table 23 The R-Square for the model involving Grade 12 English as the specified predictor of 2012 average (n=223)	104
Table 24 The t statistic for Grade 12 English as the predictor of 2012 average performance (n=223)	105
Table 25 The F statistic for the model with NBT AL and Grade 12 English as specified predictors of 2012 average (n=223)	107
Table 26 The R-Square for the model with NBT AL and Grade 12 English as specified predictors of 2012 average (n=223)	107
Table 27 The t statistics for NBT AL and Grade 12 English as predictors of 2012 Average (n=223)	108
Table 28 The F statistic for the model with PTESLAL and Grade 12 English as specified predictors of 2012 average (n=219)	109
Table 29 The R-Square for the model with PTESLAL and Grade 12 English as specified predictors of 2012 average (n=219)	109
Table 30 The t statistics for PTESLAL and Grade 12 English as predictors of 2012 average (n=219)	110
Table 31 The F statistic for the model with NBT AL, PTESLAL and Grade 12 English as specified predictors of 2012 average (n=219)	111

Table 32 The R-Square for the model with NBT AL, PTESLAL and Grade 12 English as specified predictors of 2012 average (219)	111
Table 33 The t statistics for NBT AL, PTESLAL and Grade 12 English as predictors of 2012 average	112
Table 34 The means and standard deviations for the scores on TALL, PTESLAL, Grade 12 English and 2014 average	114
Table 35 Intercorrelations for TALL, PTESLAL, Grade 12 English and 2014 average	115
Table 36 The F statistic for the model with TALL as the specified predictor of 2014 average (n=98)	118
Table 37 The R-Square for the model with TALL as the specified predictor of 2014 average (n=98)	118
Table 38 The t statistic for TALL as the predictor of 2014 average performance (n=98)	119
Table 39 The F statistic for the model with PTESLAL as the specified predictor of 2014 average (n=98)	120
Table 40 The R-Square for the model with PTESLAL as the specified predictor of 2014 average (n=98)	121
Table 41 The t statistic for PTESLAL as a predictor of 2014 average (n=98)	121
Table 42 The F statistic for the model with Grade 12 English as the specified predictor of 2014 average (n=78)	124
Table 43 The R-Square for the model with Grade 12 English as the specified predictor of 2014 average (n=78)	124
Table 44 The t statistic for Grade 12 English as predictor of 2014 average (n=78)	125
Table 45 The F statistic for the model with TALL and Grade 12 English as specified predictors of 2014 average (n=78)	128
Table 46 The R-Square for the model with TALL and Grade 12 English as specified predictors of 2014 average (n=78)	128
Table 47 The t statistics for TALL and Grade 12 English as predictors of 2014 average performance (n=78)	129
Table 48 The F statistic for the model with PTESLAL and Grade 12 English as specified predictors of 2014 average (n=78)	129
Table 49 The R-Square for the model with PTESLAL and Grade 12 English as specified predictors of 2014 average (n=78)	130
Table 50 The t statistics for PTESLAL and Grade 12 English as predictors of 2014 average performance (n=78)	130
Table 51 The F statistic for the model with TALL, PTESLAL and Grade 12 English as specified predictors of 2014 average performance (n=78)	131
Table 52 The R-Square for the model with TALL, PTESLAL and Grade 12 English as specified predictors of 2014 average (n=78)	131
Table 53 The t statistics for TALL, PTESLAL and Grade 12 English as predictors of 2014 average (n=78)	132
Table 54 The means and standard deviations of the scores on TALL, Grade 12 English and 2014 average	134
Table 55 Intercorrelations for TALL, Grade 12 English and 2014 average	135
Table 56 The F statistic for the model with TALL as the specified predictor of 2014 average (n=604)	137
Table 57 The R-Square for the model with TALL as the specified predictor of 2012 average (n=604)	137
Table 58 The t Statistic for TALL as the predictor of 2014 average performance (n=604)	138
Table 59 The F statistic for the model with Grade 12 English as the specified predictor of 2014 average performance (n=478)	139



Table 60 The R-Square for the model with Grade 12 English as the specified predictor of 2014 average performance (n=478)	140
Table 61 The t statistic for Grade 12 English as a predictor of 2014 average performance (n=478)	140
Table 62 The F statistic for the model with TALL and Grade 12 English as specified predictors of 2014 average (n=478)	142
Table 63 The R-Square for the model with TALL and Grade 12 English as specified predictors of 2014 average performance (n=478)	143
Table 64 The t statistic for TALL and Grade 12 English as predictors of 2014 average (n=478)	143
Table 65 The Benchmarks for the National Benchmark Tests	226
Table 66 Levels of risk associated with scores on TALL	228

## List of Figures

Figure 1: Measures of homogeneity and heterogeneity in TALL 2008	50
Figure 2: The Bachman and Palmer construct of language ability	52
Figure 3: The fit plot for a linear regression analysis for NBT AL as predictor of 2012 average (n=309)	100
Figure 4: The fit plot for the results of a linear regression of 2012 average on PTESLAL (n=303)	102
Figure 5: The fit plot for the results of a linear regression for Grade 12 English as the predictor of 2012 average (n=223)	105
Figure 6: The fit plot for the results of a linear regression of 2014 average on TALL (n=98)	119
Figure 7: The fit plot for PTESLAL as the predictor of 2014 average performance (n=98)	122
Figure 8: The fit plot for Grade 12 English as predictor of 2014 average performance (n=78).	126
Figure 9: The fit plot for TALL as predictor of 2014 average (n=604)	138
Figure 10: The fit plot for Grade 12 English as the predictor of 2014 average performance (n=478)	141
Figure 11: The constitutive and regulative conditions of applied linguistic designs	210
Figure 12: Messick's "Facets of validity"	212
Figure 13: McNamara and Roever's reinterpretation of Messick's matrix of validity	214
Figure 14: The relationship of a selection of fundamental considerations in language testing	215

## List of acronyms

AARP – Alternative Admissions Research Project  
ANOVA – Analysis of Variance  
APS – Admission Point Score  
BICS – Basic Interpersonal Communicative Skills  
CALP – Cognitive Academic Language Proficiency  
CAPS – Curriculum and Assessment Policy Statement  
CFL – College of Foreign Languages  
CHE – Council on Higher Education  
CLT – Communicative Language Teaching  
CTT – Classical Test Theory  
CUT – Central University of Technology  
DBE – Department of Basic Education  
DIF – Differential Item Functioning  
ECP – Extended Curriculum Programme  
ex-DET – ex-Department of Education and Training  
ex-HOA – ex-House of Assembly  
FAL – First Additional Language  
GLB – Greatest Lower Bound  
HE – Higher Education  
HESA – Higher Education South Africa  
HL – Home Language  
HSRC – Human Sciences Research Council  
ICELDA – Inter-institutional Centre for Language Development and Assessment  
IRF – Item Response Function  
IRT – Item Response Theory  
KSAs – Knowledge, Skills and Abilities  
MACH – Mathematics Achievement  
MCOM – Mathematics Comprehension  
NBT AL – National Benchmark Test in Academic Literacy  
NBTP – National Benchmark Tests Project  
NBT – National Benchmark Test  
NBTs – National Benchmark Tests

NCME – National Council on Measurement in Education  
NSC – National Senior Certificate  
NSFAS – National Student Financial Aid Scheme  
NWU – North West University  
PTEEP – Placement Test in English for Educational Purposes  
PTESLAL – Proficiency Test English Second Language Advanced Level  
SCU – Statistical Consultation Unit  
SRT – Scientific Reasoning Test  
TAG – Toets van Akademiese Geletterheidsvlakke  
TALL – Test of Academic Literacy Levels  
TALPS – Test of Academic Literacy for Postgraduate Students  
TLU – Target Language Use  
UCT – University of Cape Town  
UP – University of Pretoria  
US – University of Stellenbosch  
Wits - Witwatersrand

## Appendices

- A. Correlations: 2012 Average, NBT AL, PTESLAL and Gr. 12 English
- B. Simple regression: 2012 Average and NBT AL
- C. Simple regression: 2012 Average and PTESLAL
- D. Simple regression: 2012 Average and Gr. 12 English
- E. Multiple regression: 2012 Average, NBT AL and Gr. 12 English
- F. Multiple regression: 2012 Average, PTESLAL and Gr. 12 English
- G. Multiple regression: 2012 Average, NBT AL, PTESLAL and Gr. 12 English
- H. Correlations: 2014 Average, TALL, PTESLAL and Gr. 12 English
- I. Simple regression: 2014 Average and TALL
- J. Simple regression: 2014 Average and PTESLAL
- K. Simple regression: 2014 Average and Gr. 12 English
- L. Multiple regression: 2014 Average, TALL and Gr. 12 English
- M. Multiple regression: 2014 Average, PTESLAL and Gr. 12 English
- N. Multiple regression: 2014 Average, TALL, PTESLAL and Gr. 12 English
- O. Correlations: 2014 Average, TALL and Grade 12 English
- P. Simple regression: 2014 Average and TALL
- Q. Simple regression: 2014 Average and Gr. 12 English
- R. Multiple regression: 2014 Average, TALL and Gr. 12 English

## **Abstract**

This study focuses on the incremental validity of three assessments of academic language readiness, compared to Grade 12 English results: the National Benchmark Test in Academic Literacy (NBT AL), the Proficiency Test English Second Language Advanced Level (PTESLAL) and the Test of Academic Literacy Levels (TALL) at the end of the first year of academic study. More specifically, the study investigates the ability of any of the four assessments to predict first year academic performance better than the others. Where those that are examined do not possess this ability, the further question is asked: can they at least add to the predictive power of the best predictor? Ultimately, the aim is to determine if the assessments designed to provide additional information about first year academic preparedness are valid for this purpose, and the extent to which this is the case.

The study starts with a brief exploration of the literature on the reportedly low levels of academic language ability among first time entrants to higher education in South Africa in recent years, and the consequent need for the development and use of valid tests of academic language ability for channelling these students into academic language interventions that are aimed at dealing with this challenge. The literature on the current theories of validity is also explored in relation to the hypothesis of the study, which is that as a test designed to provide additional information about the academic language readiness of first year students, TALL will possess better incremental validity in relation to the best predictor of first year academic performance.

Subsequently, an attempt is made to account for the ability of Grade 12 English results to predict first year academic performance better than the other three assessments investigated in the present study. Similarly, an effort is made to account for the ability of

TALL to show evidence of incremental validity in relation to Grade 12 English results and the inability of NBT AL and PTESLAL to do the same. Furthermore, on the basis of the results of previous studies and the current one, a recommendation is made that Grade 12 results in general and Grade 12 English results in particular be used together with those of academic literacy tests to make access and placement decisions. The basis for this recommendation resides in the psychometric and other shortcomings of Grade 12 results that have been identified by previous studies as well as the evidence that similar studies have produced to show that tests of academic literacy possess better ability to partition test taker performance from different school backgrounds and at different levels of performance.

Finally, the implications of the results of the study for current theories of test validity are discussed. In the main, the discussion focuses on demonstrating on the basis of these results and those of previous studies that the currently popular theory of validity wherein a unitary approach to validity is upheld and the interpretation and use of test scores are regarded as the essence of validation does not hold. At the same time, the discussion focuses on demonstrating that the traditional theory of validity, wherein validity is believed to reside in the objective ability of a test to produce valid scores and a distinction is made between the three traditional types of validity, namely construct, content and criterion-related validity is, with certain obvious qualifications, still defensible. Finally, the implications of the results of the study for validity theory are dealt with in relation to the validity of courses of academic literacy.

**Key terms:** academic literacy, incremental validity, validity theory, NBT AL, PTESLAL, TALL, Grade 12 English, academic performance, academic under-preparedness.

## Opsomming

Hierdie studie fokus op die inkrementele geldigheid van drie assesserings van akademiese taalgereedheid, te wete die Nasionale Normtoets vir Akademiese Geletterdheid (NBT AL), die Proficiency Test English Second Language Advanced Level (PTESLAL), en die Test of Academic Literacy Levels (TALL), in vergelyking met die Graad 12-resultate vir Engels. Die studie ondersoek in die besonder die vermoë van elk van die genoemde toetse om die akademiese prestasie van eerstejaars beter te kan voorspel as die ander. Waar die toetse wat ondersoek word nie daardie vermoë het nie, word daar 'n verdere vraag gevra, naamlik of die toets ten minste bydra tot die voorspellingsvermoë van die beste voorspeller? Die uiteindelige doel is om te bepaal of assesserings wat ontwerp is om addisionele inligting te bied oor die voorbereidheid van studente vir eerstejaarstudie geldig is vir hierdie doel, en tot watter mate dit die geval is.

Die ondersoek begin met 'n bondige verkenning van die literatuur oor die beweerde lae vlakke van akademiese taalvaardigheid by nuwelingsstudente aan instellings van hoër onderwys in Suid-Afrika die afgelope aantal jaar, sowel as die gevolglike behoefte aan die ontwikkeling en aanwending van geldige toetse van akademiese taalvermoë ten einde daardie studente te kan kanaliseer na akademiese taalintervensies wat ten doel het om hierdie uitdaging te bowe te kom. Die literatuur oor huidige teorieë oor geldigheid word ook verken en binne konteks van die hipotese van die ondersoek. Dit is naamlik dat as 'n toets wat ontwerp is om addisionele inligting oor die akademiese taalgereedheid van studente te verskaf,



TALL 'n groter moontlikheid bied om inkrementele geldigheid te vertoon in verhouding tot die beste aanwyser van eerstejaars se akademiese prestasie.

Vervolgens word daar gepoog om 'n verklaring te verky vir die vermoë van die Graad 12-resultate in Engels om eerstejaars se akademiese prestasie beter te kan voorspel as enige ander assessering wat in hierdie studie gebruik is. Op dieselfde wyse word daar gepoog om die nodige bewyse te vind vir die vermoë van TALL om inkrementele geldigheid toe te voeg tot die Graad 12-resultate, asook vir waarom die ander twee assesserings, NBT AL en PTESLAL, dit nie het nie. Op basis van die resultate van vorige studies, asook van hierdie ondersoek, word die aanbeveling gemaak dat die Graad 12-resultate in die algemeen, en die Graad 12-resultate vir Engels in die besonder, tesame met die resultate van 'n akademiese geletterdheidtoets gebruik moet word om besluite te neem oor toegang en plasing. Die gronde vir hierdie aanbeveling lê in die psigometriese en ander tekortkomings van die Graad 12-resultate, wat geblyk het uit vorige studies, en wat ook geïdentifiseer is deur ondersoeke soortgelyk aan hierdie een. Uit hierdie ondersoeke blyk dit dat toetse van akademiese geletterdheid veral goed kan vaar om te onderskei tussen diegene wat die toets deurloop het se onderskeie opvoedkundige agtergronde en hul verskillende vlakke van prestasie.

Ten besluite word die implikasies van die studie se resultate oorweeg betreffende huidige teorieë van toetsgeldigheid. Die bespreking fokus hoofsaaklik daarop om aan te toon dat hierdie resultate en die van soortgelyke studies 'n basis bied waarteen populêre geldigheidsteorie bevraagteken sou kon word, in die besonder

aannames onderliggend aan die idee van een saambindende aanpak om geldigheid te bepaal, asook om die interpretasie van toetsresultate as die essensie van geldigmaking voor te hou. Terselfdertyd dui die bespreking daarop dat die tradisionele siening van geldigheid, waarin dit geag word verbind te wees aan die objektiewe vermoë van 'n toets om geldige resultate te lewer, steeds met sekere voorbehoude regverdigbaar is, asook die onderskeid wat getref kan word tussen die tradisionele drietal tipes geldigheid, naamlik konstruk-, inhouds- en kriteriumgeldigheid. Ter afsluiting word die implikasies van die resultate van hierdie ondersoek vir geldigheidsteorie in verband gebring met die geldigheid van akademiese geletterdheidskursusse.

**Sleutelwoorde:** akademiese geletterdheid, inkrementele geldigheid, geldigheidsteorie, NBT AL, PTESLAL, TALL, Graad 12 Engels, akademiese prestasie, akademiese voorbereidheid.

## **Chapter 1: Introduction: The importance of academic language assessment for first year academic study in South Africa**

The aim of this study is to investigate the incremental validity of four English language assessments at a university of technology in South Africa. In the context of South Africa, universities of technology are those that, unlike traditional academic universities, mainly offer diploma programmes, the admission requirements for which are lower than those for degree programmes, which are mainly offered by traditional academic universities. Furthermore, a large proportion of the programmes offered by the former covers experiential learning whose aim is to enhance immediate employability post-graduation. This is not the case with most degree graduates from traditional academic universities who often leave these institutions without the experience required for the jobs they are aiming for. The first of these differences implies that the level of academic demand placed on students at universities of technology is lower than that faced by students at traditional academic universities. The results of a study such as this one, where the focus is on a university of technology, should therefore not necessarily be generalizable to traditional academic universities.

Although they are designed and developed on the basis of different constructs, the assessments investigated in this study are used to measure students' ability to handle university education in the language of teaching and learning, a competence commonly known as academic literacy. In other words, the study is aimed at determining if any of these assessments predicts academic success better than the

others and whether such an assessment or any of the four does this from an additional or unique perspective in relation to the best predictor. The four assessments that will be investigated in this study are the Grade 12 English Home Language (HL) and First Additional Language (FAL) examinations developed by the Department of Basic Education, the Proficiency Test English Second Language Advanced Level (PTESLAL) of the Human Sciences Research Council (HSRC), the Test of Academic Literacy Levels (TALL) developed by the Inter-institutional Centre for Language Development and Assessment (ICELDA), and the National Benchmark Test in Academic Literacy (NBT AL) developed under the auspices of the National Benchmark Tests Project (NBTP) by the Centre for Educational Testing for Access and Placement at the University of Cape Town.

Some of these assessments are used for high stakes decisions such as access and others for medium to low stakes decisions such as placement on language development interventions at tertiary institutions. Both these purposes link the assessments directly to student retention and academic success at university. In other words, whether they are used for placement or access decisions, these assessments are ultimately used for predicting the academic success of the students who take them. Given the importance of academic success to the students, the universities involved, and the country at large, it is necessary that these assessments are investigated for their ability to serve the purpose for which they are used.

## 1.1 Background to the problem

The present study is undertaken in the context of the low levels of academic literacy revealed by several studies among first year students at South African universities in the past 20 to 30 years (Van Rensburg & Weideman 2002) and the resultant need for these universities to deal with this challenge. Rambiritch (2012a: 1) has observed that the low levels of academic literacy are “... a problem not specific only to students from previously disadvantaged backgrounds. Language proficiency is low even amongst students whose first language is English and Afrikaans, which are still the main languages of teaching and learning at tertiary level.”

Indeed, Van Wyk and Yeld (2013) have pointed out that gaining access to university means that students have to acquire academic literacy. In Bourdieu and Passeron’s (1990: 66) view, the ability to handle academic discourse is difficult because that kind of language is nobody’s native language. This means that newly admitted students need to learn new ways of “saying (writing) – doing – being – valuing – believing combinations” (Gee 1996: 127). Gee (1990: 1) describes the process of acquiring this ‘new’ and “secret language” (Pennycock 1999: 330) and of ultimately becoming part of the academic community as follows:

You learn the discourse by becoming a member of the group: you start as a ‘beginner’, watch what’s done, go along with the group as if you know what you are doing when you don’t, and eventually you can do it on your own.

It is this strangeness and novelty of learning a ‘new’ language that commentators often identify as the root of the low levels of mastery of academic discourse among new entrants to the world of higher education. Furthermore, these low levels of

competence in the language of academic discourse have had “a detrimental effect on students’ academic development, leading to poor pass rates” (Rambiritch 2012a: 4). Butler (2006: 2) observes that “the difficulty of engaging successfully in tertiary study in South Africa through an additional language (English) that one has not acquired adequately is well documented”. Van Dyk (2005: 38) has added that “low levels of academic literacy in the language of learning are widely seen as one of the main reasons for the lack of academic success among South African undergraduate students with high academic potential”. In support of this, Barry (2002: 106) has argued that language ability and academic achievement are

inextricably linked and the use of English as the language of learning and teaching by the majority of second language learners in South African schools should be seen as a major contributor to the poor pass rates and dropout rates of learners throughout the education system.

The challenge of low academic literacy levels among first year students has grown with the advent of a democratic dispensation in South Africa in 1994. The new democratic constitution (Act 108 of 1996) enshrines the rights of citizenship and equality for all, irrespective of race (Mdepa & Tshiwula 2012: 21). Section 29 of this constitution promotes the right of all races to access all levels of education in South Africa and Section 29.2 (c) “refers to the need to redress the results of past discriminatory laws and practices that institutionalised difference” (Mdepa & Tshiwula 2012: 21). It was in the spirit of this new constitution and the need for a reformation and restructuring of the South African education system that, in his foreword to the National Plan for Higher Education, Kader Asmal, the then Minister of Education, wrote that

The victory over the apartheid state in 1994 set policy makers in all spheres of public life the mammoth task of overhauling the social, political, economic and

cultural institutions of South Africa to bring them in line with the imperatives of a new democratic order (Ministry of Education 2001).

This meant, among others, that the whole system of education in the country would have to be overhauled. The plan for bringing this into effect was subsequently documented in the Education White paper 3: A Programme for the Transformation of Higher Education (Department of Education 1997). The major aim of the programme was “the establishment of a single, national co-ordinated system, which would meet the learning needs of our citizens and the reconstruction and development needs of our society and economy” (Department of Education 1997). The publication of White Paper 3: A Programme for the Transformation of Higher Education meant, *inter alia*, that university education became accessible to more students both from historically advantaged and disadvantaged backgrounds than was the case previously. The influx of these students added to the number of academically illiterate students who had started entering universities in greater numbers a decade or so ago, prior to the advent of a democratic order in 1994.

In a context of language competence levels that may be inadequate, and a massification of higher education, university authorities needed to respond to this dual challenge. This has prompted universities to introduce academic literacy intervention programmes for these students to boost their chances of success at university. In the words of Rambiritch (2012a: 5), “tertiary institutions, especially those considered previously advantaged, today need contingency measures to deal with this situation”. Similarly, Van Wyk and Yeld (2013: 62) have argued that the fact that the medium of instruction at South African universities is an additional

language to most students means that “universities face a significant challenge – that of providing effective and meaningful language learning and development opportunities for the great majority of their students ...” These interventions have, however, had to be preceded by academic literacy testing to channel students who need this intervention into the appropriate course. Cliff, Yeld and Hanslo (2003: 1) have justified this in their observation that

It is no longer possible (nor perhaps desirable) to assume traditional student bodies in traditional higher education systems. In order to grant access and – as far as possible – contribute to success, higher education institutions are faced with the need to identify student applicants with at least a reasonable measure of potential for coping with the demands of academic study.

Cliff, Yeld and Hanslo (2003: 1) add that the assessment of students for the purpose of access “appears to carry with it a concomitant need for institutions to understand and cater for the needs of their students in terms of curriculum structures, learning support, teaching interventions and appropriate preparation to meet assessment forms and requirements”. It is crucial, therefore, that both the academic literacy tests and the interventions offered on the basis of the scores they yield are underpinned by a defensible theory of academic literacy. In other words, it is important that the academic literacy tests used for taking this placement decision are informed by a construct of academic literacy that is justifiable and that the interventions following the assessment do exactly what they are designed for: addressing the academic literacy needs of the students. Indeed, Patterson and Weideman (2013a: 107) have argued that “constructs of academic literacy are used both for test and course design”. Van Dyk and Weideman (2004: 141) have further explained that “a construct is usually articulated in terms of a theory, in our case, a theory of language, and more specifically, a theory of academic literacy.”



## 1.2 A definition of academic literacy in the South African higher education context

The first step towards achieving a meaningful definition of academic literacy is to arrive at some understanding of the nature of academic discourse and whether it is a different type of discourse. The suggestions by Bourdieu and Passeron (1990), Gee (1990, 1996) and Pennycook (1990), referred to above, seem to validate that it is, as do those of Cummins (1984, 1996, 2009) and Cummins and Swain (1986), that will be referred to again below. In the words of Patterson and Weideman (2013a: 108), “definitions of the ability to handle academic discourse that explicitly derive from an idea of what academic discourse entails, and how it differs from other types of discourse, are not only easier to engage with critically, but also potentially more useful.” Patterson and Weideman (2013b: 126) elaborate this point further:

... there is probably no better starting point than firstly to determine whether academic discourse is a distinct type of discourse and secondly, what it is that makes it different from other lingual spheres. By a lingual sphere, we mean a distinctly different kind of language that is used within a particular social institution, so that the language of business, for example, will differ from that of an intimate relationship, or the language of worship will differ from the language of the court, or the language of literature will differ from the language of education.

Recently, Patterson and Weideman (2013b) have argued that what makes academic discourse different from other types of discourse is that it essentially requires analytical and logical thinking to be processed efficiently. In their view (Patterson and Weideman 2013b: 137), the typicality of this discourse derives from “the (unique) distinction-making associated with the analytical or logical mode of experience” and these need to be emphasized in any attempt to define academic literacy. Patterson and Weideman (2013a: 111) capture this view as follows:

It is evident that the typicality of academic discourse is stamped or guided by a specific dimension of experience – namely, the analytical. While each

academic field is circumscribed by one or more modes of reality ... academic discourse as a whole is qualified by the analytical (or logical) mode, which is usually historically grounded. In other words, work within every academic discipline ... is guided and led by the logical dimension of experience which involves analysis as its defining kernel.

Patterson and Weideman (2013b: 137) argue that while “distinction-making and analytical or logical thinking are ... a component” of the constructs of academic literacy advanced by Cliff and Yeld (2006), Van Dyk and Weideman (2004) and Cummins (1984, 1996, 2000), that are dealt with later in this chapter, the distinction-making, analytical and logical characteristics of academic literacy are not sufficiently foregrounded in the definitions of such constructs. Patterson and Weideman (2013a:138) have, for this reason, suggested three kinds of modifications to how these constructs are defined:

First, an emphasis on the analytical nature of academic language, which is missing from the initial formulation; second, an augmentation of the construct by articulating components that may have been implied, but that are certainly overt; third, a more patent grasp of the nature of academic interaction through language, which might include analytical information gathering, processing and production, or what is conventionally conceived of as listening, writing, reading, and speaking ... or what another reviewer has called cognitive processing.

This has resulted in Patterson and Weideman (2013b) reformulating the constructs of academic literacy referred to earlier. Their modified definition of academic literacy suggests that this notion should be defined as students’ ability to do the following:

- Think critically (analyse the use of techniques and arguments) and reason logically and systematically in terms of one’s own research and that of others;
- Distinguish between essential and non-essential information, fact and opinion, propositions and arguments, cause and effect, and classify, categorize and handle data that make comparisons;
- Interact (both in speech and writing) with texts; discuss, question, agree/disagree, evaluate, research and investigate problems, analyse, link texts, draw logical conclusions from texts, and then produce new texts;

- Synthesize and integrate information from a multiplicity of sources with one's own knowledge in order to build new assertions, with an understanding of academic integrity and the risks of plagiarism;
- Understand relations between different parts of a text, be aware of the logical development and organization of an academic text, via introductions to conclusions, and know how to use language that serves to make the different parts of a text hang together.
- Know what counts as evidence for an argument, extrapolate from information by making inferences, and apply the information or its implications to other cases than the one at hand;
- Think creatively: imaginative and original solutions, methods or ideas which involve brainstorming, mind-mapping, visualization, and association;
- Interpret, use and produce information presented in graphic or visual format;
- Understand and use a range of academic vocabulary as well as content or discipline-specific vocabulary in context;
- Interpret the use of metaphor and idiom in academic usage, and perceive connotation, word play and ambiguity;
- Interpret different kinds of text type (genre), and have a sensitivity for the meaning they convey, as well as the audience they are aimed at;
- Use specialized or complex grammatical structures, high lexical diversity, formal prestigious expressions, and abstract/technical concepts which can also function as agents;
- Make meaning (e.g. of an academic text) beyond the level of a sentence;
- See sequence and order, and do simple numerical estimations and computations that are relevant to academic information, that allow comparisons to be made, and can be applied for the purpose of an argument;
- Interpret and adapt one's reading/writing for an analytical/argumentative purpose and/or in light of one's experience;
- Understand the communicative function of various ways of expression in academic language (such as defining, providing examples, inferring, extrapolating, arguing); and
- Write in an authoritative manner, which involves the presence of an imagined audience of specialists/novices (young researchers)/general public/media.

(Patterson & Weideman 2013b: 139-140)

This reformulation is informed by the typicality of the nature of academic discourse and adequately foregrounds analytical and logical thinking as well as the distinction-making characteristic that was not adequately articulated in previous definitions of academic literacy (Patterson & Weideman 2013b).

Patterson and Weideman's (2013b) argument for the typicality of academic discourse and the resultant uniqueness of the literacy skills required to handle this

discourse is reminiscent of the distinction that Cummins (1980, 1984, 2009) makes between the type of language skills required for conversational language and those that enable one to handle university education in the language of learning and teaching. Cummins refers to these types of competence as the Basic Interpersonal Communicative Skills (BICS) and Cognitive Academic Language Proficiency (CALP) respectively. Cummins (1984, 1996, 2009) has argued that CALP takes six to eight years to acquire through the mother tongue at school level. Van Wyk and Yeld (2013: 66) have argued for the role of a learner's first language in the acquisition of CALP thus:

The student studying in the mother tongue makes considerable CALP gains with each passing school year, thus widening the gap between those studying in their mother tongue and those who are not. The destructive role that this plays in students' success at university is far reaching as students need a good grounding in CALP to enable them to acquire the academic literacy required in higher education.

In the words of Cummins (2009: 22), a growth in CALP at this level of schooling “requires expansion of vocabulary, grammatical and discourse knowledge far beyond what is required for social communication”. In the same breath, Alidou, Aliou, Brock-Utne, Diallo, Heugh and Wolff (2006: 15) have added that “the development of the type of literacy necessary for reading and writing about science, history and geography, or understanding problems in mathematics, becomes increasingly complex from the fourth year of school onwards.” The distinction Cummins (1984) makes between what he calls BICS and CALP prompted Cummins and Swain (1986: 151) to conclude that it is “necessary to distinguish between the processing of language in informal everyday situations and the language processing required in most academic situations”, a point that Patterson

and Weideman (2013a, 2013b) make with regard to the uniqueness of academic discourse and by extension, the skills believed to constitute academic literacy.

It is clear, therefore, that the language assessments that must be employed to measure competence in dealing with this type of discourse and which are the focus in the present study should be theoretically defensible in terms of a definition or construct of the ability to handle academic discourse as a distinct type of language. The perspective of academic discourse held by the test designer is critical because “a test is always produced for a specific purpose, and ... its results inevitably influence decisions about the future of the candidates that take it” (Van Dyk & Weideman 2004: 139). It is important, for this reason, that those who develop tests are “able to demonstrate how performance on that language test is related to language use in specific settings” (Bachman & Palmer 1996: 61). For the purpose of providing a meaningful context for the present study, the constructs of the four assessments that will be investigated in the study as well as the language perspectives informing them are explored below to clarify the extent to which they are aligned to the notion of academic literacy as a unique kind of language ability.

### **1.3 Constructs of the four assessments**

#### **1.3.1 The construct of the National Benchmark Test in Academic Literacy**

The low academic literacy levels among first year university students have led to efforts to generate constructs of academic literacy for purposes of measuring and, by extension, teaching academic literacy at universities in South Africa. Firstly, as a basis for the design and development of the National Benchmark Test in

Academic Literacy, Cliff and Yeld (2006: 19) have argued for a construct of academic literacy that focuses on

students' capacities to engage successfully with the demands of academic study in the medium of instruction of the particular study environment. In this sense, success is constituted of the interplay between the language (medium of instruction) and the academic demands (typical tasks required in higher education) placed upon students.

Their definition is, in the view of Cliff and Yeld (2006), informed by Bachman and Palmer's (1996) view of language ability. As will be demonstrated again in Chapter Two below, Bachman and Palmer (1996) view language ability as being constituted by what they call language knowledge and strategic competence. Language knowledge itself consists of two broad categories, namely, organizational and pragmatic knowledge. Bachman and Palmer (1996: 68-69) have defined these categories thus:

organizational knowledge is involved in controlling the formal structure of language for producing or comprehending grammatically acceptable utterances or sentences, and for organizing these to form texts, both oral and written .... pragmatic knowledge enables us to create or interpret discourse by relating utterances or sentences and texts to their meaning, to the intentions of language users, and to relevant characteristics of the language use setting.

Bachman and Palmer's concept of language knowledge and its constituents are captured in **Table 1** below.

**Table 1: Bachman and Palmer's areas of language knowledge**

---

**Organizational knowledge**

(how utterances or sentences and texts are organized)

*Grammatical knowledge*

(how individual utterances or sentences are organized)

Knowledge of vocabulary

Knowledge of syntax

Knowledge of phonology/graphology

*Textual knowledge*

(how utterances or sentences are organized in texts)

Knowledge of cohesion

Knowledge of rhetorical or conversation organization

**Pragmatic knowledge**

(how utterances or sentences and texts are related to the communicative goals of the language user and to the features of the language use setting)

*Functional knowledge*

(how utterances or sentences and texts are related to the communicative goals of language users)

Knowledge of ideational functions

Knowledge of manipulative functions

Knowledge of heuristics functions

Knowledge of imaginative functions

*Sociolinguistic knowledge*

(how utterances or sentences and texts are related to features of the language use setting)

Knowledge of dialects/varieties

Knowledge of registers

Knowledge of natural or idiomatic expressions

Knowledge of natural or idiomatic expressions

Knowledge of cultural references and figure of speech

---

(Bachman & Palmer 1996: 68)

Strategic competence, on the other hand, 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 of language use, as well as other cognitive activities” (Bachman & Palmer 1996: 70). These metacognitive processes involve goal-setting, assessment and planning (Bachman & Palmer 1996). The details of Bachman and Palmer’s notion of strategic competence are covered in **Table 2** below:

**Table 2: Bachman and Palmer’s areas of strategic competence**

---

**Goal setting**

(deciding what one is going to do)

Identifying the test tasks

Choosing one or more tasks from a set of possible tasks (sometimes by default, if only one task is understandable)

Deciding whether or not to attempt to complete the task (s) selected

**Assessment**

(taking stock of what is needed, what one has to work with, and how well one has done)

Assessing the characteristic of the test task to determine the desirability and feasibility of successfully completing it and what is needed to complete it

Assessing our own knowledge (topical, language) components to see if relevant areas of knowledge are available for successfully completing the test task

Assessing the correctness or appropriateness of the response of the test task

**Planning**

(deciding how to use what one has)

Selecting elements from the areas of topical knowledge and language knowledge for successfully completing the task

Formulating one or more plans for implementing these elements in a response to the test task

Selecting one plan for initial implementation as a response to the test task

In Bachman and Palmer's (1996: 70) view, language knowledge, strategic competence, topical knowledge and affective schemata interact to make language use possible. Bachman and Palmer (1996: 70) explain this as follows:

using language involves the language user's topical knowledge and affective schemata, as well as all other areas of knowledge discussed above. What makes language use possible is the integration of all these components as language users create and interpret discourse in situationally appropriate ways.

Cliff, Yeld and Hanslo (2003: 3) have argued along similar lines that academically literate students are "by implication, those who are able to negotiate the grammatical and textual structure of the language of instruction and to understand its functional and sociolinguistic bases". It is on the basis of this that Cliff and Yeld (2006: 20) have formulated the construct underpinning the NBT AL as a student's ability to do the following:

- negotiate meaning at word, sentence, paragraph and whole-text level;
- understand discourse and argument structure and the text 'signals' that underlie this structure;
- extrapolate and draw inferences beyond what has been stated in text;
- separate essential from non-essential and super-ordinate from sub-ordinate information;
- understand and interpret visually encoded information, such as graphs, diagrams and flow-charts;
- understand and manipulate numerical information;
- understand the importance and authority of own voice;
- understand and encode the metaphorical, non-literal and idiomatic bases of language; and
- negotiate and analyse text genre.

Clearly, this construct captures the views presented earlier on the nature of academic discourse and how the ability to handle it should be defined. As pointed out earlier, however, the construct does not clearly foreground the analytical,



logical and distinction-making character of the ability to handle academic texts that Patterson and Weideman (2013b) emphasize. The details of the specifications and task types arising from the construct underpinning the NBT AL are provided later in Chapter Three.

### **1.3.2 The construct of the Test of Academic Literacy Levels**

The Test of Academic Literacy Levels was conceptualized and developed mainly on the basis of the construct of academic literacy advanced by Van Dyk and Weideman (2004). In agreement with the view of Cliff and Yeld (2006) presented earlier, Van Dyk and Weideman (2004) have also formulated a construct of academic literacy that is informed by the Bachman and Palmer (1996) model of language ability. Furthermore, Van Dyk and Weideman (2004) agree with the view of Bachman and Palmer (1996) that academic literacy should not be interpreted in terms of ‘skills’. Bachman and Palmer (1996) have argued against a skills-oriented definition of language ability on the grounds that while language tasks such as face to face conversation and listening to a radio newscast, for example, all involve listening, they involve other abilities associated with language use in general and can therefore not be confined to the ‘skill’ of listening. Bachman and Palmer (1996: 75) argue on this basis, therefore, that

We would thus not consider language skills to be part of language ability at all, but to be the contextualized realization of the ability to use language in the performance of specific language use tasks. We would argue ... that it is not useful to think in terms of ‘skills’, but to think in terms of specific activities or tasks in which language is used purposefully.

Weideman (2013: 13) has similarly argued that “we no longer stick to the behaviourist belief, so ably embodied in the audio-lingual method and its conventional predecessors, that listening, speaking, reading and writing are separate

or even separable language ‘skills’.” Kumaravadivelu (2003: 226) has also argued that the integrated nature of listening, speaking, reading and writing makes it a worthless exercise to try and separate them into ‘skills’.

The Bachman and Palmer (1996) model of language ability and the views advanced by Blanton (1994), Cummins (1984, 1996, 2009) and Patterson and Weideman (2013b) on the nature of academic language ability echo the communicative approach to language teaching that came into being in the late 1960s and 1970s to replace the structural-situational and audio-lingual methods that preceded it (Richards 2001). Communicative Language Teaching (CLT) refers to “a broad approach to teaching that resulted from a focus on communication as the organizing principle for teaching rather than a focus on mastery of the grammatical system of the language” (Richards 2001: 36). CLT introduced to the language teaching profession a focus on “how language is used by speakers in different contexts of communication” as opposed to the focus “on grammar as the core component of language abilities” of its predecessors (Richards 2001: 36). To this end, CLT is consistent with an open as opposed to a restrictive view of language ability (Van Dyk & Weideman 2004) that underpinned its predecessors. The details of these perspectives are dealt with in Chapter Two below.

Van Dyk and Weideman (2004) have added further that Blanton’s(1994) view of academic literacy was also useful in their formulation of the construct of academic literacy underpinning TALL. Along the lines of Bachman and Palmer (1996),

Blanton (1994: 228) has argued that academic literacy involves a student's ability to interact with academic texts:

Whatever else we do with L2 students to prepare them for the academic mainstream, we must foster the behaviour of 'talking' to texts, talking and writing about them, linking them to other texts, connecting them to their own lives and experiences, and then using their experiences to illuminate the text and the text to illuminate their experiences.

Van Dyk and Weideman (2004) are of the view that, like the Bachman and Palmer (1996) perspective of language ability, Blanton's construct of academic literacy also contradicts a restrictive and outdated view of language ability which foregrounds the teaching of grammar and vocabulary. Specifically, Blanton (1994: 226) argues that an academically literate student should be able to do the following:

1. Interpret texts in the light of their own experience and their own experience in the light of texts;
2. Agree or disagree with texts in the light of that experience ;
3. Link texts to each other;
4. Synthesize texts, and use their synthesis to build new assertions;
5. Extrapolate from texts;
6. Create their own texts, doing any or all of the above;
7. Talk and write about texts doing any or all of the above;
8. Do numbers 6 and 7 in such a way as to meet the expectations of the audience.

(Blanton 1994: 226)

On the basis of all these perspectives, Van Dyk and Weideman (2004) have generated a construct of academic literacy which underpins TALL. According to this construct, academically literate students should be able to do the following:

- Understand a range of academic vocabulary in context;
- Interpret and use metaphor and idiom, and perceive connotation, word play and ambiguity;
- Understand relations between different parts of a text, be aware of the logical development of (an academic) text, via introductions to conclusions, and know how to use language that serves to make the different parts of a text hang together;
- Interpret different kinds of text type (genre), and show sensitivity for the meaning that they convey, and the audience that they are aimed at;
- Interpret, use and produce information presented in graphic or visual format;
- Make distinction between essential and non-essential information, fact and opinion, propositions and arguments; distinguish between cause and effect, classify, categorize and handle data that make comparisons;

- See sequence and order, do simple numerical estimations and computations that are relevant to academic information, that allow comparisons to be made, and can be applied for the purpose of an argument;
- Know what counts as evidence for an argument, extrapolate from information by making inferences, and apply the information or its implications to other cases than the one at hand;
- Understand the communicative function of various ways of expression in academic language (such as defining, providing examples, arguing); and
- Make meaning (e.g., of an academic text) beyond the level of sentence.

Clearly, the construct underpinning TALL is also informed by current views on how academic literacy should be defined. Like that informing the NBT AL, however, TALL's current construct also does not foreground the analytical, logical and distinction making characteristics of academic language ability (Patterson & Weideman 2013b). The specifications arising from this construct are also dealt with in detail in Chapter Three below.

### **1.3.3 The construct of the English HL and FAL examinations**

The Curriculum and Assessment Policy Statement (CAPS) stipulates that the curriculum underpinning the National Senior Certificate (NSC) examinations should help students become participants in “society as citizens of a free country”, gain “access to higher education” and help them transition from “education institutions to the workplace” (Department of Basic Education 2011: 4). In addition, CAPS provides a list of specific aims for Home Language (HL) and First Additional Language (FAL) learning at school level. Learning an HL and FAL should, according to CAPS, enable learners to:

- Acquire the language skills required for academic learning across the curriculum;
- Listen, speak, read/view and write/present the language with confidence and enjoyment. These skills and attitudes form the basis for life-long learning;
- Use language appropriately, taking into account audience, purpose and context;
- Express and justify, orally and in writing, their own ideas, views and emotions confidently in order to become independent and analytic thinkers;

- Use language and their imagination to find out more about themselves and the world around them. This will enable them to express their experiences and findings about the world orally and in writing.
- Use language to access and manage information for learning across the curriculum and in a wide range of contexts. Information literacy is a vital skill in the ‘information age’ and forms the basis for life-long learning; and
- Use language as a means for critical and creative thinking; for expressing their opinions on ethical issues and values; for interacting critically with a wide range of texts; for challenging the perspectives, values and power relations embedded in texts; and for reading texts for various purposes, such as enjoyment, research and critique.

(Department of Basic Education 2011: 9).

Furthermore, with regard to language teaching and learning in particular, CAPS aims at promoting learners’ achievement at two levels of language proficiency, namely, the social and educational or academic levels. The social level focuses on “the mastery of basic interpersonal communication skills required in social situations” while the educational or academic level targets “cognitive academic skills essential for learning across the curriculum” as well as “literary, aesthetic and imaginative ability”. To this end, the construct of language ability that underpins CAPS “articulates the intention to develop in learners differentiated language ability so that by the end of their school careers they have mastery of language(s) in a wide range of contexts and situations (educational and academic; aesthetic, political; economic; social and informational; ethical)” (Du Plessis, Steyn & Weideman 2016: 7).

The construct of language ability described in CAPS is without question consistent with the way academic discourse and academic literacy are conceptualized by Cliff and Yeld (2006), Van Dyk and Weideman (2004) as well as Cummins (1984, 1996 & 2009) above. Like the construct of the two tests dealt with earlier, however, the analytical, logical and distinction-making dimensions of academic literacy that

Patterson and Weideman (2013b) propose are not clearly foregrounded in the construct of language ability espoused in CAPS for HL and FAL instruction. As a matter of logic, the language view informing the curriculum dimension of CAPS should also be the foundation on which the assessment of the English HL/FAL is based in the National Senior Certificate (NSC) examinations. The very name of this policy statement presupposes an alignment between the curriculum and assessment accompanying it. Whether this is, in fact, the case is a matter I shall return to in subsequent chapters of this study. The specifications that derive from the construct advanced for the English HL and FAL examinations are also provided in detail in Chapter Three below.

#### **1.3.4 The construct of the Placement Test English Second Language Advanced Level**

The Placement Test English Second Language Advanced Level is a test of English proficiency that was designed by the Human Science Research Council (HSRC) “in response to the perceived needs of education departments and various sectors of South African society” to measure the test takers’ “level of general language development” (HSRC 1991: 15). The HSRC (1991: 15) defines the purpose of a proficiency test such as PTESLAL as follows:

The purpose of a proficiency test is to determine a testee’s knowledge and skill regarding a defined field of experience or subject matter not attached to a specific syllabus. It is fairly self-evident that language proficiency levels are not attained solely as a result of curricular activities, but also as a result of extra-curricular language contact and use.

It is deducible from this that while the test has been used for deciding whether first year students possess the competence required to handle academic discourse successfully at one South African university, its designers describe it as a test of

general language development and not necessarily that of the ability to handle academic discourse in the language of teaching and learning. The construct of this test is therefore not consistent with the definitions of academic discourse and the constructs of academic literacy proposed for the three tests dealt with above. The specifications arising from the construct of the PTESSLAL are also dealt with in details in Chapter Three below.

#### **1.4 Problem statement**

Language testing is a fundamental component of the English language teaching profession at all levels of education worldwide. At tertiary institutions in particular, language test scores are often used for the selection and placement of students, and as a tool to assess learning progress and diagnose learning difficulties (Bachman & Palmer 1996: 96-97). Language test scores are also used by some universities for student certification, language programme evaluation, and teacher professional development (Bachman & Palmer 1996: 96-97). These are high stakes purposes for testing that make it incumbent upon language test designers and developers to ensure that their tests have a high degree of validity, which begins with a construct or definition of the ability to be measured. This is extremely important, furthermore, because this kind of testing has consequences for educational systems, individuals within those systems and society at large (Bachman & Palmer 1996: 34). The developers of the tests investigated in this study are no exception to these concerns.

The English HL and FAL examinations are high stakes tests that are developed by the Department of Basic Education and used to decide whether Grade 12 learners in

South Africa graduate from high school. They also signify if these learners' performance in these exams is good enough to enable them to gain admission to post-Grade 12 institutions of higher learning such as colleges and universities. Also, starting from 2014, the University of Pretoria uses these examinations to channel first year students enrolling in degree programmes in the Humanities Faculty into appropriate academic literacy courses. In this faculty, students whose English HL scores are at level 4 or lower are required to register for two academic literacy modules, namely, ALL 110 and ALL 125 while those whose English FAL scores are at level 5 and lower have to do the same. This is the case notwithstanding the observation by Cliff, Yeld and Hanslo (2003: 2) that

In a country such as South Africa, for instance, school-leaving certification has had a particularly unreliable relationship with Higher Education academic performance especially in cases where this certification intersects with factors such as mother tongue versus medium-of-instruction differences, inadequate school-backgrounds and demographic variables such as race and socio-economic status.

It is for reasons similar to those Cliff, Yeld and Hanslo (2003) advance above that Du Plessis, Steyn and Weideman (2016: 2) have observed that the English HL and FAL exams “cannot be regarded as fair and equal assessments” and add that it is clear from the several reports commissioned by Umalusi, the Council for Quality Assurance in General and Further Education and Training, on these examinations that “the quality and standard of the assessment in the exit-level examinations needs urgent scrutiny”. The primary reason for this shortcoming is, as I will again argue in Chapter Five, that the construct underlying these exams is not clearly defined (Du Plessis, Steyn & Weideman 2016). This lack of clarity in what it is that these exams were developed to measure is likely to result in a misalignment



between them and the curriculum, and constitutes a clear threat to the validity of the exams.

TALL is a test of academic literacy developed by the Inter-institutional Centre for Language Development and Assessment (ICELDA), a partnership of four multilingual universities, namely, Pretoria, Stellenbosch, North West and Free State (Le, du Plessis & Weideman 2011). The test has been used by the four partners and other South African universities to measure the levels of academic literacy of first year students for the purpose of placement and, on a small scale, admission at such universities.

The NBT AL was an outcome of the National Benchmark Tests Project (NBTP) that was initiated by Higher Education South Africa (HESA) and currently operates within the Centre for Educational Testing for Access and Placement of the University of Cape Town. The original aim of the National Benchmark Tests (NBTs) was to measure the test taker's levels of academic, quantitative and mathematical literacy. Ultimately, the purpose was to provide information to tertiary institutions regarding the level of academic preparedness of school-leavers in order to assist such institutions to determine their curriculum needs and ensure that they are properly placed within the institutions. The NBTs are, however, as I note again in Chapter Six, administered in the year preceding a student's admission to university and have, for that reason, been used by some universities for making access decisions, and have consequently assumed the status of high-stakes tests.

The NBT AL is written by applicants for admission to all programmes offered by a university.

Finally, the PTESLAL is a test of English proficiency that was designed by the HSRC. This is a high stakes test used by the Central University of Technology (CUT) for selecting students who apply for admission to the university but whose performance in the Grade 12 examinations does not satisfy the admission requirements of the university. At CUT, the requirement for straight admission to most programmes is 27 points on the Grade 12 average marks index, the Admission Points Score (APS). These points are arrived at on the basis of performance across subjects in the Grade 12 examinations. Applicants whose points range between 22 and 26 are required to take the PTESLAL and those who pass are granted access to the university while those who fail are not.

The research problem for this study is two-fold. Firstly, while performance in the English HL and FAL language examinations and the PTESLAL have been used for student admission at CUT, to the researcher's best knowledge no research has been done to determine the ability of these assessments to predict academic success at this institution. Secondly, studies investigating the predictive validity of TALL and the NBT AL have mainly been carried out at traditional academic universities but not at universities of technology in South Africa. Cohen and Swerdlik (2010: 172) have argued that, "no test or measurement technique is 'universally valid' for all time, for all uses, with all types of testtaker populations." It is therefore necessary

that the four tests selected for this study are researched for their predictive validity in different situations with different groups of test takers.

## **1.5 Aim of the study**

The aim of the study is to investigate the incremental validity of four English assessments of academic literacy, namely, the English HL and FAL examinations, the PTESLAL, TALL and the NBT AL. The key research question of the study therefore is the following: Do any of these tests possess incremental validity?

## **1.6 Hypothesis of the study**

TALL possesses incremental validity in relation to the other three tests, namely, the English HL/FAL language examinations, the PTESLAL and the NBT AL.

## **1.7 Chapter outline**

The remaining chapters of this study will be organized as follows:

### **1.7.1 Chapter 2**

This will be a review of the current literature on academic literacy testing. The chapter focuses on how the concept of validity has been defined and the scholarly debate around the meaning of this term. It argues for the hypothesized incremental validity of TALL on the basis that it is a well-researched test and that its empirical properties of reliability and validity are well established in the public domain as compared to those of the other three assessments. It also argues for the incremental validity of TALL in relation to the other three assessments because its impact, justice and fairness have also been established. It anticipates that the study may lead to some further insight into how we view validity and validation, a topic that I shall return to in the final chapter.

### **1.7.2 Chapter 3**

This chapter will outline the research methodology of the study. Firstly, the chapter points out that the procedure used for selecting the sample is convenience sampling. Secondly, the chapter discusses ethical considerations that were relevant for the study and argues that paying attention to such considerations was not necessary in the data collection process involving the Grade 12 English HL/FAL examinations, the NBT AL and the PTESLAL, because data from these assessments were already available at the time the study was undertaken. It also argues that it was necessary for the participants to be deceived about the purpose of their taking TALL so that the validity of this test could be protected. Thirdly, the chapter discusses the model used for determining academic success. Fourthly, it demonstrates that this study was mainly quantitative in nature and that descriptive and inferential statistics were used to analyse the data. Finally, the chapter gives a description of the specifications that derive from the constructs of the tests investigated in this study as well as the tasks used to achieve the measurement of the constructs underpinning them.

### **1.7.3 Chapter 4**

In this chapter, the descriptive and inferential statistics arising from the analysis of the data from the four assessments are presented and discussed.

#### **1.7.4 Chapter 5**

This chapter focuses on the interpretation and discussion of the results of this study. Firstly, the discussion focuses on the results of the study in relation to the predictive validity of Grade 12 English. Secondly, it deals with the results of the study with regard to the incremental validity of the PTESLAL. Thirdly, the discussion focuses on the results of the study in relation to the incremental validity of NBT AL and TALL. Finally, it deals with the results of the whole study in relation to recent studies on the predictive validity of NBT AL, TALL and Grade 12 results.

#### **1.7.5 Chapter 6**

This chapter starts by summarising this study. It then moves on to discuss its limitations as well as recommendations and suggestions for further study. Finally, the chapter provides a discussion of the low graduation rates at South African universities as a result of academic under preparedness and as the motivation for predictive validity studies of the kind carried out in this thesis.

#### **1.7.6 Chapter 7**

This chapter deals with the implications of the results of the present study for the theories of validity that are dealt with in Chapter Two. It starts with a brief exploration of these theories and then moves on to discuss the framework of language test design that has been proposed within the field of applied linguistics in response to one of the theories. Finally, the chapter focuses on a discussion of the implications of the analysis of the results of this study for the validity of academic language tests and courses.

## 1.8 Conclusion

The aim of this chapter has been to provide an introduction to the study as whole. It starts by explaining the aim of the study and names the variables involved. It then makes the point that levels of academic literacy among students at South African universities have been low in recent years and that this has given rise to the need for academic literacy testing and intervention. The chapter then argues for the need for the design and development of both academic literacy courses and the tests used for placing students in such courses to be informed by constructs of academic literacy that are reflective of the unique nature of academic discourse. It moves on to focus on the constructs of academic literacy underpinning the four assessments being studied and asks whether these constructs are compliant with the definitions of academic discourse and academic literacy. Finally, the chapter presents and explains the research problem and question as well as aim of the study. It ends with an outline of the remaining chapters of the study.

The next chapter is a review of the relevant literature on language testing. It discusses the debate around the meaning of the term validity and pursues a research-based argument in support of the hypothesis of this study that TALL possesses incremental validity in relation to the other three tests.

## **Chapter 2: Literature Review**

### **2.1 Introduction**

In this chapter, we focus on reviewing the language testing literature as it relates to the hypothesis of this study. In other words, the chapter examines the extent to which the literature supports the present study's hypothesis that the Test of Academic Literacy Levels (TALL) stands a chance to possess incremental validity with regard to first year academic performance in the chosen context of study as compared to the other three predictor assessments, namely the Grade 12 English examination, the Proficiency Test English Second Language Advanced Level (PTESLAL) and the National Benchmark Test in Academic Literacy (NBT AL). Prior to this and with no intention to delve extensively into the debates involved, however, the chapter begins by briefly exploring the various ways in which the concept of validity has been conceptualized and defined in both the fields of educational and psychological testing. It is appropriate that an exploration of this kind should precede a validity study such as this one.

### **2.2 Validity**

The concept of validity is probably the most crucial and contested of all principles governing the design and development of tests. In the words of Rambiritch (2012a: 62), "One would be forgiven for assuming that all questions find their answers in the concept of validity, for it is the concept of validity that seems to dominate the literature on language testing." Traditionally, validity has been used to refer to the question of whether a test measures what it is intended to measure. Traditionalists view validity "to be an inherent attribute or characteristic of a test, that a

psychologically real construct or attribute exists in the minds of the test taker – this implies that if something does not exist, it cannot be measured” (Van der Walt & Steyn 2007: 139). From the point of view of this definition, a test is valid if it measures what it purports to measure (Kelly 1927; Cattell 1946; Lado 1961). In other words, such a test restricts itself to “measuring only what it is intended to test and not extraneous or unintended abilities” (Weir 1993: 19). In this sense, validity is a property of nothing else but the test involved.

This definition has not gone unchallenged, though. For example, Messick (1989) has associated validity with how test scores are interpreted and used, and not necessarily with the test yielding such scores. His definition of the term captures this notion very well:

Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of the interpretation of the *inferences* and *actions* based on test scores or other modes of assessment. (Messick 1989: 13)

Messick’s view of validity has received support from a number of scholars in both educational and psychological measurement. For example, Lynch (2003: 146) points out that while people conveniently refer to the validity of a test, “it is important to remember that validity is a property of the conclusions, interpretations or inferences that we draw from the assessment instruments and procedures, not the procedures themselves.” Similarly, Bachman (2004: 259) has argued against the common tendency to attach validity to a test instead of associating it with how the scores yielded by such a test are interpreted and used. Chapelle and Brindley (2002: 270) maintain the same position:



Test users are always interested not in test performance and test scores themselves, but in what the scores mean, that is, the inferences that can be drawn from them and what they can do with the scores.

This perspective of validity implies that any interpretation of a set of scores cannot have validity for all times, situations and test takers (Cohen & Swerdlik 2010: 179). Indeed, Bachman (2004: 260) argues that every test should be developed bearing in mind the use for which it is intended, how its scores will be interpreted and the characteristics of the test takers for whom it is intended. It is for this reason that McNamara (2004) has argued that the interpretation of a test's scores be validated every time such a test is used with a new group of test takers, in a new context and for a different purpose. In other words, it is incumbent upon "test users to define precisely what information they wish to obtain from a test before they can decide whether or not it is valid" (Van Els, Bongaerts, Extra, Van Os & Janssen-van Dieten 1984: 318).

A dimension of Messick's (1989) framework, which introduced a new perspective to the way validity had been defined, was his further association of this concept with the consequences of how test scores are interpreted and used. Messick (1980: 1012) has contended that "not only should tests be evaluated in terms of their measurement properties, but that testing applications should be evaluated in terms of their potential social consequences". In support of this, McNamara and Roever (2006: xiv) have argued that language tests should not solely be validated psychometrically because "language is rooted in social life and nowhere is this more apparent than in the ways in which knowledge of language is assessed."

Considering these “extravalidity concerns” is, in Gregory’s (2007: 139) view, a test designer’s way of acknowledging that testing has consequences that are unrelated to a test’s psychometric soundness. Bachman and Palmer (1996: 30) have further advanced this view by arguing that

The very acts of administering and taking a test imply certain values and goals, and have consequences. Similarly, the uses we make of test scores imply values and goals and these uses have consequences.

The consequences Bachman and Palmer (1996) refer to above relate directly to the stakes that accompany a decision taken on the basis of a particular test’s scores (Miller, Linn & Gronlund 2009) and therefore determine the type of measurement instrument used and the quantity of the resources expended on the development of such a tool. The higher the stakes attached to a test, the more important it is that its consequences be taken into account in the assessment of its overall validity (Messick 1989). As the phrase implies, “high stakes” decisions include those that may have a negative impact on a large number of people (Bachman & Palmer 1996: 96-97). In most cases, such decisions cannot be rescinded and can therefore have a lifetime negative impact on the lives of those involved (Bachman & Palmer 1996: 97). It is important therefore that high stakes decisions taken on the basis of test scores be taken wisely and in harmony with the purpose for which a test was designed (Stoynoff and Chapelle 2005: 165).

While Messick’s (1989) consequential dimension of validity has not generated any opposition from scholars in the fields of educational and psychological assessment, his inclination to associate validity solely with test scores, and not the test through

which such scores are generated, has been challenged by language testing scholars especially. For example, Davies and Elder (2005: 279) have argued that

... through acquiring over time, and through repeated validation arguments, an adequate reputation, any test must eventually present a principled choice to those wishing to use it, and that choice can be attributed to nothing else than its known validity.

Borsboom, Mellenbergh and Van Heerden (2004: 279) have similarly argued that a test that is used many times for a similar purpose meets the psychometric requirement of validity if no evidence exists to show that it is used for purposes it was not designed for. In the words of Borsboom et al. (2004: 279), it should be possible to “speak of the validity of that particular test – as a characteristic of it”. Furthermore, Weideman (2012) has challenged Messick’s (1989) insistence on associating validity with test scores and not the measurement instrument itself by arguing that this drives attention away from the importance of the psychometric soundness of such an instrument. Weideman (2012) rightly points out that no matter how good the interpretation of a set of test scores is, if the measurement instrument is not technically sound, this interpretation is of no utility to the test user. In Weideman’s (2012: 4) words, “No amount of interpretation can improve the measurement result (score) obtained from an inadequate instrument that gives a faulty and untrustworthy reading.” In the light of this, Weideman (2012: 6) has argued for the need for one to distinguish between the objective effect of a test and the subjective interpretation of its scores. Weideman (2012) has further argued that through his use of the word ‘adequacy’ in his definition of validity, Messick inadvertently attaches validity to the measurement instrument and not the interpretation of the scores from such an instrument as he claims he does.

Adequacy is, in Weideman's (2012) thinking, a word that is conceptually appropriate to describe a test and not the interpretation of its scores. As Weideman (2012: 6) puts it, "... using validity as descriptive of a test therefore merely returns in another guise, that of adequacy ...". In other words, while he does not deny that scores require subjective interpretation, or that their use has a social impact, Weideman (2012) believes that Messick's definition of validity simply constitutes a circumlocution aimed at obfuscating the traditional definition of validity as an objective property of a test. To put Weideman's (2012) point differently, one would surely not consider using an invalid, inadequate instrument or testing object to measure language ability responsibly.

Traditionally, validity has been categorized into three types. These are the content, construct and criterion-related types. These concepts have also been viewed differently by scholars in the fields of educational and psychological measurement. In the following section, we briefly look at how each of the concepts was traditionally defined and the current debate on what they mean. All three may still figure and contribute to what may be called the validation process, i.e. the argument-based procedure (Kane 1992) that has now replaced less sophisticated and less complex notions of validity. I shall return below to an application of the notion that the subjective validation of a test is a process that is distinguishable from its objective validity (Van der Walt & Steyn 2007; Weideman 2012). The traditional categories of validity are articulated here, however, since there is agreement that they may provide evidence for the validation of a particular use of a test (Stoynoff & Chapelle 2005).

### 2.2.1 Content validity

Content validity is a term traditionally used to refer to the degree to which tasks in a test are adequately representative of the universe of the content that allows the test designer to capture enough of the construct or knowledge that they wish to measure (Cohen & Swerdlik 2010). Validating a test's content involves a scientific examination of the degree to which its items represent the targeted content domain and the use of content experts to rate the extent to which this is the case (Kurpius & Standford 2006: 147). Content validation is in this sense essentially a content sampling exercise that needs to be carried out with care if any claim is to be made that a test possesses content validity. In the words of Miller et al. (2009: 75),

The essence of content consideration in validation, then, is determining the adequacy of the sampling of the content that the assessment results are interpreted to represent. More formally, the goal in the consideration of content validation is to determine the sample of the domain tasks about which interpretations of assessment results are made.

In language testing, this sampling exercise involves a consideration of the characteristics of the language tasks typical of what Bachman and Palmer (1996) call the Target Language Use (TLU) domain. The term TLU refers to the particular real life situation in which the test taker will use language. This means that for the purpose of ensuring content validity, language test designers are obliged to ensure that the characteristics of their test tasks are a mirror of those typical of the tasks inherent to a TLU domain. Bachman and Palmer (1996) have referred to this correspondence between test tasks and the specified TLU domain as authenticity. In the words of Bachman and Palmer (1996: 23), authenticity is “the degree of

correspondence of the characteristics of a given language test task to the features of the TLU task”.

### **2.2.2 Construct validity**

Construct validity is probably the most important of all traditional classifications of the concept of validity. This is a term used to refer to the degree to which a theory underpinning the ability a test was designed to measure can be justified. In the words of Stoyhoff and Chapelle (2005: 17), construct validity relates to the “extent to which evidence suggests that the test measures the construct it is intended to measure, in other words, that inference specified as one facet of test purpose is justified”. This means therefore that testers “need to be precise about what a test is intended to measure” and should “develop the conceptual apparatus to do so” (Chapelle & Brindley 2002: 269). In other words, a construct first has to be defined and evidence subsequently produced to demonstrate that a test measures the ability it purports to measure. In language testing, Bachman and Palmer’s (1996) notion of TLU is crucial, once again, to both the definition and validation of a construct. Validating a language test’s construct essentially involves ensuring that the test’s tasks are aligned to the targeted TLU domain. Thus, not only is authenticity a function of content validity, it is inherent to construct validity as well.

### **2.2.3 Criterion-related validity**

The type of validity that is of particular interest to the present study is criterion-related. Criterion-related validity refers to the judgment of the degree to which a test is equivalent to another measure, also known as a criterion, of the same or

related ability or knowledge. A criterion is therefore another measurement requirement used as a standard against which the accuracy and appropriateness of another similar or related assessment tool is evaluated. Concurrent and predictive validity are two types of validity that are subsumed under criterion-related validity. On the one hand, concurrent validity is an estimation of the degree to which test scores correlate with those obtained in an equivalent measure or criterion that is administered around the same time. On the other hand, predictive validity refers to the extent to which test scores can predict performance on another measure or criterion that will be administered at a later stage.

Messick (1980) has argued against this traditional categorization of validity into the three types dealt with so far. Instead, he views validity as a single unifying concept that does not need to be compartmentalized in this manner. In Messick's (1980: 1014) thinking, there is a problem with this classification:

Many test users focus on one or another of the types of validity as though any one would do, rather than on the specific inferences they intend to make from the scores. There is an implication that once evidence of one type of validity is forthcoming, one is relieved of the responsibility for further enquiry.

In Messick's framework, construct validity is the umbrella concept while the traditional categories of content and criterion-related validity are sources of evidence for this unitary conception of the notion of validity (Stoynoff & Chapelle 2005). A construct validation study would in Messick's view involve an "overall evaluative judgment" (Bachman 2004: 260) that requires that, in the validation process, all available evidence be advanced to support the appropriateness, meaningfulness and usefulness of the interpretation of test scores. Some of such evidence includes "a consideration of the content measured, the ways in which

students respond, the relationship of individual items to the test scores, the relationship of performance to other assessments, and the consequences of using and interpreting assessment results” (Miller et al. 2009: 73).

Weideman (2012) has contested Messick’s unitary approach to validity by arguing that it is a conflation of what he terms the regulative and constitutive concepts of responsible test design. Weideman’s (2009: 1) constitutive requirements include systematicity, reliability, the three traditional types of validity, and the meaningfulness of test results while the regulative conditions are constituted by accessibility, transparency and accountability. Weideman (2009, 2012) has argued against subsuming all these conditions under a single concept such as construct validity as Messick (1980, 1989) does. He believes that conceptual clarity is achievable only if each of the constitutive and regulative conditions is recognizable as a critical factor in responsible test design and appraisal. In the words of Rambiritch (2012a: 60), Weideman’s (2009, 2012) framework underlines a number of “important concepts in language testing” and “allows for a more open and flexible way of designing and using tests rather than the restriction of an overarching or unified ... concept” such as Messick’s (1980, 1989) construct validity or Bachman and Palmer’s (1996) test usefulness. In the words of Weideman (2009: 249) himself, the value of his framework “lies in separating out what is conceptually distinct, and, by so doing, enriching our theoretical understanding of the constitutive and regulative, necessary and sufficient conditions of language testing”. Finally, Weideman (2012) observes that efforts by scholars such as Kane (1992), Bachman and Palmer (1996) and Kunnan (2000) to



reinterpret Messick's unitary concept of validity are evidence of a conceptual disunity, so that the need to distinguish between constitutive and regulative conditions of test design is underlined. Weideman (2012: 8) adds that these attempts at reinterpreting Messick's framework do not achieve conceptual clarity and that instead, "they may help more to confuse" us. I will return to a more detailed discussion of Weideman's argument for the constitutive and regulative conditions for test design later in Chapter Seven of this study.

### **2.3. Hypothesis of the study**

As pointed out in Chapter One, the hypothesis of the present study is that TALL possesses better incremental validity when compared to Grade 12 English, the PTESLAL and the NBT AL. This hypothesis is mainly based on the degree to which TALL, unlike the other three assessments, has been researched and validated in refereed journals. Validation is, according to Van der Walt and Steyn (2007: 142; 2008: 194), a process of operationalizing validity and which mainly involves assembling and presenting empirical evidence from multiple sources in support of a test's validity. Kane (1992: 527) has referred to this process as the building of "an interpretive argument". The evidence collected for validating a test may include construct validity, content validity, criterion-related validity and reliability coefficients, test-taker feedback, test consequences, test ethics, social responsibility, washback and impact (Van der Walt & Steyn 2007: 142; 2008: 194). In line with Van der Walt and Steyn's (2007, 2008) approach to test validation, and in support of the hypothesis of the current study, I will now examine a set of claims about the psychometric and consequential properties of TALL in the current

literature on language testing and relate these claims to current opinion about test quality.

### **2.3.1 TALL is a reliable test of academic literacy**

The first of the empirically established properties of TALL referred to earlier is its reliability. Reliability is a concept that refers to consistency in measurement. Reliability has, like validity, traditionally been viewed as a property of the measurement instrument. From this point of view, a test is reliable if it yields consistent performance when administered to the same group of test takers more than once. Also, a test is reliable if its alternate versions yield consistent results from the same group of test takers. In the words of Du Plessis (2012: 31), “test scores may be deemed to be reliable if they remain consistent from one set of tests and tasks to another. Reliability is thus a function of score consistency between different administrations of tests and tasks.” Messick (1989), however, does not view reliability in the same manner. In his view, reliability is a property of the interpretation of test scores and not necessarily the test itself. Also, Messick does not view reliability as a technical aspect of test design in its own right. Rather, he regards it as a source of evidence for his unitary concept of validity. As pointed out earlier, however, Weideman (2009, 2012) has contested Messick’s failure to distinguish between what he calls constitutive aspects of test design such as validity and reliability from those he refers to as regulative conditions.

The approach to reliability in the design and development of TALL has been the one proposed by Weideman above. In other words, in the theoretical

conceptualization of this test, reliability is viewed as a constitutive condition of test design in its own right and – though it may contribute to or support it - not necessarily only as a source of evidence for validity. Studies on the reliability of TALL have consistently reported high reliability indices for the test. For example, the reliability coefficient of TAG, the Afrikaans version of the test, was found to be .86 and .87 while the reliability indices of TALL, the English version of the same test, was .96 and .92 after it was administered to 10 000 students at the Universities of Pretoria and North West (Weideman 2003).

Furthermore, a total of 6,310 first year undergraduate students at the University of Pretoria were administered TALL while 3,277 of them took TAG for the purpose of assessing their levels of academic literacy at the beginning of 2004 (Van de Slik & Weideman 2005). SPSS (Statistical Package for Social Sciences) and TiaPlus are the two statistical packages that were used to determine the psychometric efficiency of the test on this occasion. The software packages were used to compute Cronbach's alpha and Greatest Lower Bound (GLB) to determine the reliability of the two versions of TALL at test level (Van der Slik & Weideman 2005). Cronbach's alpha is a statistic that is used under Classical Test Theory to determine the amount of error in test scores, also known as internal consistency. Unlike Cronbach's alpha, GLB does not assume a unidimensional construct and may be higher than the former if used to determine the reliability of a test with a heterogeneous construct (Van der Slik & Weideman 2005). The 2004 administration of TALL at the University of Pretoria yielded a Cronbach's alpha

statistic of .92 and .83 for TALL and TAG respectively, and a GLB of .95 and .90 for the two versions of the test respectively (Van der Slik & Weideman 2005).

In a study investigating the consistency of TALL and TAG over several administrations and over a period of four years (2005-2008) at the Universities of Pretoria, Stellenbosch and North West, Van der Slik and Weideman (2009) further confirmed the high reliability indices of the test. Computing Cronbach's alpha and Greatest Lower Bound (GLB) after these administrations, reliability indices ranging from .86 to .94 were recorded for TALL while TAG scored between .81 and .94 on the two statistics (Van der Slik & Weideman 2009; Van der Slik, 2008). Actually, TALL and TAG have yielded an average Cronbach's alpha reading of .90 and .85 at the three universities over the years (Van der Slik & Weideman 2009). Selected psychometric properties of TALL for these administrations are captured in **Table 3** below:

**Table 3: Selected properties of the academic literacy test (2005-2008) (standard deviations in italics)**

TALL	UP	US	NWU	Overall
N	15,202	13,886	675	29,793
Mean proportion correct ( <i>difficulty</i> )	.65 (0.05)	.69(0.05)	.49(0.13)	.61(0.12)
Mean Cronbach's alpha ( <i>reliability</i> )	.92 (0.01)	.88(0.01)	.91(0.03)	.90(0.02)
Mean Average Rit ( <i>discrimination index</i> )	.45(0.01)	.38(0.01)	.45(0.02)	.43(0.04)

(Van der Slik & Weideman 2009: 257)

Furthermore, in a study aimed at investigating whether TALL was robust enough to have reliability and validity in contexts other than South African universities, Le,

du Plessis and Weideman (2011) investigated these properties of the test after it was administered to a total of 197 students in the College of Foreign Languages (CFL) at the University of Da Nang in Vietnam as compared to those after its administration to a total of 1819 first year students at the University of Pretoria. The Cronbach's alpha measure for the CFL students was found to be 0.774 and 0.831 for the test takers from the University of Pretoria. These measures, in the view of Le et al. (2011), are acceptable because TALL is a low-stakes test used for placement and not for making high-stakes decisions like admission. The combined GLB measure for the two groups of test takers on this occasion was recorded at .91. Clearly, from the point of view of reliability, of all the four assessments to be investigated in this study, TALL is the only one whose Cronbach's alpha and GLB readings have been researched and are known to have been consistently high across various administrations. In the words of Le et al. (2011: 118), this test

has consistently measured test scores under different testing conditions and thus meets the quality of reliability .... This implies that test scores remain consistent from one set of tests and tasks to another.

To this end, TALL has demonstrated evidence of what Whiston (2013) refers to as reliability generalization. As Whiston (2013: 56) explains, "reliability generalization involves combining reliability estimates across studies, which allows researchers to characterize and explore variance in score reliability". This is the first basis on which it is hypothesized in this study that this test will score higher on incremental validity than the Grade 12 English examination, PTESLAL and NBT AL. As will again be noted in Chapter Five of this study, Du Plessis (2016: 7) has argued that "to date, no use has been made of statistical methods to determine the scoring validity of the Grade 12 examination papers, and no subtest correlation data

are available either”. Similarly, while “a significant amount of research has been done to establish the reliability and validity of the NBTs” and the Cronbach’s alpha of the NBT AL is reported to be .88 (NBTP 2013: 10-11), these findings have, as far as the researcher knows, never been published in refereed journals. Furthermore, in a study focusing on how the inferences made from performance on the NBT AL might translate into teaching and learning, Cliff (2015) fleetingly reports acceptably high reliability statistics of seven NBT AL test forms. On the basis of this, Cliff (2015: 12) observes that the high Cronbach’s alpha coefficient for these tests support his argument “for the coherence of the test and the reliability of test scores”. To date, this is the only study, however, through which psychometric information of this kind has been made available to the scholarly public. Finally, the reliability of the PTESLAL has been reported to be .89 (HRSC 1991: 19). To the researcher’s best knowledge, however, there is no peer-reviewed study available in the public domain on the empirical properties of this test.

### **2.3.2 TALL has construct validity**

A critical factor underpinning a hypothesis of this kind of study is the strength of the evidence assembled and presented to demonstrate a test’s construct validity. While information regarding the validity of the Grade 12 English examination is not existent and very little is publicly available regarding the construct validity of the PTESLAL and the NBT AL, like its reliability, TALL has been the most researched of the four tests in this respect as well (cf. Patterson & Weideman 2013a, 2013b). The following section covers various further sources of research generated evidence that have attested to the construct validity of this test.

### 2.3.3 TALL possesses acceptable item difficulty and discrimination levels

At a very basic level, construct validity is a function of a test's item difficulty, also known as *p*-value as well as its discrimination power, also known as *d*-value. A test that is too easy or too difficult for a specified group of students cannot provide meaningful information about the levels of the ability being tested and fails to meet a construct validity criterion. Indeed, Chapelle and Brindley (2002: 277) have argued, as we further observe in Chapters Six and Seven of this study, that when test "difficulty is interpreted in view of the construct that an item of a test is intended to measure, it can be used as one part of a validity argument". The difficulty index of an item is synonymous with its mean score. The higher the mean score or difficulty index of an item, the easier the item and vice versa. For a criterion-referenced test with a purpose similar to that of TALL, item *p*-values of .20 to .75 or .20 to .80 are recommended (Bachman 2004).

Item difficulty relates very closely to item discrimination. Item discrimination refers to a statistical description of how well an item separates high from low performers in a test. In other words, item discrimination "shows the relationship between examinees' performance on a single item and their performance on the test as a whole" (Stoynoff & Chapelle 2005: 20). Item discrimination indices range from -1 to +1. An item that most high performers in a test as a whole get right and which most weak performers get wrong discriminates positively and therefore has the desired discrimination power. In other words, "a good item is one that low-ability test takers tend to answer incorrectly and that the high-ability test takers answer correctly" (Stoynoff & Chapelle 2005: 20). Conversely, an item which

most poor performers in the test as a whole answer correctly and most high performers answer incorrectly discriminates negatively and has an undesired discrimination index. This kind of item falls short of meeting the psychometric criteria of both reliability and construct validity. As Cohen and Swerdlik's (2010: 258) explain,

... the higher the value of  $d$ , the greater the number of high scorers answering the item correctly. A negative  $d$  value on a particular item is a red flag because it indicates that low-scoring examinees are more likely to answer the item correctly than high scoring examinees.

Finally, an item that all test takers answer incorrectly or correctly has a discrimination index of zero. Such an item is psychometrically just as defective as one that has a low or negative discrimination index. The item fails to provide differential information about test taker ability and therefore impacts the validity of a test negatively. For criterion-referenced tests such as those investigated in this study, discrimination indices of .30 and higher are recommended (Bachman 2004).

While  $p$ -values of .50 are generally desirable for the four assessments chosen for the present study, the developers of TALL and TAG would like their  $p$ -values to range between .20 and .80 and  $d$ -values to be .30 and more (Weideman & Van der Slik 2005; Weideman 2011). After three administrations of the two versions of the test to students at the Universities of Pretoria, Stellenbosch and North West, Van der Slik and Weideman (2009) recorded average  $p$ -values of .61 and .57 and average  $d$ -values of .43 and .36 for TALL and TAG respectively. These statistics are recorded in **Table 3** above. The item difficulty and discrimination statistics of TALL are therefore compliant with those recommended by measurement



specialists. Furthermore, in a study aimed at comparing performance by students at the Universities of Pretoria and Da Nang in Vietnam on TALL, Le et al. (2011) recorded an average  $p$ -value of 53.73 and an average Rit-value of .24 for this test. Rit is an item-total correlation that can also be used as a discrimination measure. These values indicate that the test was of acceptable difficulty for the two groups of students and that the correlation between performance on each item and the test as a whole was satisfactory (Le et al. 2011). From the point of view of item difficulty and discrimination, this constitutes evidence for the test's construct validity.

For the purpose of validating the construct underpinning TALL and TAG further from the point of view of item difficulty in particular, some researchers investigating the empirical properties of TALL have used what is known as Item Response Theory (IRT). While Classical Test Theory focuses solely on gauging the psychometric properties of test items, IRT models focus on assessing such qualities in relation to the test taker's level of targeted ability. In other words, from the point of view of IRT, the facility values and discrimination indices of test items, as well as the ability level of those who are tested are critical factors in the validation of a test's construct. The IRT model that has been used to validate the construct of TALL is known as the Rasch analysis (Van der Walt & Steyn 2008). Firstly, a program known as FACETS is used to run a Rasch analysis to help the test developer identify items that are too easy or too difficult for the targeted group of test takers (Bachman 2004). FACETS makes an estimation of item difficulty on a logit scale, which represents the ability levels of the targeted test takers and whose standardized mean is zero and the values of which range on a continuum

from -3 to +3 (Bachman 2004: 147). High positive values on this scale indicate high item difficulty levels while negative values are indicative of low item difficulty indices for test takers at different levels of the ability measured by a test.

Furthermore, the Rasch analysis enables the test designer to determine the extent to which the particular IRT model used fits test data and thereby to compute what is known as “fit statistics” (Van der Walt & Steyn 2007, 2008). Fit statistics are established by computing what is technically known as “infit mean square”. The expected infit mean square value is 1. McNamara (1996: 173) has suggested that acceptable infit mean square values are those ranging from 0.75 to 1.3 and that those above 1.3 are indicative of significant misfit while those below 0.75 indicate significant overfit. Using the FACETS program to run a Rasch analysis, Van der Walt and Steyn (2007, 2008) found that the items in both TALL and TAG were of appropriate difficulty for the targeted group of test takers and that their infit mean square values ranged from 0.97 to 1.04. This means that both versions of TALL have shown appropriate fit for the test takers for which the test was designed. This known outcome of an IRT analysis regarding TALL items is not publicly known about the Grade 12 English examination, PTESAL and NBT AL. This was part of the motivation for the researcher to hypothesize that TALL will, for the purpose of and in the context of this study, possess better incremental validity when compared to the other three predictor assessments.

#### **2.3.4 Factor analysis has attested to the construct validity of TALL**

A standard procedure for validating a test’s construct is factor analysis. In the words of Stoyloff and Chapelle (2005: 21), factor analysis functions “to reduce a

large number of variables (e.g test or questionnaire items) to a smaller number (thought to represent the underlying abilities the test developer is seeking to measure) of variables”. The procedure is premised on the understanding that a construct consists of traits, some of which intercorrelate and can therefore be reduced into a single factor or dimension. A discovery of such factors is a suggestion to the test developer that the construct aimed at exists and that it can be measured. The ultimate purpose of factor analysis, however, is to determine whether these factors are all related to one construct and that such a construct is therefore homogeneous or unidimensional. In other words, a factor analysis of test data should produce evidence that a test’s construct constitutes a homogeneous ability.

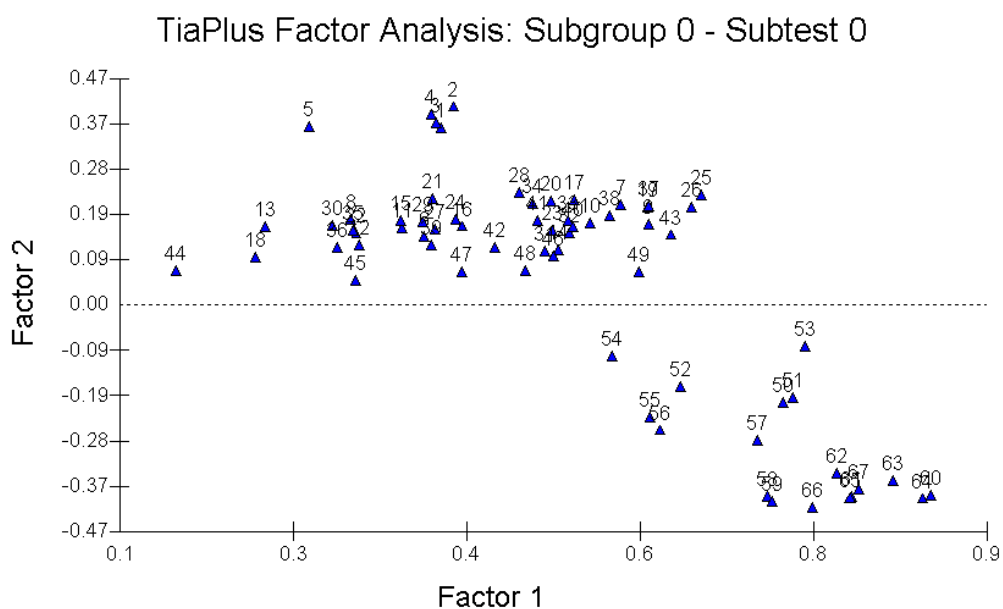
To the researcher’s best knowledge, no peer reviewed studies are available in the public domain about whether the constructs underpinning both the Grade 12 English examination and PTESAL are homogenous or not.

The NBT AL is different from these two assessments in this regard. In a study focusing on how the diagnostic information yielded by the NBT AL might translate into curriculum development in higher education, Cliff (2015: 12) refers to a factor analysis of performance on this test and on the basis of which he concludes that the test’s construct is “highly unidimensional, with essentially one factor with an eigenvalue greater than 1”. In the observation of Cliff (2015: 12), “these data appear to support the internal coherence of the test, but also suggest the presence of more than one factor – the classification of the test construct into a number of sub-constructs appear somewhat justified by the factor analytic structure...” So far, this

is the only study which provides evidence of the construct validity of the NBT AL that is available to the scholarly public.

TALL is slightly ahead of the NBT AL in this regard. Several studies (e.g. Van der Slik & Weideman 2005; Van der Walt & Steyn 2007, 2008; Le et al. 2011) have been conducted and published in which a factor analysis of this test and its Afrikaans counterpart was carried out. These studies have revealed - against the traditional expectation that a test's construct should be homogeneous – that TALL and TAG's construct is multidimensional. This finding has been justified on the grounds that in the case of “an ability as richly varied and potentially complex as academic language ability, one would expect, and therefore have to tolerate, a more heterogeneous construct” (Weideman 2009: 5). The heterogeneous nature of the construct of TALL is demonstrated in its factor analysis output presented in **Figure 1** below. The fact that some items of this test are further away from the zero line bears testimony to the heterogeneous nature of the test's construct:

**Figure 1: Measures of homogeneity and heterogeneity in TALL 2008**



The multidimensional nature of the construct underpinning TALL resides in the fact that the test's construct mainly derives from Bachman and Palmer's (1996: 61-62) model of language ability. Bachman and Palmer (1996: 63) view language ability as being constituted by a complex interaction of the language user's topical knowledge, language knowledge, personal characteristics and the characteristics of the language use situation. As Bachman and Palmer (1996: 61-62) rightly point out, in real-life language use, the interaction of these factors translates into

the creation or interpretation of intended meanings in discourse by an individual, or ... the dynamic and interactive negotiation of intended meanings between two or more individuals in a particular situation.

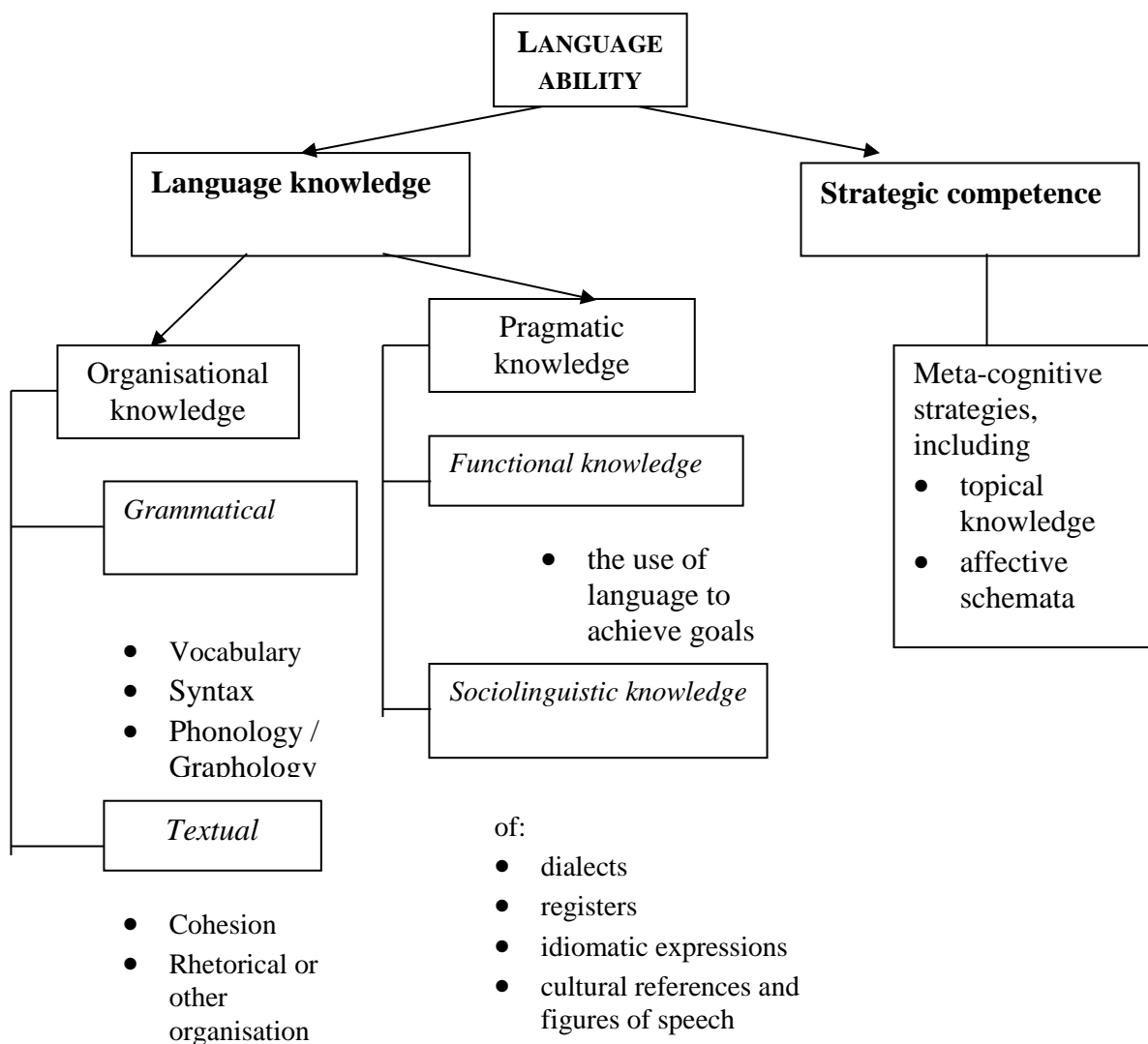
On the basis of their framework of language ability, Bachman and Palmer (1996) have contended that tests of language ability be designed to elicit language use performance that entails the interaction of all the constituent processes of language ability mentioned earlier. In other words, the construct validation of a test developed according to the Bachman and Palmer (1996) framework should entail a consideration of the degree to which such a test is interactive. In the words of Bachman and Palmer (1996: 25):

The interactiveness of a given language test task can thus be characterized in terms of the ways in which the test taker's areas of language knowledge, metacognitive strategies, topical knowledge, and affective schemata are engaged by the test task.

The interaction of the factors that Bachman and Palmer (1996) have identified to constitute language ability are the essence of the reading, writing and thinking abilities that students typically need to succeed at university and are the basis on

which TALL is designed and developed. The Bachman and Palmer (1996) model is captured in **Figure 2** below:

**Figure 2: The Bachman and Palmer construct of language ability**



(Van Dyk & Weideman 2004: 143)

The Bachman and Palmer (1996) model captures what Weideman (2003, 2004) refers to as an ‘open’ as opposed to a ‘restrictive’ view of language ability. In the words of Weideman (2004: 139), a restrictive view of language “limits it to a combination of sound, form, and meaning, or, in technical linguistic terms, phonological, morphological, syntactic and semantic elements” while an open view “maintains that language is not only expressive, but communicative, intended to

mediate and negotiate human interaction”. These two opposite views of language ability are captured in **Table 4** below:

**Table 4: Two perspectives on language ability**

Restrictive	Open
Language is composed of elements: <ul style="list-style-type: none"> <li>• Sound</li> <li>• Form, grammar</li> <li>• meaning</li> </ul>	Language is a social instrument to: <ul style="list-style-type: none"> <li>• mediate</li> <li>• negotiate human interaction</li> <li>• in specific contexts</li> </ul>
<i>Main function</i> : expression	<i>Main function</i> : communication
<i>Language learning</i> = mastery of structure	<i>Language learning</i> = becoming competent in communication
<i>Focus</i> : language	<i>Focus</i> : process of using language

(Van Dyk & Weideman 2004: 5)

From the point of view of writing, Blanton (1994: 228) has, in like fashion, further argued that the kind of language-related skills that second language students need to function at university involves the ability to interact with texts:

Whatever else we do with L2 students to prepare them for the academic mainstream, we must foster the behaviour of ‘talking’ to texts, talking and writing about them, linking them to other texts, connecting them to their own lives and experiences, and then using their experiences to illuminate the text and the text to illuminate their experiences.

This is also the kind of language-related competence that students need to possess in order to succeed at university and which is naturally constituted by a complex interaction of their topical knowledge, language knowledge, personal characteristics and the characteristics of the language use situation.

From the perspective of academic discourse being the interaction between one or more lingual subjects, we should acknowledge that the assessment of the ability to handle it may, like real-life interaction in texts, also be remote or asynchronous. In fact, testing this particular ability to communicate is certainly not testing oral, reading or writing proficiency – all of which have in the literature been reduced to single traits. As Weideman (2013: 9) remarks,

If CLT (Communicative Language Teaching) is not limited to speaking, ..., then surely testing communicative language ability cannot be restricted to testing oral proficiency either. Communication implies interaction amongst two or more individuals, and this interaction may even be displaced (non-simultaneous) and remote, depending on the communicative medium. In fact, interaction with academic texts is what is most commonly and justifiably thought to constitute the appropriate source that provides material for tests of academic literacy.

This justifies the multidimensional nature of TALL's construct and should logically make it the most likely to possess incremental validity when compared to the other three assessments.

### **2.3.5 Decision Theory has been used to identify misclassification in TALL**

Another way to validate a test's construct involves making use of what is, in the language of psychometrics, known as Decision Theory. Applying this theory to the design and development of tests involves using established procedures to determine the accuracy of the decisions taken on the basis of test scores. In admission and placement tests such as those investigated in this study, the test developer is enabled to determine the frequency with which these tests accurately classify the test takers, also known as *hits*, and how often they classify such test takers inaccurately, also known as *misses* (Whiston 2013; Erford 2013). In other words,



through Decision Theory procedures, the test developer is able to identify what are known as false positives and false negatives. “A false positive occurs when the instrument predicts that individuals have ‘it’ (the criterion) when in fact they do not” while “a false negative occurs when the instrument predicts that the test takers do not have it when in fact they do” (Whiston 2013: 68). In essence, false positives and false negatives are misclassifications that are inherent to testing. No test is 100% reliable. It is necessary therefore that test designers find ways to handle such misclassifications if any claim is to be made about the validity of their tests (Van der Slik & Weideman 2005).

The analysis of the scores from TALL has involved the use of the TiaPlus software package to identify these misclassifications. The software has enabled the developers of the test to use two types of scenarios that are derived from Cronbach’s alpha and Greatest Lower Bound (GLB) statistics to identify false positives and negatives (Weideman 2011). These scenarios are the correlation between TALL and a hypothetical parallel test as well as the correlation between observed and ‘true’ scores (Van der Slik & Weideman 2005, 2009; Weideman 2011). Based on the results of this analysis, false negatives in particular are given a second chance to demonstrate their academic literacy levels, and parameters are set for determining the size of such false negatives (Van der Slik & Weideman 2009). This is the extent to which the developers and users of TALL attempt to ensure that the test is fair to all those who take it. In the words of Cohen and Swerdlik (2010: 203), test fairness is “the extent to which a test is used in an impartial, just, and equitable way”. Contrary to their (Cohen & Swerdlik 2010) position that test

fairness and validity be treated as separate issues, however, a test whose results are unfairly used especially for those who are misclassified as not having the relevant criterion falls short of meeting the regulative criterion of fairness (cf. Kunnan 2000; Weideman 2009). Logically, this impacts such a test's regulative condition of fairness and by extension, its constitutive condition of consequential validity negatively. The statistics of the potential misclassifications of the writers of TALL at the Universities of Pretoria, Stellenbosch and North West from 2005 to 2008 are shown in **Table 5** below:

**Table 5: Potential misclassifications on the English version of the academic literacy test**

(Percentage of this tests population). In italics the corresponding interval (in terms of standard deviations) around the cut-off points.

TALL	UP	US	NWU
Alpha based: Correlations between test and hypothetical parallel test			
2005	432(13.0%) 63-74 (.31)	246 (14.2) 63 -74 (.41)	16 (11.8%) 64 – 71 (.18)
2006	439 (12.0%) 51 – 59 (.25)	432 (11.7%) 52-58 (.25)	20 (13.7%) 45 – 54 (.26)
2007	448 (11.5%) 47 – 55 (.19)	604 (14.5%) 54 – 61 (.24)	18 (12.8%) 43 – 52 (.19)
2008	179 (4.1%) 30 – 35 (.15)	152 (3.6%) 34 – 42 (.24)	26 (10.0%) 37 – 43 (.15)
Average % (Average sd)	(10.0%) (.23)	(11.0%) (.28)	(12.0%) (.20)

(Van der Slik & Weideman 2009: 258)

As can be seen from the last row in **Table 5** above, in TALL, false negatives have generally been found to “occur more or less within the expected range of scoring points around the cut-off point, i.e. around 0.25 standard deviations around the cut-off point” (Van der Slik & Weideman 2009: 258). This adds to the already existing

evidence for the construct validity of TALL and constitutes another basis for the hypothesis of this study.

### **2.3.6 TALL items function similarly for all test takers**

While some scholars (e.g. Cohen & Swerdlik 2010) have argued that validity and test bias should be treated as separate issues in test design and appraisal, ensuring that a test is free from bias has, in recent years, become closely associated with the validity of a test and the appropriateness of the interpretation and use of test scores obtained on the test. From the point of view of psychometrics, test bias “is a factor inherent in a test that systematically prevents accurate and impartial measurement” (Cohen & Swerdlik 2010: 199). Test bias is therefore a consistent and systematic failure by a test to provide a reliable and justifiable measurement of an ability a test was designed to measure, as a result of some factor that is a function of the background of the test takers involved and that is unrelated to the construct underpinning the test. In other words, a test is biased in favour of test-takers of a common background such as males if it discriminates against another group of test-takers such as females. Such a test would be male-oriented in some way and would make it more difficult for its user to make meaningful inferences about the female students involved. The fairness of such a test would be under question, since its construct would be giving a measure unrelated to what it can validly test. In the words of Jensen (1980: 444), the essence of a test that has validity from the point of view of bias is that

... any person showing the same ability as measured by the whole test should have the same probability of passing any given item that measures that ability, regardless of the person’s race, social class, sex, or any other

background characteristics. In other words, the same proportion of persons from each group should pass any given item of the test, provided that the persons all earned the same total score on the test.

In the interest of validating the construct of TALL from the point of view of test bias, Van der Slik (2008) conducted a study to establish if there was any evidence of gender bias in the TALL and TAG tests administered to undergraduate students at the Universities of Pretoria, Potchefstroom and Stellenbosch from 2005 to 2008. He used the TiaPlus software program to run T-tests and Differential Item Functioning (DIF) analyses to determine if male and female students performed differently on the two versions of the test. Furthermore, Van der Slik (2008) also used the StatsDirect package to perform meta-analyses on this test to determine the effect size of the difference in performance by males and females on the test throughout the four years. The general finding was that the two versions of the test did not exhibit evidence of significant differences of performance by males and females both at subtest and whole test levels. The conclusion Van der Slik (2008) made was that the negligible DIF evident at both these levels of the test was probably attributable to the difference between male and female cognitive functioning. In other words, Van der Slik (2008) concluded that the DIF he found was a probable result of gender differences that are related to cognition and not necessarily gender-related bias in the content of the test.

For the purpose of determining the construct validity of TALL further from the point of view of test bias, Van der Slik and Weideman (2010) conducted a study to investigate if the test would function differently for students from three first language backgrounds namely, African languages, English and Afrikaans at the

Universities of Pretoria, Stellenbosch and North West. T-tests were used and Differential Item Functioning (DIF) analyses were carried out by means of the Mantel-Haenszel statistic to determine this. The outcomes of the T-tests and DIF analyses of performance by the three groups of students are shown in **Table 6** below:

**Table 6: T-values of differences between mean scores on TALL of first year students who have an African language, English, or Afrikaans as their first language**

Study	1 versus 2			1 versus 3			2 versus 3		
	<i>T</i>	<i>DF</i>	<i>p</i> <sup>[1]</sup>	<i>T</i>	<i>DF</i>	<i>p</i> <sup>[1]</sup>	<i>T</i>	<i>DF</i>	<i>p</i> <sup>[1]</sup>
2005	39.62	2462	< .001	26.13	1521	< .001	.34	1887	> .05
2006	39.83	2675	< .001	28.31	1713	< .001	-.37	1994	> .05
2007	37.39	3179	< .001	27.72	1467	< .001	-3.12	2540	< .01
2008	35.23	3505	< .001	27.60	1625	< .001	-1.87	2935	> .05

[1]: with Bonferroni adjustment

(Van der Slik & Weideman 2011: 111)

The T-tests and DIF statistics in **Table 6** above show that there were negligible differences in performance between the three different native language groups that took this test. The overall finding, however, was that the DIF could be accounted for by the less proficient test takers' lack of ability to complete all the test tasks and that the DIF was evidently not related to the content of the test items. Van der Slik and Weideman (2010: 115) explain this finding as follows:

The primary reason for the occurrence of DIF is not the biased content of the test items, but because they are situated at the end of the test, a test that students less capable of handling the demands of academic discourse at this level are less able to complete than those who can competently and fluently handle the demands of cognitive processing and language associated with tertiary education.

To the best of the researcher's knowledge, no finding of this kind has been reported about the Grade 12 English examination, PTESAL and NBT AL in the language

testing literature. This is another piece of research information on which this study hypothesizes that TALL will possess incremental validity which is better than that of the other three assessments.

### **2.3.7 There is an acceptable degree of internal correlations between the TALL test tasks**

Another procedure that has been used to generate evidence for construct validity involves computing the correlation of the scores obtained on all test sections (Bachman 1990: 258; Alderson, Clapham & Wall 2005: 184). This is the case because each test task should measure a different aspect of the construct a test is intended to measure and should therefore contribute towards the measurement of the test-takers' overall possession of such a construct (Van der Walt & Steyn 2008: 196). Alderson et al. (2005: 184) propose three types of correlations for the purpose of establishing construct validity from this point of view. These are the correlation between each pair of test tasks, the correlation between a test task and the whole test, and the correlation between each test section and the entire test minus that section (Alderson 2005: 184). The criteria that Alderson et al. (2005: 184) set for construct validation using this procedure is that firstly, the first type of correlation should be fairly low - 0.3 to 0.5 - because a high correlation – 0.8 to 0.9 - might mean that the two test tasks are testing the same dimension of the construct involved. They have also argued that the correlation between a test section and the whole test should be higher - .07 or more – because a test taker's score from the whole test represents their mean score from all the test sections. Finally, Alderson et al. have suggested that the correlation between each test task

and the entire test minus the test task should be lower than those between each test task and the entire test.

Van der Walt and Steyn (2007, 2008) investigated the construct validity of TAG using the criteria of Alderson et al. (2005) and found that eight of fifteen correlations satisfied the first criterion, three of six correlations met the second criterion and all correlations satisfied the last criterion. Evidently, TAG failed to achieve a complete satisfaction in terms of these criteria, and some effort needs to be made to improve it in this regard. However, some room for this situation needs to be allowed because, as measurement specialists have consistently argued, no test is 100% perfect. Unlike the Grade 12 English examination, PTESAL and NBT AL, the extent of the construct validity of TALL from this point of view has been researched and published. This is another of the bases on which this study hypothesizes that TALL will have better incremental validity than the other three assessments.

### **2.3.8 Studies have been conducted to obtain feedback from TALL test-takers**

In the main, tests are designed, developed and administered to measure the test taker's mastery of the ability that the test user is interested in. To use the words of Davies (1990: 17), tests are "intended above all to clarify the difference in the matter under test, in what is being tested (proficiency, aptitude, achievement) among the candidates". In language testing, however, studies have generated evidence to show that variance in test scores is also affected by the different processes, experiences and strategies that test takers engage in when taking a test

(Bachman 2004). The role played by these processes should therefore be considered when a test's construct validity is under scrutiny (Messick 1989). Bachman (2004: 276) raises questions that point to the relevance of these experiences to the validity of a test's construct:

To what extent are the processes that test takers use to answer a task typical of the processes that language users would employ in responding to similar tasks in the TLU domain? Are these processes included in our construct definition?

Measurement researchers have addressed this concern by asking test-takers to give a report of their own experiences of taking a test (Van der Walt & Steyn 2007, 2008). Such a report can be generated by the test taker while in the process of responding to test tasks in what is known as "think aloud" protocols (Bachman & Palmer 1996; Van der Walt & Steyn 2008). Alternatively, the report can be compiled after the test is taken in what is called a retrospective verbal report (Bachman 2004). Records of these verbal reports are known as verbal protocols and can subsequently be qualitatively and quantitatively analyzed by the test developer in what is known as verbal protocol analysis (Bachman 2004).

In a bid to establish the construct validity of TAG from the angle of test taker experience, Van der Walt and Steyn (2007) distributed questionnaires to extract feedback from a group of 754 test takers at the Potchefstroom campus of the North West University regarding their familiarity with the tasks used in the test. The feedback the two researchers received was that the test was not adequately transparent and that its developers had to make some effort to make the test and its format more familiar to test takers. Secondly, using the same questionnaire, Van



der Walt and Steyn (2007) elicited information from the test takers regarding their perception of the conduciveness of the conditions under which the test was administered. The general perception of the test takers was that such circumstances were not ideal and that this could impact the validity of the test's scores negatively. Thirdly, the researchers wanted to establish the test takers' perception of whether the test seemed relevant to their studies. Only 45% of the respondents felt that the test had relevance to their studies. Finally, Van der Walt and Steyn (2007) aimed at finding out through the questionnaire whether the test takers were clear about what was required of them by the test tasks. Only 68% of the respondents indicated that they were confident about how they were expected to respond to most tasks. Asked if they could finish taking the test in the allotted time, only 14 percent indicated they had been able to do so. All these shortcomings of the test notwithstanding, TALL remains the only one of the four tests whose psychometric properties have been researched and are known. As argued through this chapter, this is the primary basis for the hypothesis advanced in this study.

## **2.4 Conclusion**

This chapter was a review of the literature that is particularly relevant to the hypothesis of the present study. Firstly, the review reveals that, in relation to the Grade 12 English examination, PTESAL and NBT AL, so much is known about the reliability and validity of TALL that this makes it the ideal test to possess better incremental validity with regard to first year academic performance when compared to the other three assessments. Secondly, the review shows, however, that TALL is not a perfect test that meets all the requirements of psychometric

soundness. This is no reason for one to hypothesize otherwise though. Measurement specialists are in total agreement that no test is completely flawless. In the researcher's view, what would enable TALL to possess the best possible incremental validity with regard to academic success in the chosen context of the study is the extent to which it has been researched and its – albeit not completely perfect – known psychometric qualities. In the researcher's view, it would not make research sense to predict that assessments whose qualities are not publicly known such as the Grade 12 English examination, PTESAL and the NBT AL would possess incremental validity better with regard to a variable like academic success than one whose psychometric qualities have been researched and are as known such as those of TALL.

While the current discussion has focused particularly on those aspects of test validity and validation that relate to the articulation of and motivation for the hypothesis of the thesis, it remains to be seen whether the study as a whole may be able to contribute to our general insight into these critically important dimensions of test quality. I therefore envisage returning to the potential contribution that this study might make to validity theory in the final chapter (Chapter Seven) below.

In the following chapter, we focus on the methodology used to accomplish the aim of the present study. Specifically, the chapter explains the procedure used for sampling, the relevant ethical considerations, the procedure for defining academic success and the quantitative method used for analysing the data.

## **Chapter 3: Research methodology**

### **3.1 Introduction**

This chapter describes the procedure followed to collect the data, the data sampling method that was used, the ethical research issues that were considered and the statistical techniques that were used for data analysis in this study. Finally, it describes the specifications of the four assessments, namely the National Benchmark Test in Academic Literacy (NBT AL), the Proficiency Test in English Second Language Advanced Level (PTESLAL), the Test of Academic Literacy Levels (TALL) and Grade 12 English, whose incremental validity will be investigated in the study. All these are dealt with one after another in the sections that follow.

### **3.2 Data collection**

As pointed out in Chapter One, the key focus of this study was three comparisons of the incremental validity of Grade 12 English examinations, PTESLAL, NBT AL and TALL. The first of these comparisons involved scores obtained on the PTESLAL written by the participants at the end of 2011, and who were applying for admission to the Central University of Technology (CUT) for the 2012 academic year. These scores were compared to those obtained on Grade 12 English in November 2011 as well as those on the NBT AL that was administered at the beginning of 2012 at this university. The outcome variable for these predictors was the participants' average academic performance at the end of 2012, their first year of academic study.

The next stage of this study involved a comparison of the incremental validity of the scores on Grade 12 English, PTESLAL and TALL. In this case, the participants had taken PTESLAL at CUT late in 2013 and had subsequently been admitted to different programmes offered by this university in 2014. TALL was then administered to the same group of test takers at the beginning of 2014 and the scores on the test were provided by ICELDA in March of the same year. In this case, the outcome variable was the participants' average scores at the end of 2014, their first year of academic study at this university.

The last stage of this study was a comparison of the incremental validity of the subset of the TALL scores referred to above and the Grade 12 English results from the end of 2013. The outcome variable for the two predictors was the participants' average performance at the end of 2014, also their first year of academic study. Except for those on TALL, the scores on the rest of the variables were officially requested from the university. Permission to use these scores was sought from and granted by the Registrar's Office.

### **3.3 Sampling**

More than it was the case with the other predictor assessments involved in this study, determining the validity of PTESLAL was, in the researcher's view, the first priority for the study, especially in the context of the high stakes purpose for which the test was used at CUT. As pointed out in Chapter One, at this university, performance on the test was alternatively used for making access decisions where Grade 12 results' ability to provide information about readiness for university education was deemed inadequate. This made it important that the relationship of

this performance and academic success at this university was established. For this reason, the process of sampling the data for this study first depended on the availability of scores on PTESLAL for the 2012 and 2014 intake cycles. In other words, the participants' scores on PTESLAL were first obtained and compiled and those on the other variables, both predictive and predicted, were obtained and compiled alongside those of PTESLAL. This means that the participants first had to have a score on the latter test before they were selected for participation in the study. Similarly, and as will again be shown in Chapter Four, these scores were used only if the participants had scores on the other predictor variables referred to earlier as well as on the two outcome variables, namely, end of 2012 and 2014 average scores.

Typically, not a large and equal number of applicants are required to take PTESLAL every year at CUT. As will again be shown in Chapter Four, more of these participants wrote this test for the 2012 intake cycle than they did for admission in 2014. Nevertheless, the data were deemed sufficient for the analyses carried out in this study to be accomplished. As will again be shown in Chapter Four, the sample size for the statistical comparisons carried out was acceptable for a study of this nature. Dornyei (2007: 99) has described sample size requirements for applied linguistics research in general in the following terms:

The following rough estimates of sample sizes for specific types of quantitative methods have also been agreed on by several scholars: correlational research – at least 30 participants; comparative and experimental procedure – at least 15 participants in each group; factor analytic and other multivariate procedures – at least 100 participants.

Mackey and Gass (2005: 124) have added that “we must remember, however, that research in general education tends to have access to (and to utilize) larger pools than second language research. In second language studies, small groups are sometimes appropriate as long as the techniques for analysis take the numbers into account”

The kind of sampling used in this study is known as convenience sampling, “the selection of individuals who happen to be available for a study” (Mackey & Gass 2005: 122). In the words of Dornyei (2007: 98-99), convenience sampling happens

where an important criterion of sample selection is the convenience of the researcher: members of the target population are selected for the purpose of the study if they meet certain practical criteria, such as geographical proximity, availability at a certain time, easy accessibility, or the willingness to volunteer.

The disadvantage of convenience sampling is that “it is likely to be biased and should not be taken to be representative of the population” (Mackey & Gass 2005: 122). Mackey and Gass (2005) point out at the same time, however, that this kind of sampling is very common in second language and applied linguistics research in general. Given the shortcoming of convenient sampling referred to above, “we need to describe in sufficient detail the limitations of such samples when we report the results, while also highlighting the characteristics that the particular sample shares with the defined target population. In a similar vein, we also have to be particularly careful about the claims we make about the more general relevance of our findings” (Dornyei 2007: 99).

### 3.4 Ethical considerations

Most books on research methods in the fields of social sciences and medicine devote space and time to the importance of research ethics for the purpose of protecting those who participate in research from possible harm (Dornyei 2007). Salkind (2006) argues, for example, that researchers should always be mindful of how they treat those who participate in their studies and how these participants will benefit from such studies. In the words of Salkind (2006: 58), “subjects must ... be prevented from physical or psychological harm. If there is any doubt at the outset that there is a significant risk involved ... then the experiment should not be approved.” In general, a way recommended for ensuring that harm is not experienced by those who participate in a study involves obtaining informed consent from them. As Salkind (2006: 59) argues: “Without question, every research project that uses human participants should have an informed consent form that is read and signed by each participant ...” Obtaining this consent requires that the participants be familiarized with all aspects of the research study, including the possible risks and benefits involved (Mackey and Gass 2005). On the basis of this, the participants can decide whether they willingly want to participate in a study or not. Thus, not only does obtaining informed consent help ensure that research participants are not harmed, but it also functions as a tool for protecting them from being coerced into participating in a study.

Researchers in second language research in particular and applied linguistics research in general tend to argue, however, that obtaining informed consent from those participating in a study is unnecessary in the case of these disciplines

because, in their view, studies conducted in these fields are beneficial instead of posing any harm to participants (Mackey and Gass 2005). For the same reason, Johnson and Christensen (2004: 111) have argued, for example, for the need for this kind of educational research to be exempted from having to comply with the informed consent criteria:

Fortunately, studies conducted by educational researchers seldom if ever run the risk of inflicting such severe mental and physical harm on participants. In fact, educational research has historically engaged in research that imposes either minimal or no risk to the participants and has enjoyed a special status with respect to formal ethical oversight.

This is the approach that was adopted in the data collection process involving TALL in particular in the present study. In other words, obtaining the informed consent of the participants was deliberately omitted from the data collection process. This was done for two reasons. Firstly, except that it would cost them time and unrewarded effort to participate in the study, the study itself was not considered to pose any serious harm to the participants. Secondly, TALL was one of the measurement variables in the study, the reliability and validity of which and, by extension, of the study itself, depended on how genuinely the participants received and responded to the test. In other words, it was possible that the participants would deliberately not apply themselves in taking the test had they been informed that it was administered solely for research purposes. The variance in the scores from the test could be a consequence of factors irrelevant to the ability targeted by the test. Compiling an informed consent form and asking the test takers to complete it would, in other words, divulge this information and would as Mackey and Gass (2005: 30) have put it, result in “giving away the goals of the



study”. Indeed, as Rounds (1996) has pointed out, in second language research, withholding this information from the participants is often necessary: “sometimes ... a research design requires that the researcher conceal her real interests, and perhaps use small deceptions to deal with the classic ‘observer’s paradox’ (p. 53).”

Dornyei (2007: 70) adds to this as follows:

It does not require much justification that sometimes researchers cannot provide full disclosure of the nature and purpose of the study without causing participant bias or even invalidating the study, and in some (rare) cases the researcher needs not only to withhold some information but to actively mislead the participants.

In order to protect the reliability and validity of TALL from the potential threat posed by obtaining informed consent from the participants in this study, they were simply informed that it was a requirement by CUT rules that first year students take the test at the beginning of their first year of enrolment and that the scores would count towards their end of semester academic literacy course mark. Deceiving the participants in this manner is acceptable when it is absolutely necessary but should be followed by a debriefing of the participants about the real purpose of the study (Dornyei 2007). The deception should, however, not compromise the welfare of the participants in any significant way (Dornyei 2007) and in the case of TALL, the deception did not pose any harm to the participants. The participants were briefed on the real purpose of the testing following their harmless deception by the researcher.

At the time the data collection for this study started, it was not necessary to obtain informed consent from the participants regarding Grade 12 English, PTESLAL and NBT AL. Firstly, the Grade12 English examination is an official requirement for

obtaining the National Senior Certificate (NSC). Secondly, taking PTESLAL is, as pointed out in Chapter One, a requirement for those students applying for admission to CUT but whose Grade 12 results do not meet the required 27 points for straight admission to academic programmes. Lastly, the NBT AL administered at CUT at the beginning of 2012 was an official arrangement of the university that was mandatory for all first year students. The scores on these three assessments were therefore already available at the time the data collection process for the study started.

Another ethical consideration in the research involving human subjects relates to how the confidentiality of the data is handled and how the participants' anonymity is protected. As Dornyei (2007: 68) argues, every participant has the right "to remain anonymous and if the participants' identity is known to the research group, it is the researcher's moral and professional obligation ... to maintain the level of confidentiality that was promised at the outset". Consistent with the design of the present study and for the reasons already given, no informed consent was sought from the participants and neither was any open commitment to protect the confidentiality of the data and the anonymity of the participants made by the researcher to the participants with regard to the administration of TALL. The latter step would, just as obtaining informed consent from the participants, jeopardize the reliability and validity of the scores the test would yield. In any event, given the primary aim of this study, the findings of the research would, without making any reference to the individual identities of the participants, be made available to the various department of the CUT community and eventually, to the academic public.

While the groups may therefore have been momentarily identifiable, no individual student's identity was either revealed or compromised.

### **3.5 Procedure for defining academic success**

In order to compare the incremental validity of the four assessments with regard to the first year academic success of the sample used, it was necessary that the concept of 'academic success' was defined for the study. Several of these definitions are possible. Smit, Boraine and Owen (2006) offer three types of such definitions. The first of these is called "pcredit" and involves dividing the number of credits a participant successfully completes by the total number of credits for which they register at the beginning of the year. The second procedure these authors propose is "pprogram". In this method, academic success is determined by dividing the number of credits a participant successfully obtains by the total number of credits they are required to complete in a year. The last procedure is called "pmod". This one involves dividing the number of modules a participant passes by the total number of modules for which they enrol at the beginning of a year. Yet another model commonly used for defining academic success involves the end-of-course average performance of a participant on all the courses they enrol for in a programme. Since the procedures referred to above were all to a degree contestable, average performance was the model used in the present study. The reason for choosing the model was that both this average and the scores obtained on the variables used as predictors in the present study are typically reported as percentages. This means that all the variables would be dealt with at the same level

of measurement and that this would enhance the accuracy of the statistical computations to be carried out.

### **3.6 Procedure for data analysis**

The present study was mainly quantitative in nature. The basis for quantitative research involves three stages, namely, identifying a research problem, generating a hypothesis based on the identified problem and testing this hypothesis by using scientific methods to collect and analyse the data (Dornyei 2007: 31). This hypothesis becomes accepted as a scientific theory or law when it has been successfully tested and validated by replicating the initial study (Dornyei 2007: 31). The procedure therefore offers “a tool to explore questions in an ‘objective’ manner, trying to minimize the influence of any researcher bias or prejudice, thereby resulting in what scholars believed was an accurate and reliable description of the world” (Dornyei 2007: 31). Dornyei (2007: 32-34) has characterised quantitative research methodology as one that involves the use of numbers, specifies categories and values before the beginning of a study, focuses on the features of groups of people and less on those of individuals, uses statistics to analyse data, uses standardized procedures to assess empirical data and aims for the generalizability of the findings of a study. Mackey and Gass (2005) observe, furthermore, that quantitative research is associational and experimental in nature and that a common feature of these types of analysis is that researchers’ aim in both cases is to investigate a relationship between or within variables. Mackey and Gass (2005: 137) have contrasted the foci of associational and experimental research in the following words:

The goal of associational research is to determine whether a relationship exists between variables and, if so, the strength of that relationship... In experimental studies, researchers deliberately manipulate one or more variables (independent variables) to determine the effect on another variable (dependent variable). This manipulation is usually described as a treatment....

Furthermore, associational researchers do not concern themselves with the causal relationship between variables but with the co-occurrence of such variables, and correlation is the statistic they commonly use to determine this (Mackey & Gass 2005). In contrast, the experimental researcher's aim is to investigate causal relationships between variables and the procedures they use to do this involve a comparison of pre-treatment and post-treatment performance (Mackey & Gass 2005).

It has been pointed out several times so far that the aim of the present study was to investigate the incremental validity of four language assessments of readiness for university education, namely, Grade 12 English, PTESLAL, NBT AL and TALL. In this sense, the study was associational by nature and as Hunsley and Meyer (2003: 450) observe, the typical manner in which incremental validity is assessed in correlational designs such as this one, "is by using hierarchical multiple regression analyses to determine the contribution of one measure to the prediction of the criterion after one or more other variables have been entered into the analysis." Haynes and Lench (2003: 461) have similarly argued that the first step of an incremental validity analytic sequence "is to examine a zero-order correlation matrix that includes all predictor and criterion variables". Haynes and Lench (2003:462) also add that once the zero-order correlations among measures have been determined, "the incremental validity of a new measure is most often

examined through a hierarchical linear regression analysis”. For the purpose of data analysis in the present study, however, advice was also sought from and provided by the Statistical Consultation Unit of the University of the Free State which recommended, in line with the views of Hunsley and Meyer (2003) and Haynes and Lench (2003) presented above, the use of descriptive statistics and inferential statistics such as correlation, linear regression and multiple regressions (cf. Morgan, Leech, Gloeckner & Barrett, 2011; Bachman 2004).

Descriptive statistics are a tool used by the researcher to summarize data in order to generate an overall understanding of a data set (Mackey & Gass 2005). As Dornyei (2007: 213) further points out, descriptive statistics “are indispensable when we share our results ... and they also form the basis of further inferential statistics”. The descriptive statistics computed and reported in the present study were the Mean, a measure of central tendency, and the Standard Deviation, a measure of variability. The Mean is “a type of average where scores are summed and divided by the number of observations” (Salkind 2011: 433-435) while the Standard Deviation is the average distance of all observed scores from the Mean (Dornyei 2007). The Mean is the commonly used measure of central tendency because it takes all the scores into account (Dornyei 2007). The Standard Deviation is high and contains extreme scores when computed for heterogeneous samples and low for homogeneous samples where scores are clustered around the Mean (Dornyei 2007).

The inferential statistics used for the analysis of data in the present study were correlations, linear and multiple regressions. Correlation is the term used to refer

to the nature and degree of association between variables. Also, correlation is the basis for regression, an investigation of the degree to which the value of one variable can be used to predict that of another. Linear regression is a less sophisticated form of regression where the correlational relationship between one independent variable, also known as the predictor variable, and one dependent variable, also referred to as the outcome variable, is examined (Cohen and Swerdlik 2010: 133). Multiple regression is a form of regression analysis that allows the researcher to understand the degree to which two or more independent variables can be used to predict one dependent variable. In the words of Bachman (2004: 110), multiple regression allows a data analyst to “regress a given dependent variable not on a single but, on multiple predictors, or independent variables” (Bachman 2004: 110). Bachman (2004: 110) further explains this procedure in the following terms:

... if we know something about the relationship between variables X and Y, we can make a more accurate prediction about unknown values of Y than if we only knew the mean of Y. It follows that if we knew the relationship between a given variable, Y, and several other variables, X1, X2, ... Xn, then we might be able to predict future values of Y even more accurately.

When determining the incremental validity of variables using the multiple regression procedure, predictor variables that correlate highly with the outcome variable are given more weight because this means that the regression coefficients and, by extension, the predictive efficiency of such variables are high (Cohen and Swerdlik 2010: 135). Furthermore, predictor variables that do not correlate with others but correlate highly with the outcome variable may be given relatively even

more weight because such predictors probably provide predictive information from a unique angle (Cohen and Swerdlik 2010: 135).

In the case of the present study, however, it was necessary for an adjustment to be made with regard to the second condition for incremental validity suggested by Cohen and Swerdlik (2010) above. The predictor variables used were all language assessments of different constructs of English language performance and logically, a zero correlation of the scores obtained on these variables was unlikely. It was necessary therefore to determine the cut-score for the acceptable degree of multicollinearity between predictor variables, below which they could be judged to possess incremental validity. Dornyei (2007: 223) has argued that in applied linguistic research, correlations of .60 and above mean that the degree of overlap between the tests involved show that they measure almost the same construct. This was the definition of multicollinearity that was used in this study. As will again be shown in Chapter Four, this means that any of the tests used would be judged to have satisfied the second condition of incremental validity set by Cohen and Swerdlik (2010) above if its correlation with other predictors was below .60 and if they relate significantly with the outcome variable.

Like other types of correlational analyses, regression coefficients in studies involving both linear and multiple regressions can range from +1 to -1 where +1 denotes a perfect positive correlation while -1 represents a perfect association of the variables in the opposite direction (Mackey and Gass 2005). Mackey and Gass (2005: 286) explain the meaning of the difference between positive and negative correlation and by extension, regression coefficients thus:



... correlation coefficients can be expressed as positive and negative values. A positive value means that there is a positive relationship; for example, the more talk, the taller the child. Conversely, a negative value means a negative relationship – the more talk, the shorter the child.

Prior to the presentation and discussion of the results of this study in the next chapter, it is necessary that exactly what the four predictor instruments purport to measure, how they measure it and the extent to which they do so are first examined. The specifications on the basis of which NBT AL, PTESLAL, TALL and Grade 12 English were developed are therefore dealt with in the section below.

### 3.7 Test Specifications

#### 3.7.1 The National Benchmark Test of Academic Literacy

The NBT AL comprises 75 multiple choice items that are aligned with the construct informing this test and that are based on texts that mirror those that test takers are typically required to process in academic settings (Cliff 2015). The items require the test taker to “choose the most inclusive or plausible or reasonable answer from four options, where distractors have been specifically designed to be indicative of reading and reasoning misconceptions” (Cliff 2015: 11). The subdomains of academic literacy that underpin these items are presented and explained in **Table 7** below:

**Table 7: The subdomains of NBT AL**

<b>Subdomain</b>	<b>Description</b>
Communicative function	Students’ abilities to ‘see’ how parts of sentences / discourse define other parts; or are examples of ideas or are supports for arguments; or attempts to persuade.
Inferencing	Students’ capacities to draw conclusions and apply insights, either on the basis of what is stated in texts or is implied by these texts.
Vocabulary	Students’ abilities to derive/work out word meanings from their context
Relations	Students’ capacities to ‘see’ the structure and

1. Cohesion 2. Discourse	organisation of discourse and argument, by paying attention – within and between paragraphs in text – to transitions in argument; superordinate and subordinate ideas; introductions and conclusions; logical development.
Essential/non-essential	Students’ capacities to ‘see’ main ideas and supporting detail; statements and examples; facts and opinions; propositions and their arguments; being able to classify, categorise and ‘label’.
Grammar/syntax	Students’ abilities to ‘see’ / analyse the way in which sentence structure / word, phrase order affects meaning and emphasis in language
Metaphor	Students’ abilities to understand and work with metaphor in language. This includes their capacity to perceive language connotation, word play, ambiguity, idiomatic expressions, and so on
Text genre	Students’ abilities to perceive ‘audience’ in text and purpose in writing, including an ability to understand text register (formality / informality) and tone (didactic / informative / persuasive / etc.).

(NBTP 2015a)

The NBT AL items are systematically designed to measure these subdomains to varying degrees of length and complexity as determined by the developers of this test. The proportions of items for these subdomains in the test as well as the different levels of cognition at which the assessment of the subdomains are pitched are captured in **Table 8** below.

**Table 8: Levels of cognitive challenge for the NBT AL**

Specifications	Reproducing orientation				Transformative orientation			
	Level 1 Knowing		Level 2 Applying routine procedures in familiar contexts		Level 3 Applying multi-step procedures in a variety of contexts		Level 4 Reasoning and reflecting	
	1+	1-	2+	2-	3+	3-	4+	4-
Vocabulary (10%)								
Metaphorical (15%)								
Inferencing								

(15%)								
Communicative function (15%)								
Relations cohesion (5%)								
Relations discourse (10%)								
Grammar/syntax (5%)								
Text genre (5%)								
Essential / non-essential (20%)								
Number of items								

(NBTP 2015a)

From **Table 8**, it is evident that the highest proportion of the test focuses on Essential/non-essential (20%), Metaphor (15%), Inferencing (15%) and Communicative Function (15%) while relatively less of it proportionally comprises items focusing on Vocabulary (10%), Discourse Relations (10%), sentence-level Cohesion (5%), Grammar/syntax and Text Genre (5%). It is also clear from **Table 8** that all these sub-domains are assessed at four levels of cognitive difficulty, the first two of which involve producing and applying knowledge while the last two involves transforming and applying this knowledge. Clearly, the two categorizations derive from the now well-known Taxonomy of Educational Objectives by Bloom et al. (1956), which recommends that educational assessment starts with the basic knowledge and comprehension of information and then moves onto focusing on the application, analysis, synthesis and evaluation of that information. The plus (+) and minus (-) signs that are presented alongside the cognitive levels of assessment associated with NBT AL in **Table 8** above are

indicative of the levels of difficulty associated with every item used in the test. So, for example, a vocabulary item allocated a minus sign is considered less difficult than one with a positive sign.

### 3.7.2 Proficiency Test English Second Language Advanced Level

PTESLAL comprises a total of 40 items presented in multiple-choice format and in the reading mode. The specifications on the basis of which these items are developed are captured in **Table 9** below.

**Table 9: Test specifications for PTESLAL**

Skill being tested	No. of items	Item numbers
<ul style="list-style-type: none"> <li>Recognizing paraphrased meaning of common idioms</li> </ul>	2	9;20
<ul style="list-style-type: none"> <li>Making <i>general</i> inferences based on the given text</li> </ul>	8	1;4;5;6;8;10;11;14
<ul style="list-style-type: none"> <li>Making inferences related to <i>diction</i> – writer’s choice of words in the context.</li> </ul>	1	29
<ul style="list-style-type: none"> <li>Making inferences related to the writer’s intention</li> </ul>	3	12;13;15
<ul style="list-style-type: none"> <li>Making inferences related to <i>setting</i> or atmosphere</li> </ul>	1	7
<ul style="list-style-type: none"> <li>Selecting appropriate language for audience/situation/circumstance</li> </ul>	2	3;22
<ul style="list-style-type: none"> <li>Accurately communicating summary of intended meaning: headlines, recognizing redundancy</li> </ul>	2	25;30
<ul style="list-style-type: none"> <li>Editing: being consistent about time, i.e. recognizing incorrect use of tenses</li> </ul>	3	16;24
<ul style="list-style-type: none"> <li>Combining of simple sentences to form complex sentences</li> </ul>	1	32;35;36
<ul style="list-style-type: none"> <li>Meaningful paraphrasing – selecting best opening or concluding sentence or arranging</li> </ul>	2	31

sentences meaningfully		
• Selecting precise word to describe something in context	1	17
• Selecting words/phrases used deliberately to express or stir emotions	1	23
• Recognizing correct idiomatic and functional use of verbs	3	33;34;40
• Recognizing correct idiomatic and functional use of conjunctions	1	18
• Prefixes and suffixes	1	19
• Punctuation	2	37;38
• Word order	2	28;39
• Changing actives to passives	1	21
• Changing statement to questions.	1	26
TOTAL	40	

(HSRC 1991: 16-17)

It is evident from **Table 9** above that the subdomains for PTESLAL are each also allocated an unequal proportion of focus in the test. Clearly, the ‘making *general* inferences based on the given text’ subconstruct of this test is allocated the highest number (8) of items (20%) while the allocation for the rest of the other subdomains ranges from 1 (2,5%) to 3 items (7,5%).

### 3.7.3 Test of Academic Literacy Levels

From the construct of academic literacy that they formulated, Van Dyk and Weideman (2004: 141) proposed the specifications and task types into which the construct of TALL may be operationalized. These are presented in **Table 10** below.

**Table 10: The subdomains and task types proposed for TALL**

<b>Specifications</b>	<b>Task Types</b>
Vocabulary comprehension	Vocabulary knowledge tests Cloze procedure
Understanding metaphor & idiom	Text comprehension
Textually (cohesion and grammar)	Scrambled text Close procedure Text comprehension
Understanding text type (genre)	Register and text type tasks Interpreting and understanding visual & graphic information Scrambled text Cloze procedure Text comprehension
Understanding visual and graphic information	Interpreting and understanding visual & graphic information (potentially:) Text comprehension
Distinguishing essential/non-essential	Text comprehension Interpreting and understanding visual & graphic information
Numerical computation	Interpreting and understanding visual & graphic information Text comprehension
Extrapolation and application	Text comprehension Interpreting and understanding visual & graphic information
Communicative function	Text comprehension passages (possibly also:) Cloze procedure, Scrambled text
Making meaning beyond the sentence	Text comprehension Register and text type tasks Scrambled text Interpreting and understanding visual & graphic information

The version of TALL that was used for the present study comprised 62 multiple-choice items that were distributed over six sections and that measured the subdomains presented in **Table 10** above to varying extents. In other words, some subdomains were allocated more focus in the test than others. This is reflected in the different number of items and the different points allocated to these items in **Table 11** below.

**Table 11: Task types, number of items and mark allocation for items in TALL**

<b>Section</b>	<b>Task type</b>	<b>Number of items</b>	<b>Total mark</b>
1	Scrambled text	5	5
2	Knowledge of academic vocabulary	10	20
3	Interpreting graphs and visual information	7	7
4	Text type	5	5
5	Understanding texts	20	47
6	Text editing	16	16
Total		62	100

Clearly, the biggest proportion of this test comprised items focusing on Understanding texts (20 items), Text editing (16 items) and Knowledge of vocabulary (10 items) while a relatively smaller proportion of the test comprised items in the sections on Scrambled text (5 items), Interpreting graphs and visuals (7 items) and Text type (5 items). Candidates are required to complete the test in 60 minutes and the maximum points they could earn equal 100, with approximately half of the items carrying 2 or 3 points instead of 1 (Van der Slik & Weideman 2009).

### **3.7.4 The Grade 12 English Home and First Language examinations**

The Grade 12 English Home and First Additional Language final assessments take two forms. These are the end of year written examinations and the oral assessment carried out throughout the year. The written examinations consist of three papers that are assessed out of 250 marks. In both examinations, Paper 1 is intended to measure Language in context and comprises three sections: Comprehension,

Summary and Language structures and conventions. The latter section focuses on Vocabulary knowledge and language use, Sentence structures and Critical language awareness. Paper 2 assesses literary appreciation and also has three sections focusing on Poetry, Fiction and Drama. For this section, First Additional Language candidates are required to complete tasks on any two of these genres while Home Language candidates are required to respond to questions on all three. Paper 3 assesses Writing. For Home Language candidates, this paper consists of two sections on Essay and Transactional text production. For First Additional Language candidates, the same section consists of three sections on Essay as well as Longer and Shorter Transactional text writing. For both, longer transactional texts include, among others, friendly/formal letters, formal and informal letters to the press, a curriculum vitae and covering letter, an obituary, agenda and minutes of meeting, and a report. For First Additional Language candidates, shorter transactional texts include advertisements, diary entries, postcards, filling in forms, and so on. Another slight difference between the two examinations is that marks are allocated slightly differently in the sections of the three papers referred to above. Finally, the oral assessment for both examinations is assessed out of 50 marks. The tasks constituting this assessment are prepared speech, unprepared speech and listening. The framework informing the Grade 12 English Home and First Additional Language final assessments is summarised in **Table 12** below.



**Table 12: Specifications for Grade 12 English HL and FAL final assessments**

<b>Paper</b>	<b>Task name</b>	<b>English HL mark allocation</b>	<b>English FAL mark allocation</b>
1. Language in context	Comprehension	30	30
	Summary	10	10
	Language structures and conventions	30	40
	<b>Total</b>	<b>70</b>	<b>80</b>
2. Literature	Poetry	30	35
	Novel	25	35
	Drama	25	
	<b>Total</b>	<b>80</b>	<b>70</b>
3. Writing	Essay	50	50
	Longer transactional texts Shorter transactional texts	50	30 20
	<b>Total</b>	<b>100</b>	<b>100</b>
4. Oral	Prepared speech	20	20
	Unprepared speech	15	20
	Listening for comprehension	15	10
	<b>Total</b>	<b>50</b>	<b>50</b>

It is evident from **Table 12** above that in the Grade 12 Home Language assessment the largest proportion of the marks is allocated to Writing (100 marks), Literature (80 marks) and Language in context (70 marks) while a relatively smaller portion is allocated to Oral assessment (50 marks). This pattern is the same for First Additional Language assessment: A large proportion of marks is allocated to Writing (100 marks), Language in context (80 marks) and Literature (70 marks) while a relatively smaller portion is allocated to Oral assessment (50 marks). In both these examinations, the same framework is used for ensuring that learner

achievement assessment is pitched at varying levels of cognitive demand. This framework is presented in **Table 13 below**.

**Table 13: Levels of cognitive challenge for Grade 12 English HL and FAL assessments**

<b>Cognitive Levels</b>	<b>Activity</b>	<b>Percentage of task</b>
<b>Literal (Level 1)</b>	Questions that deal with information explicitly stated in the text	40%
<b>Reorganization (Level 2)</b>	Questions that require analysis, synthesis or organisation of information explicitly stated in the text	
<b>Inference (Level 3)</b>	Questions that require a candidate's engagement with information explicitly stated in the text in terms of his /her personal experience.	40%
<b>Evaluation (level 4)</b>	Questions that deal with judgements concerning value and worth. These include judgement regarding reality, credibility, facts and opinions, validity, logic, and reasoning, and issues such as the desirability and acceptability of decisions and actions in terms of moral values	20%
<b>Appreciation (Level 5)</b>	Questions intended to assess the psychological and aesthetic impact of the text on the candidate. They focus on emotional responses to the content, identification with characters or incidents, and the reactions to writer's use of language (such as word choice and imagery)	

(Department of Basic Education 2011: 76)

### 3.8 Conclusion

This chapter has focused on explaining the data collection and sampling procedure used for the present study. It has made the point that data collection was mainly

dependent on available data for one of the tests and that convenient sampling was therefore used. The chapter has also focused on the possible models of defining academic success and indicated that the participants' end of first year average score was the definition of academic success for this study. Also, the chapter described the statistical procedure used for analysing the data in the study. Finally, it described the specifications on the basis of which the predictor assessments investigated have been developed.

The next chapter will present the results of this study, starting with a brief recap of the statistical procedures used for data analysis.

## **Chapter 4: Data analysis**

### **4.1 Introduction**

This chapter starts by identifying the data that were collected for the present study and explains how these were subsequently arranged for analysis. It then specifies and justifies the statistical procedures employed to analyse these data. Finally, it presents and discusses the results of all the analyses carried out. The results are presented and discussed in response to the key question of the study: Does the National Benchmark Test in Academic Literacy (NBT AL), Proficiency Test English Second Language Advanced Level (PTESLAL), Test of Academic Literacy Levels (TALL) or Grade 12 English possess incremental validity?

### **4.2 Results of the study**

Five sets of data were collected for analysis in this study. The first of these included scores obtained by a total of 352 first year students in 2012, on the PTESLAL and Grade 12 English examinations written in November 2011 and on the NBT AL administered in March 2012. The second set consisted of scores obtained by 102 first year students in 2014, on the PTESLAL and Grade 12 English examinations written in November 2013 and on the TALL administered in March 2014. The third set of data comprised scores obtained by a total of 637 first year students in 2014, on the Grade 12 English examinations written in November 2013 and on the TALL administered in March 2014. The last two data sets comprised the participants' average scores in their different programmes of study, in their first year at the tertiary institution being investigated, at the end of the 2012 and 2014 academic years, respectively. For the purpose of analysis, these data were

organized into three sets. The first set included NBT AL, PTESLAL and Grade 12 English scores as predictors of 2012 average. The second set comprised TALL, PTESLAL and Grade 12 English scores as predictors of 2014 average. The last set consisted of TALL and Grade 12 English scores also as predictors of 2014 average. The second of these data sets was a subset of the third. These data were analysed using the SAS software package (SCU 2015).

Four types of statistical analyses were carried out on the three data sets identified above. These were descriptive statistics, pairwise Pearson correlations, simple linear regression and multiple regression. These statistics were appropriate for a predictive validity study such as this one for several reasons. In the first place, descriptive statistics form the basis of all inferential statistics (Mackey & Gass 2005; Dornyei 2007). Secondly, correlations are the essence of predictive statistical procedures such as linear and multiple regressions. Cohen and Swerdlik (2010: 126) have indeed argued that “if we know that there is a high correlation between  $X$  and  $Y$ , then we should be able to predict – with various degrees of accuracy, depending on other factors – the value of these variables if we know the value of the other.” Of importance and relevance to a study such as the present one, however, is the fact that correlations provide a preliminary indication of the strength of relationships between the predictors involved on the one hand, and the strength of association between each of such predictors and the criterion variable on the other (Haynes & Lench 2003: 461). These are the first two conditions for incremental validity which all predictor variables must satisfy (Cohen & Swerdlik 2010: 184). There are two main reasons for these conditions. In the first place,

logic dictates that a predictor variable must have a positive association with the criterion variable before it can add to the efficiency of others to predict that variable. Secondly, the degree of overlap or collinearity between the predictors used should not be too high to render them redundant for incremental validity in relation to each other. In the words of Haynes and Lench (2003: 462):

With a high degree of collinearity ... the measures are redundant, and each is unlikely to show significant increases in the proportion of variance accounted for in the criterion variable when added to a regression formula that includes the other.

Thirdly, linear regression functions to determine the individual predictive validity of measurement instruments so that their potential to contribute incrementally to the predictive efficiency of other predictors is first established. Put differently, linear regression is a preliminary indicator of whether an individual variable has predictive ability and ultimately, whether it might possibly add - in a regression analysis simultaneously involving other predictors - to the predictive efficiency of the predictor(s) already in use. This is the second requirement that additional predictors included in an incremental validity study must satisfy (Cohen & Swerdlik 2010: 184). Finally, multiple regression is one of the methods ultimately used to determine and measure the incremental validity of more than one predictor simultaneously in relation to the one already being used. In the present study, the last three statistics were computed only on the scores obtained by participants who had complete data on all the predictor and predicted variables involved. In other words, participants who did not have scores for any of these variables were excluded from these analyses. The results of the four statistical analyses carried out are presented in the three sections below.

### 4.3 NBT AL, PTESLAL, Grade 12 English and 2012 average

#### 4.3.1 Descriptive statistics

The first data set to be analysed included scores on NBT AL, PTESLAL and Grade 12 English as predictors of 2012 average. The descriptive statistics for the scores obtained on all these variables are presented in **Table 14** below.

**Table 14: The means and standard deviations for the scores on NBT AL, PTESLAL, Grade 12 English and 2012 average**

Variable	N	M	SD	Minimum	Maximum
2012 Average	309	57.96	10.32	0	81
NBT AL	352	43.37	10.03	22	81
PTESLAL	345	46.64	22.02	5	99
GR 12 English	225	57.40	8.43	37	81

As can be seen in **Table 14** above, the mean scores for the participants on NBT AL and PTESLAL were the lowest (M=43.37 and M=46.64 respectively) while those for their 2012 average and Grade 12 English were the highest (M=57.96 and M=57.40 respectively). The low mean scores for NBT AL and PTESLAL indicate that on average, the two tests were more challenging than the other two assessments for these participants. The probable reason for this is that the two tests were mainly designed for assessing readiness for university education which should logically be expected to be more demanding than the Grade 12 English examination. Also, the mean scores for Grade 12 English and 2012 average show that the participants' 2012 average was more attuned to their performance on Grade 12 English than it was to their performance on the other two predictors. For the reason provided

earlier, one would expect tests of academic readiness like NBT AL and PTESLAL to be more aligned with performance at university than a high school exit examination like Grade 12 English. The latter assessment mainly focuses on measuring learner achievement on high school curricula which, as will again be argued later in Chapter Five, does not necessarily translate into readiness for higher education as regards language. It can also be seen in **Table 14** that the standard deviation for the scores was the highest for PTESLAL (SD=22.02), higher for both 2012 average and NBT AL (SD=10.32 and SD=10.03 respectively) and the lowest for Grade 12 English (SD=8.43). The higher standard deviations for PTESLAL and NBT AL should be expected, because these are tests aimed to serve as additional sources of information to that obtainable from Grade 12 English and ideally, to facilitate the placement of students in various programmes of a university. This purpose of testing requires that the instruments used be able to provide greater variability in the scores they generate. In contrast, an achievement assessment such as Grade 12 English is, as pointed out earlier, mainly aimed to provide information on the extent to which the objectives of the high school curriculum have been achieved. Naturally, this kind of assessment is not expected to spread scores out the way placement tests typically do.

#### **4.3.2 Intercorrelations: NBT AL, PTESLAL, Grade 12 English and 2012 average**

The pairwise Pearson intercorrelations of the variables constituting the first data set - scores obtained by the participants on NBT AL, PTESLAL and Grade 12 English as predictors of 2012 average - are captured in **Table 15** below.



**Table 15: Intercorrelations for NBT AL, PTESLAL, Grade 12 English and 2012 average**

Variable	N	2012 average	NBT AL	PTESLAL	Gr 12 English examination
2012 Average	309	--	.10	.04	.24*
NBT AL	309		--	.53*	.50*
PTESLAL	303			--	.32*
Gr 12 English examination	223				--

\* $p < .05$

It is clear from **Table 15** above that the correlation with 2012 average was the highest for Grade 12 English ( $r = .24$ ), lower for NBT AL ( $r = .10$ ) and the lowest for PTESLAL ( $r = .04$ ). All these correlations were positive. It can also be seen in **Table 15** that the correlation coefficient for Grade 12 English and 2012 average was statistically significant ( $p = < .05$ ). In applied linguistics research, the accepted  $p =$  value for inferential statistics is .05 and below: “A  $p$ -value of .05 indicates that there is only 5% probability that the research findings are due to chance, rather than to an actual relationship between and among variables” (Mackey & Gass 2005: 265). It is probable therefore that for this population and data set, the association between Grade 12 English and 2012 average was not a result of chance. This means that participants who performed well on Grade 12 English on average tended to do the same on 2012 average and vice versa. Once again, one would expect the association in performance on 2012 average and the two tests used for academic readiness to be higher than that of performance on Grade 12 English and 2012 average. As pointed out earlier, this expectation resides in the fact that the purpose for which the two tests are used associates them more with performance at

university and that Grade 12 English is essentially an assessment of achievement at high school. The strength of association between Grade 12 English and 2012 average, however, is an early indication that if any of the predictor variables in this data set will predict academic performance better than the others at all, it will be Grade 12 English.

It is also clear from **Table 15** that the correlation coefficients for 2012 average and NBT AL on the one hand, and 2012 average and PTESLAL on the other, were not statistically significant ( $p = > .05$ ). The association between these variables may therefore have been the result of chance. Again, this happens against the logical expectation that the two tests are the ones that should have a statistically significant relationship with academic performance. Furthermore, **Table 15** shows that the correlations among the predictors, namely NBT AL, PTESLAL and Grade 12 English were all positive and statistically significant ( $p = < .05$ ). This is not a surprising result, because the three assessments are measures of different abilities of the same language. Performance on these assessments should therefore be expected to show evidence of collinearity. The positive but lower correlations for NBT AL and 2012 average on the one hand, and PTESLAL and 2012 average on the other, are an early indication that these tests stand less chance of predicting academic success better than Grade 12 English.

### **4.3.3 Linear regression analyses: NBT AL, PTESLAL, Grade 12 English and 2012 average**

#### **4.3.3.1 NBT AL as a predictor of 2012 average**

The first linear regression analysis of the first data set – NBT AL, PTESLAL and Grade 12 English as predictors of 2012 average – involved NBT AL as a predictor

of 2012 average. The first result of this analysis worth reporting is the  $F$  statistic. In both linear and multiple regressions, the  $F$  statistic is an outcome of an Analysis of Variance (ANOVA) which partitions the variation in the predicted variable that can be explained by the changing levels of the specified predictor(s) (NBT AL in this case) and the one that cannot be accounted for by changes in this predictor (also known as residual variance) (Montgomery, Peck & Vining 2006: 24). Residual variance is in this sense constituted by other possible predictors that are not specified for a study. Both these specified and unspecified predictors constitute a predictor model for which the  $F$  statistic is generated by a regression analysis. The  $F$  statistic from the current linear regression analysis was  $F(1,307) = 3.41$ ,  $p = 0.067$ . The  $p$ -value for this statistic was higher than .05, meaning that both the specified predictor (NBT AL in this case) and other unspecified predictors constituting this model did not have a positive relationship with the predicted variable. This is reflected in the last two columns of the first row in **Table 16** below.

**Table 16: The F statistic from the linear regression of 2012 average on NBT AL (n=309)**

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
<b>Model</b>	1	361.18700	361.18700	3.41	0.0657
<b>Error</b>	307	32494	105.84478		
<b>Corrected Total</b>	308	32856			

The next result worth reporting from this analysis is the R-Square. This is an indication of the extent to which the variance in performance on the predicted variable could be accounted for by the model involving NBT AL as the specified predictor (Morgan, Leech, Gloeckner & Barret 2011: 145). The R-Square value for this model equalled 0.0110, meaning that the model could account for 1.1% of the variance in the participants' average performance at the end of 2012. This is captured in the last column of the first row of **Table 17** below.

**Table 17: The R-Square for the model with NBT AL as the specified predictor of 2012 average performance (n=309)**

<b>Root MSE</b>	10.28809	<b>R-Square</b>	0.0110
<b>Dependent Mean</b>	57.96117	<b>Adj R-Sq</b>	0.0078
<b>Coeff Var</b>	17.74997		

Also worth reporting from the linear regression analysis is the *t* statistic. This is a test of whether the slope of a specified predictor variable is significantly different from zero (Montgomery, Peck & Vining 2006: 24). If statistically significant, the *t* statistic means that the null hypothesis can be rejected. For NBT AL, the *t* statistic equalled 1.85 and was not statistically significant ( $p = .07$ ). The *t* statistic for this test and its accompanying *p*-value are reflected in the last two columns of **Table 18** below.

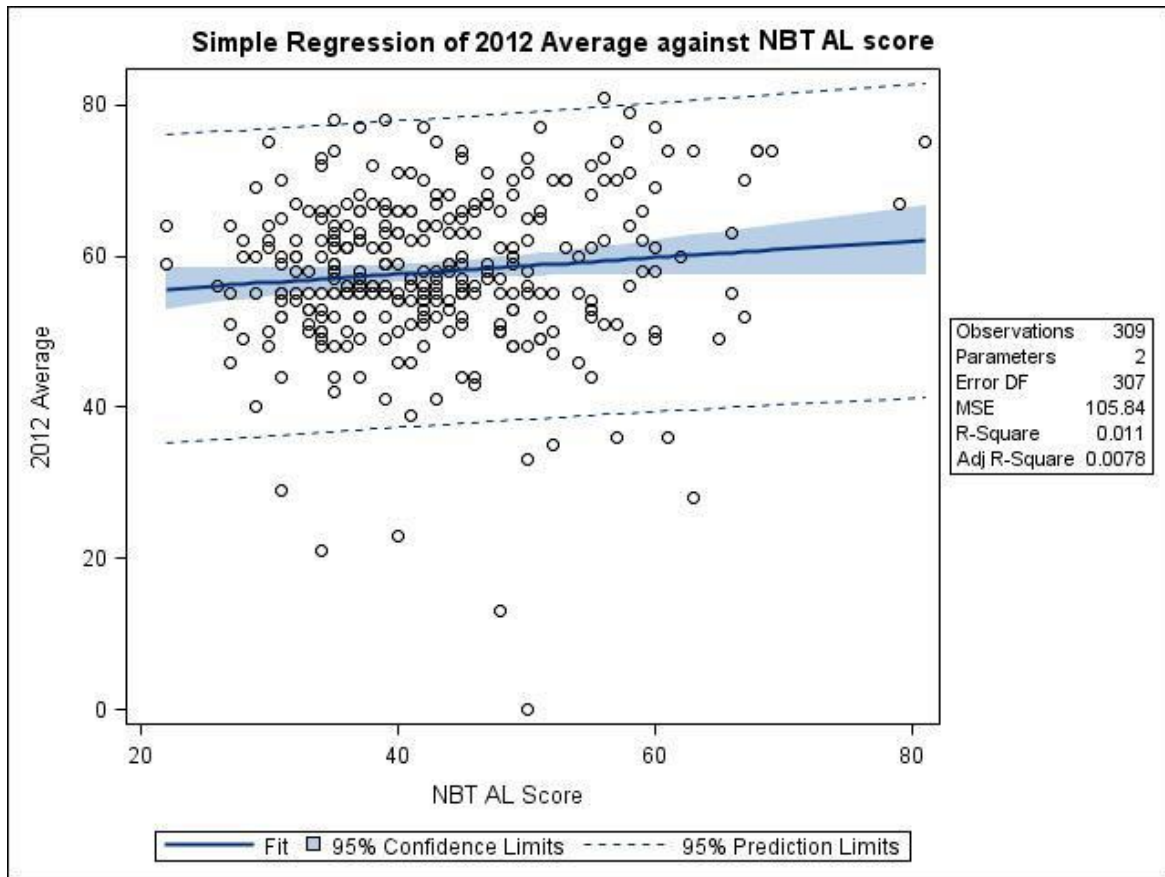
**Table 18: The t statistic for NBT AL as a predictor of 2012 average (n=309)**

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	T Value	P Value
Intercept	Intercept	1	53.26454	2.60896	20.42	<.0001
NBT AL	NBT AL score	1	0.10910	0.05906	1.85	0.0657

Also worth brief mentioning in **Table 18** above is what is reported as the Intercept. This refers to the mean of the distribution of the predicted variable when the range of scores on the predicted variable(s) includes a zero (Montgomery, Peck & Vining 2006: 12). If the latter is not the case, however, the Intercept has no interpretation of any value (Montgomery, Peck & Vining 2006: 12). In the current study, the scores obtained by the participants in the predicted variables used were all greater than zero. The Intercept will therefore not have any interpretive utility to the results of this study as a whole.

The results of a linear regression of 2012 average performance on the participants' performance on NBT AL presented above are graphically summarized in the fit plot in **Figure 3** below.

**Figure 3: The fit plot for a linear regression analysis for NBT AL as predictor of 2012 average (n=309)**



As can be seen in **Figure 3** above, the slope of the regression line for the two variables was marginally inclined in the positive direction and the bulk of the data points were mainly located away from this line.

These results mean that on its own, NBT AL was a poor predictor of 2012 average performance. The  $t$  statistic for this test from the analysis testifies to this. While this statistic was positive, it was not statistically significant, meaning that this could have happened by chance. Also, the fit plot for the results of a linear regression of 2012 average on the test confirms that the test had a very weak linear relationship with the participants' end of 2012 average performance.

#### 4.3.3.2 PTESLAL as a predictor of 2012 average

The next linear regression analysis of the first data set – NBT AL, PTESLAL and Grade 12 English as predictors of 2012 average – involved PTESLAL as a predictor of 2012 average. The  $F$  statistic from this analysis equalled  $F(1,301) = 0.67$ ,  $p = 0.4128$ , meaning that the predictor model involving PTESLAL had a positive but not statistically significant relationship with the predicted variable. This  $F$  statistic and its  $p$ -value are captured in the last two columns of **Table 19** below.

**Table 19: The F statistic for the model involving PTESLAL as the specified predictor of 2012 average (n=303)**

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	72.22254	72.22254	0.67	0.4128
Error	301	32323	107.38687		
Corrected Total	302	32396			

Secondly, the R-Square from this analysis equalled 0.0022, meaning that the model could explain 0.2% of the variance in the participants' average performance in 2012. This is reflected in the last column of the first row of **Table 20** below.

**Table 20: The R-Square for the model including PTESLAL as the specified predictor of 2012 average (n=303)**

Root MSE	10.36276	R-Square	0.0022
Dependent Mean	57.96700	Adj R-Sq	-0.0011
Coeff Var	17.87701		

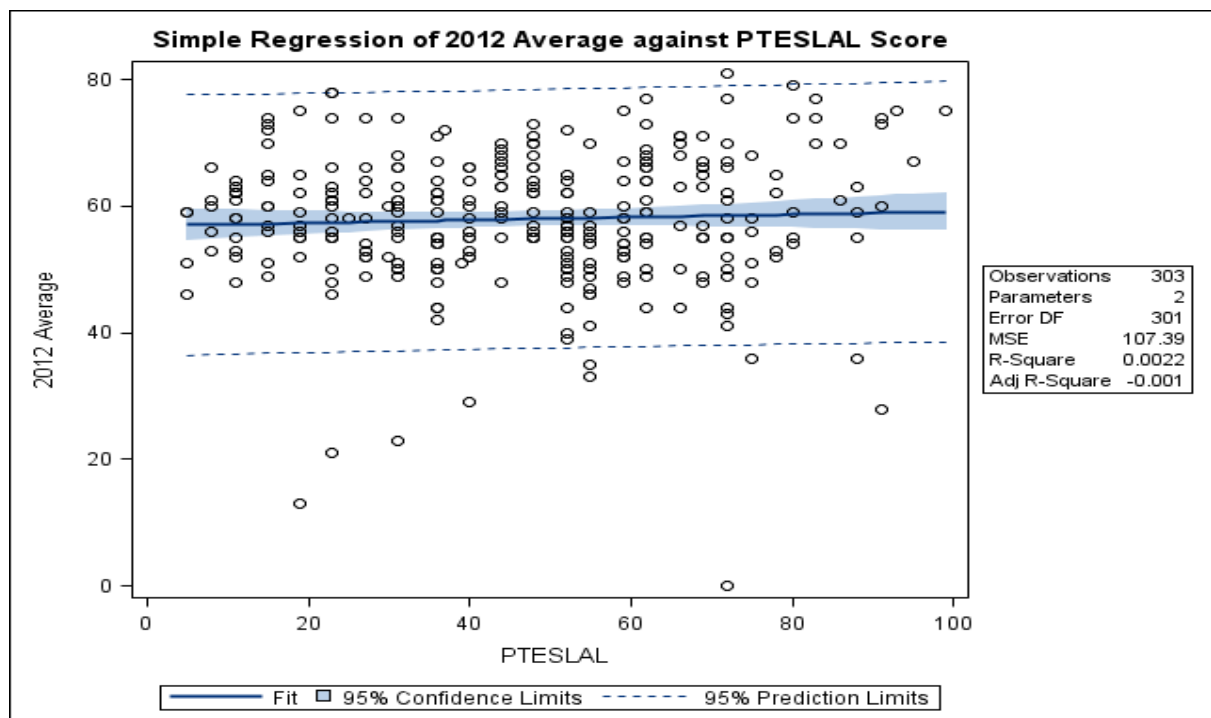
Lastly, the  $t$  statistic for PTESLAL equalled 0.82, and was not statistically significant ( $p=0.4128$ ). This means that PTESLAL had a positive but statistically not significant relationship with the participants' 2012 average scores. This is captured in the last two columns of **Table 21** below.

**Table 21: The  $t$  statistic for PTESLAL as a predictor of 2012 average (n=303)**

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	T Value	P Value
Intercept	Intercept	1	56.92521	1.40291	40.58	<.0001
PTESLAL	PTESLAL score	1	0.02234	0.02725	0.82	0.4128

The fit plot for the linear regression of 2012 average on PTESLAL is presented in **Figure 4** below.

**Figure 4: The fit plot for the results of a linear regression of 2012 average on PTESLAL (n=303)**





It is evident from both the slope of the regression line and the location of the data points in **Figure 4** above that PTESLAL had a marginally positive and extremely weak relationship with the predicted variable.

The overall meaning of these results is that on its own, PTESLAL was an even weaker predictor of 2012 average than NBT AL. This is evident in the lower  $t$  statistic and its higher than .05  $p$ -value for the test. The fit plot for the linear regression of 2012 average on the test further confirms that it had a very weak and almost parallel relationship with 2012 average. This means therefore that at linear regression analysis level, PTESLAL was a very weak predictor of the outcome variable.

#### ***4.3.3.3 Grade 12 English as a predictor of 2012 average***

The third linear regression on the first data set – NBT AL, PTESLAL and Grade 12 English as predictors of 2012 average – involved Grade 12 English as a predictor of 2012 average. The  $F$  statistic for this model equalled  $F(1,221) = 14.28$ ,  $p = 0.0002$ . This indicates that at least one of the predictors constituting this model had a positive relationship with the predicted variable. The  $F$  statistic and its  $p$ -value from this analysis are captured in the last two columns of **Table 22** below.

**Table 22: The F statistic for the model including Grade 12 English as the specified predictor of 2012 average (n=223)**

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1438.09210	1438.09210	14.28	0.0002
Error	221	22256	100.70417		
Corrected Total	222	23694			

Secondly, the R-Square from the analysis equalled 0.0607, meaning that the model could explain 6.1% of the variance in the participants' average scores. The R-Square resulting from this analysis is captured in the last column of the first row of **Table 23** below.

**Table 23: The R-Square for the model involving Grade 12 English as the specified predictor of 2012 average (n=223)**

Root MSE	10.03515	R-Square	0.0607
Dependent Mean	57.96413	Adj R-Sq	0.0564
Coeff Var	17.31269		

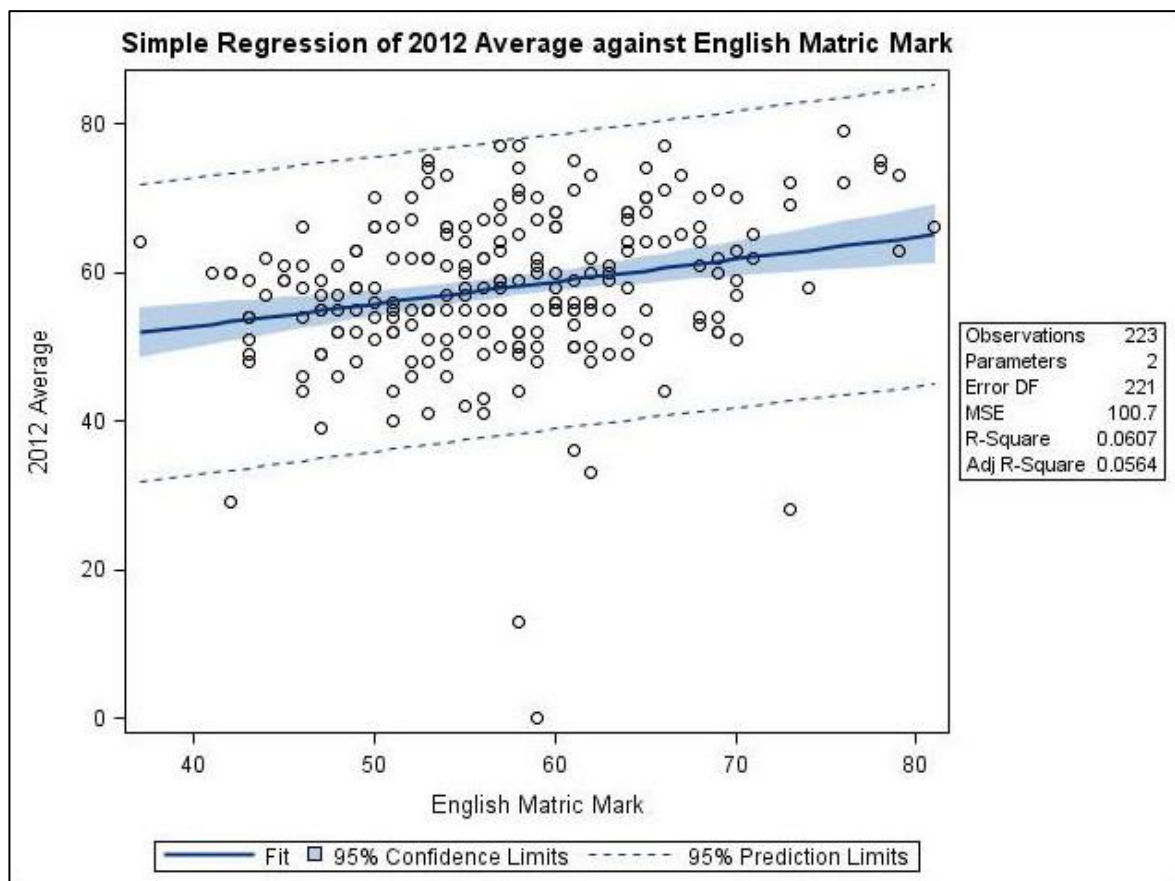
Lastly, the *t* statistic for Grade 12 English equalled 3.78 and was statistically significant ( $p=0.0002$ ). This means that considered alone, Grade 12 English had a positive relationship with the participants' 2012 average scores. The *t* statistic and *p* value for performance on this examination are presented in the last two columns of **Table 24** below.

**Table 24: The t statistic for Grade 12 English as the predictor of 2012 average performance (n=223)**

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	40.69853	4.61806	8.81	<.0001
Grade 12 English Examination	Grade 12 English score	1	0.30106	0.07967	3.78	0.0002

The results of a linear regression of 2012 average performance on Grade 12 English are graphically depicted in the fit plot in **Figure 5** below.

**Figure 5: The fit plot for the results of a linear regression for Grade 12 English as the predictor of 2012 average (n=223)**



As can be seen in **Figure 5** above, the slope of the regression line and the spread of the data points around it indicate that Grade 12 English had a positive and stronger relationship with the predicted variable than the other two predictors.

Overall, these results mean that on its own, Grade 12 English was a significant predictor of 2012 average. This is evident in the higher and statistically significant  $t$  statistic for this assessment. The fit plot for the linear regression involving Grade 12 English as the predictor confirms that at the level of linear regression, this assessment was a better predictor of 2012 average than NBT AL and PTESLAL.

#### **4.3.4 Multiple regression analysis: NBT AL, PTESLAL, Grade 12 English and 2012 Average**

##### ***4.3.4.1 NBT AL and Grade 12 English as predictors of 2012 average***

Following the three linear regressions on the first data set, – NBT AL, PTESLAL and Grade 12 English scores as predictors of 2012 average – multiple regression analyses of 2012 average on NBT AL and Grade 12 English on the one hand and PTESLAL and Grade 12 English on the other were computed. These permutations were chosen to determine if NBT AL and PTESLAL would add significantly to the predictive information provided by Grade 12 English, the predictor already in use and one with evidently a stronger ability to predict the current group of participants' academic success so far. Cohen and Swerdlik (2010: 184) have indeed argued that an incremental validity study starts with establishing the best predictor and then using multiple regression to “examine the usefulness of other predictors”. For the purpose of investigating if and which of the additional predictors in the first data set was incrementally useful, the first multiple regression analysis focused on NBT AL and Grade 12 English as predictors. The  $F$

statistic from this analyses equalled  $F(2,220) = 7.29$ ,  $p = 0.0009$ . This means that at least one of the predictors in this model had a positive relationship with the predicted variable. The  $F$  statistic and its  $p$  value for the model are presented in the last two columns of **Table 25** below.

**Table 25: The F statistic for the model with NBT AL and Grade 12 English as specified predictors of 2012 average (n=223)**

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1473.43831	736.71915	7.29	0.0009
Error	220	22220	101.00125		
Corrected Total	222	23694			

Secondly, the R-Square from this analysis was 0.0622, meaning that 6.2% of the variance in the participants' 2012 average could be accounted for by the model with NBT AL and Grade 12 English as specified predictors. The R-Square for this model is captured in the last column of the first row of **Table 26** below.

**Table 26: The R-Square for the model with NBT AL and Grade 12 English as specified predictors of 2012 average (n=223)**

Root MSE	10.04994	R-Square	0.0622
Dependent Mean	57.96413	Adj R-Sq	0.0537
Coeff Var	17.33820		

Lastly, the  $t$  statistic for NBT AL equalled -0.59 and was not statistically significant ( $p=0.5547$ ) while the  $t$  statistic for Grade 12 English was 3.57 and was statistically

significant ( $p=0.0004$ ). This means that when regressed on both these variables at the same time, 2012 average had a positive relationship with Grade 12 English and a slightly negative relationship with NBT AL. The  $t$  statistics and their  $p$  values for the two predictors are captured in the last two columns of **Table 27** below.

**Table 27: The  $t$  statistics for NBT AL and Grade 12 English as predictors of 2012 Average ( $n=223$ )**

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	T Value	Pr >  t
<b>Intercept</b>	Intercept	1	41.14487	4.68600	8.78	<.0001
<b>NBT AL</b>	NBT AL score	1	-0.04741	0.08015	-0.59	0.5547
<b>Grade 12 English examination</b>	Grade 12 English score	1	0.32822	0.09205	3.57	0.0004

#### **4.3.4.2 PTESLAL and Grade 12 English as predictors of 2012 average**

The next multiple regression of the first data set – NBT AL, PTESAL and Grade 12 English as predictors of 2012 average – involved PTESLAL and Grade 12 English as predictors of 2012 average. The  $F$  statistic from this analysis was  $F(2,216) = 7.72$ ,  $p = 0.0006$ , meaning that at least one of the predictors in this model had a positive relationship with the predicted variable. The  $F$  statistic from this analysis is captured in the last two columns of **Table 28** below.

**Table 28: The F statistic for the model with PTESLAL and Grade 12 English as specified predictors of 2012 average (n=219)**

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1574.95603	787.47802	7.72	0.0006
Error	216	22029	101.98420		
Corrected Total	218	23604			

The R-Square from the analysis was 0.0667, meaning that 6.7% of the variance in the 2012 average performance could be accounted for by the predictor model involved. This R-Square is captured in the first row of the last column of **Table 29** below.

**Table 29: The R-Square for the model with PTESLAL and Grade 12 English as specified predictors of 2012 average (n=219)**

Root MSE	10.09872	R-Square	0.0667
Dependent Mean	58.04566	Adj R-Sq	0.0581
Coeff Var	17.39789		

Lastly, the *t* statistic for PTESLAL from this analysis equalled -1.04 and was not statistically significant ( $p=0.3018$ ) while the *t* statistic for Grade 12 English equalled 3.92 and was statistically significant ( $p=0.0001$ ). This means that when simultaneously regressed on the two variables, 2012 average had a positive relationship with Grade 12 English and that its relationship with PTESLAL was

negative. The  $t$  statistics and their  $p$  values for the two predictors are captured in the last two columns of **Table 30** below.

**Table 30: The  $t$  statistics for PTESLAL and Grade 12 English as predictors of 2012 average (n=219)**

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	T Value	Pr >  t
Intercept	Intercept	1	40.50606	4.67715	8.66	<.0001
PTESLAL	PTESLAL score	1	-0.03529	0.03409	-1.04	0.3018
Grade 12 English examination	Grade 12 English score	1	0.33398	0.08513	3.92	0.0001

The last multiple regression analysis of the first data set – NBT AL, PTESLAL and Grade 12 English scores as predictors of 2012 average – involved all the three tests as predictors of 2012 average. The reason for including all predictors in the analysis was to determine whether NBT AL and PTESLAL combined would add significantly to the predictive ability of Grade 12 English. The  $F$  statistic resulting from this analysis was  $F(3,215) = 5.14$ ,  $p = 0.0019$ . This means that at least one of the predictors in this model had a positive relationship with the predicted variable. The  $F$  statistic and its  $p$  value for the model are captured in the last two columns of **Table 31** below.



**Table 31: The F statistic for the model with NBT AL, PTESLAL and Grade 12 English as specified predictors of 2012 average (n=219)**

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1579.00953	526.33651	5.14	0.0019
Error	215	22025	102.43969		
Corrected Total	218	23604			

The R-Square generated by this analysis equalled 0.0669, meaning that 6.7% of the variance in the participants' 2012 average could be accounted for by the model involving the three predictors combined. The R-Square for this model is shown in the last column of the first row of **Table 32** below.

**Table 32: The R-Square for the model with NBT AL, PTESLAL and Grade 12 English as specified predictors of 2012 average (219)**

Root MSE	10.12125	R-Square	0.0669
Dependent Mean	58.04566	Adj R-Sq	0.0539
Coeff Var	17.43670		

Lastly, the *t* statistic for NBT AL equalled -0.20 and was not statistically significant ( $p=0.8425$ ), the same statistic for PTESLAL was -0.85 and was also not statistically significant ( $p=0.3959$ ) and the same statistic for Grade 12 English equalled 3.64 and was statistically significant ( $p=0.0003$ ). This means that regressed simultaneously on the three variables, 2012 average performance had a positive relationship with Grade 12 English while the other two tests had a negative

relationship with it. The  $t$  statistics and their  $p$  values for the three predictors are captured in the last two columns of **Table 33** below.

**Table 33: The  $t$  statistics for NBT AL, PTESLAL and Grade 12 English as predictors of 2012 average**

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	T Value	Pr >  t
<b>Intercept</b>	Intercept	1	40.67189	4.76114	8.54	<.0001
<b>NBT AL</b>	NBT AL score	1	-0.01787	0.08983	-0.20	0.8425
<b>PTESLAL</b>	PTESLAL score	1	-0.03210	0.03774	-0.85	0.3959
<b>Grade 12 English examination</b>	Grade 12 English score	1	0.34172	0.09377	3.64	0.0003

The results of the multiple regression analyses of 2012 average on NBT AL, PTESLAL and Grade 12 English for all the permutations used mean that Grade 12 English had a strong predictive relationship with the average while NBT AL and PTESLAL were slightly negative predictors of this average. These results mean that the latter tests could not predict 2012 average better than Grade 12 English. Having established this, the next question was whether NBT AL and PTESLAL scores possessed any degree of incremental validity in relation to Grade 12 English scores, the best predictor of academic success in the first set of data. Haynes and Lench (2003: 462) define the index of incremental validity as the additional difference that a predictor variable makes to the R-Square of the most efficient predictor among those used. Haynes and Lench (2003: 463) also observe, however, that the adjusted R-Square provides “a more robust estimate [of the incremental validity index] ... to reflect sample size and the number of predictor variables”.

From the results of linear and multiple regressions of 2012 average on NBT AL, PTESLAL and Grade 12 English, it can be seen that the adjusted R-Square for Grade 12 English decreases from 5.64% in a linear regression to 5.37% in a multiple regression with NBT AL. In contrast, the same adjusted R-Square increases to 5.81% in a multiple regression with PTESLAL. Finally, the same adjusted R-square decreases to 5.34% in a multiple regression involving the three predictors. This suggests that on its own, NBT AL added 0% and PTESLAL 0.2% respectively to the ability of Grade 12 English to predict 2012 average. This means that PTESLAL slightly added to the predictive validity of Grade 12 English and that NBT AL did not. Furthermore, in a multiple regression involving the three assessments as predictors of 2012 average, the adjusted R-Square for Grade 12 English decreases from its initial 5.64% to 5.34%. This means that considered simultaneously, NBT AL and PTESLAL did not add to the predictive ability of Grade 12 English and that they therefore did not possess incremental validity in relation to the best predictor in the model (SCU 2015).

#### **4.4 TALL, PTESLAL, Grade 12 English and 2014 average**

##### **4.4.1 Descriptive statistics**

The second data set to be analysed comprised the participants' scores on TALL, PTESLAL and Grade12 English as predictors of their 2014 averages. The descriptive statistics for the scores obtained on all these variables are presented in **Table 34** below.

**Table 34: The means and standard deviations for the scores on TALL, PTESLAL, Grade 12 English and 2014 average**

Variable	N	Mean	SD	Minimum	Maximum
<b>2014 Average</b>	98	56.80	14.10	6	90
<b>TALL</b>	102	41.91	14.45	11	76
<b>PTESLAL</b>	102	51.07	19.99	3	97
<b>GR 12 English</b>	82	60.08	9.48	41	86

As can be seen in **Table 34** above, the mean score for the participants was the lowest on TALL (M=41.91), higher on PTESLAL (M=51.07), still higher on 2014 average (M=56.80) and the highest on Grade 12 English (M=60.08). In the first place, this shows that TALL and PTESLAL were probably more challenging than Grade 12 English for this group of participants. Once again, this should be expected for these tests, because the purpose for which they are used associates them more with the demands of higher education than Grade 12 English. In the second place, the higher mean score on Grade 12 English for the participants shows that this assessment was less challenging for them compared to TALL and PTESLAL. This should also be expected because, as pointed out earlier, this assessment is not necessarily designed to measure readiness for university education in respect of language. Also, the small difference between the mean scores on Grade 12 English and 2014 average shows that performance on the former was at first glance slightly more aligned to it than it was to TALL and PTESLAL. This was not expected, however, because, as also pointed out earlier, the two tests have mainly been used to determine academic readiness and should

therefore be expected to resonate more with academic performance than Grade 12 English. It can also be seen in **Table 34** that the standard deviation was the highest for PTESLAL (SD=19.99), higher for both 2014 average and TALL (SD=14.10 and SD=14.45 respectively) and the lowest for Grade 12 English (SD=9.49). Once again, the lowest standard deviation of the scores on Grade 12 English when compared to the other predictors should be expected and can be accounted for by the kind of information that this assessment aims to provide. As indicated earlier, Grade 12 English is an achievement assessment while TALL and PTESLAL are geared towards providing mainly placement information for academic study. As also pointed out earlier, the latter are inherently more expected to spread scores out while an achievement assessment such as Grade 12 English is not.

#### 4.4.2 Intercorrelations: TALL, PTESLAL, Grade 12 English and 2014 average

The correlations of the scores obtained by the participants on the second data set – PTESLAL, TALL and Grade 12 English as predictors of 2014 average – are presented in **Table 35** below.

**Table 35: Intercorrelations for TALL, PTESLAL, Grade 12 English and 2014 average**

Variable	N	2014 average	TALL	PTESLAL	Gr 12 English examination
2012 Average	98	--	.30*	.33*	.47*
TALL	98		--	.68*	.35*
PTESLAL	98			--	.50*
Gr 12 English examination	78				--

\* $p < .05$

As can be seen in **Table 35** above, the correlation with 2014 average was the highest for Grade 12 English ( $r = .47$ ), higher for PTESLAL ( $r = .33$ ) and slightly lower for TALL ( $r = .30$ ). All these correlations were statistically significant ( $p < .05$ ). The positive and statistically significant correlations between the predictor and outcome variables mean that to varying degrees, participants who performed well on any of them tended to do the same on the others and vice versa. The moderately high correlation for Grade 12 English and 2014 average means that the former associated with the latter better than the other two tests and that Grade 12 English will very likely predict 2014 average better than TALL and PTESLAL. Once again, given the purpose for which PTESLAL and TALL are used, however, one would expect that scores on the two tests correlate better with academic performance than Grade 12 English. As can also be seen in **Table 35** above, the correlations among the predictors were the highest for TALL and PTESLAL ( $r = .68$ ), higher for Grade 12 English and PTESLAL ( $r = .50$ ) and low for TALL and Grade 12 English ( $r = .35$ ). These correlations were also all statistically significant ( $p < .05$ ). From this, it seemed likely that TALL would provide incremental predictive information because of its positive and statistically significant correlation with 2014 average and lowest correlation with Grade 12 English, the predictor variable that associated the highest with the predicted variable in this case. Additional predictors that correlate less with the best predictor but correlate significantly with the criterion variable are likely sources of incremental validation. In the words of Salkind (2011: 280), for an incremental validity study, one wants “only independent or predictor variables that are related to the dependent variable and are unrelated to each other. That way, each one makes as unique a contribution

as possible in predicting the dependent or predicted variable.” Furthermore, the correlation between TALL and PTESLAL was high enough to constitute what is known as multicollinearity, a condition that violates the most important conditions of multiple regression and by extension, compromises the potential for the predictors involved to add differentially to the predictive efficiency of the best predictor (Morgan et al. 2011: 141). This means that the two tests were not likely to provide different additional predictive information because of the evidently high overlap in what they appeared to measure in the case of the participants involved. In the words of Haynes and Lench (2003: 462), “the degree of collinearity among predictor variables suggests the degree of overlap and amount of independent information in each ...” and “shows the likely increment in predictive efficacy that would occur if the two variables were combined versus if each variable were used as an independent predictor”.

#### **4.4.3 Linear regression analyses: TALL, PTESLAL, Grade 12 English and 2014 average**

##### ***4.4.3.1 TALL as a predictor of 2014 average***

Following the computation of correlations among the variables involved in the second data set – TALL, PTESLAL and Grade 12 English scores as predictors of 2014 average – a simple linear regression of 2014 average was carried out on each of the three assessments. The first of these analyses focused on TALL. Firstly, the  $F$  statistic from this analysis equalled  $F(1,96) = 9.61, p = 0.0025$ . This means that at least one of the predictors involved in this model had a positive relationship with the predicted variable. The  $F$  statistic and its  $p$  value from the analysis are captured in the last two columns of **Table 36** below.

**Table 36: The F statistic for the model with TALL as the specified predictor of 2014 average (n=98)**

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1754.87553	1954.87553	9.61	0.0025
Error	96	17528	182.58792		
Corrected Total	97	19283			

Secondly, the R-Square from the analysis equalled 0.0910, meaning that 9.1% of the variance in the participants' average scores could be explained by the model.

The R-Square from this analysis is captured in the last column of **Table 37** below.

**Table 37: The R-Square for the model with TALL as the specified predictor of 2014 average (n=98)**

Root MSE	13.51251	R-Square	0.0910
Dependent Mean	56.80612	Adj R-Sq	0.0815
Coeff Var	23.78707		

Furthermore, the *t* statistic for TALL equalled 3.10 and was statistically significant ( $p=0.0025$ ). This indicates that on its own, TALL had a positive relationship with 2014 average performance. Both the *t* statistic and its *p* value for this test are captured in the last two columns of **Table 38** below.

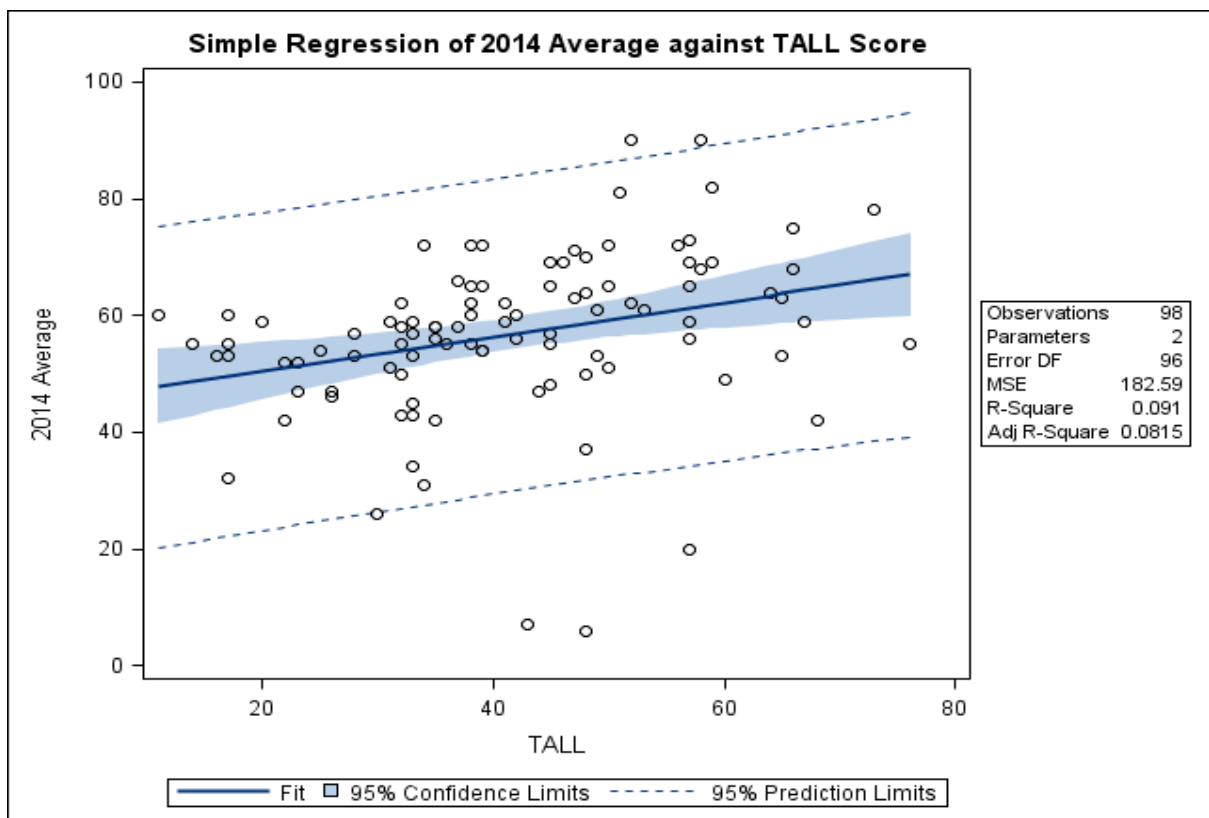


**Table 38: The t statistic for TALL as the predictor of 2014 average performance (n=98)**

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	T Value	Pr >  t
Intercept	Intercept	1	44.48881	4.20102	10.59	<.0001
TALL	TALL score	1	0.29571	0.09539	3.10	0.0025

The results of a linear regression of 2014 average on TALL are graphically captured in the fit plot in **Figure 6** below.

**Figure 6: The fit plot for the results of a linear regression of 2014 average on TALL (n=98)**



In **Figure 6** above, the slope of the regression line and the location of the data points around it confirm that TALL had a positive but moderate relationship with 2014 average performance.

The overall meaning of these results is that considered alone, TALL was a significant predictor of the participants' average performance at the end of 2014. This is evident in the positive and statistically significant  $t$  statistic value from the linear regression analysis involving this test as the predictor. This is further confirmed by the fit plot for the two variables. The slope of the regression line and the location of the data points in relation to it in this plot mean that TALL was a moderate but significant predictor of 2014 average.

#### 4.4.3.2 PTESLAL as a predictor of 2014 average

The next linear regression of 2014 average involved PTESLAL as the predictor. The  $F$  statistic from this analysis was  $F(1,96) = 11.44$ ,  $p = 0.001$ . This means that at least one of the predictors in the model had a positive relationship with the predicted variable. See the last two columns of **Table 39** below for this result.

**Table 39: The F statistic for the model with PTESLAL as the specified predictor of 2014 average (n=98)**

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2053.97004	2053.97004	11.44	0.0010
Error	96	17229	179.47236		
Corrected Total	97	19283			

The R-Square from this analysis was 0.1065, meaning that the predictor model could account for 10.7% of the variance in the participants' average scores at the end of 2014. See the last column of the first row of **Table 40** for this result.

**Table 40: The R-Square for the model with PTESLAL as the specified predictor of 2014 average (n=98)**

<b>Root MSE</b>	13.39673	<b>R-Square</b>	0.1065
<b>Dependent Mean</b>	56.80612	<b>Adj R-Sq</b>	0.0972
<b>Coeff Var</b>	23.58325		

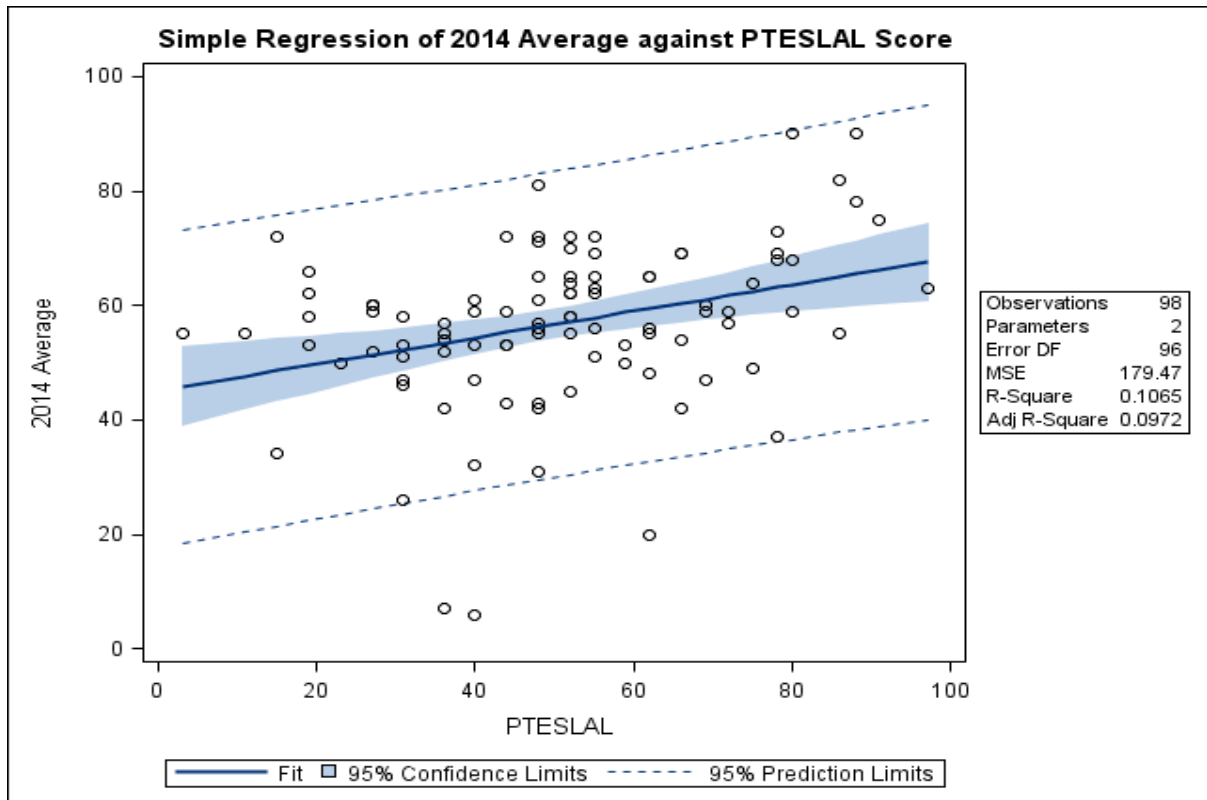
Furthermore, the  $t$  statistic for PTESLAL equalled 3.38 and was statistically significant ( $p=0.001$ ). This means that on its own, PTESLAL was a significant predictor of the participants' 2014 average scores in their programmes of study. The  $t$  and  $p$  values for this test are presented in the last two columns of **Table 41** below.

**Table 41: The  $t$  statistic for PTESLAL as a predictor of 2014 average (n=98)**

<b>Parameter Estimates</b>						
<b>Variable</b>	<b>Label</b>	<b>DF</b>	<b>Parameter Estimate</b>	<b>Standard Error</b>	<b>T Value</b>	<b>Pr &gt;  t </b>
<b>Intercept</b>	Intercept	1	45.13544	3.70577	12.18	<.0001
<b>PTESLAL</b>	PTESLAL score	1	0.23096	0.06827	3.38	0.0010

The results of a linear regression of the participants' 2014 average on their performance on PTESLAL are graphically captured in the fit plot in **Figure 7** below.

Figure 7: The fit plot for PTESLAL as the predictor of 2014 average performance (n=98)



It is clear from the slope of the regression line and the location of the data points towards it in **Figure 7** above that performance on PTESLAL had a positive and moderate relationship with 2014 average performance.

The results of a linear regression of 2014 average on PTESLAL mean that on its own, PTESLAL was also a significant predictor of the former. This is also evident in the statistically significant  $t$  statistic and its  $p$ -value for this test from this analysis. The slope of the regression line in the fit plot for the linear regression involving the two variables confirm that like TALL, PTESLAL was also a moderate predictor of 2014 average performance. While the  $t$  statistic for PTESLAL was higher than that of TALL, this difference was not significantly

large. Worth pointing out in relation to these results, however, is that as a predictor of 2012 average in the first data set already analysed in this study, neither did PTESLAL possess predictive validity nor add to the ability of the best predictor to predict this average. The probable indication of these inconsistent results is that performance on the test was sample dependent. This implies that the test was not consistent in its assessment of what it purported to measure for the two sample groups used. In a way, this raises questions about the reliability of the test. This appears to be the case when one considers that Grade 12 English, another predictor used for the same samples was consistently the best predictor on both these occasions. Worth pointing out in this regard also is that PTESLAL showed evidence of predictive validity in a linear regression involving the second data set where the sample size was smaller (n=98) but failed to do so with the first data set where the sample size was larger (n=303). The difference in the two sample sizes cannot solely be responsible for the inconsistent predictive information yielded by the test in the two linear regression analyses in which it was involved, however. In the context of second language research in particular, Fraenken and Wallen (2003) have set the minimum sample size for correlational studies such as the present one at 50 participants. Mackey and Gass (2005: 124) have also observed that “in second language studies, small groups are sometime appropriate as long as the techniques for analysis take the numbers into account.”

#### *4.4.3.3 Grade 12 English as a predictor of 2014 average*

The third linear regression on the second data set – TALL, PTESLAL and Grade 12 English scores as predictors of 2014 average – was carried out on Grade 12 English

as the specified predictor of 2014 average. The  $F$  statistic yielded by this analysis equalled  $F(1,76) = 21.71$ ,  $p = <.0001$ . This means that at least one of the predictors in this model had a positive relationship with the predicted variable. The  $F$  statistic and its  $p$  value for the model are captured in the last two columns of **Table 42** below.

**Table 42: The F statistic for the model with Grade 12 English as the specified predictor of 2014 average (n=78)**

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3761.36633	3761.36633	21.71	<.0001
Error	76	13166	173.23793		
Corrected Total	77	16927			

Secondly, the R-Square from this analysis equalled 0.2222, meaning that 22% of the variance in the participants' 2014 average scores could be accounted for by this predictor model. The R-Square for the model is captured in the last column of **Table 43** below.

**Table 43: The R-Square for the model with Grade 12 English as the specified predictor of 2014 average (n=78)**

Root MSE	13.16199	R-Square	0.2222
Dependent Mean	55.47436	Adj R-Sq	0.2120
Coeff Var	23.72626		

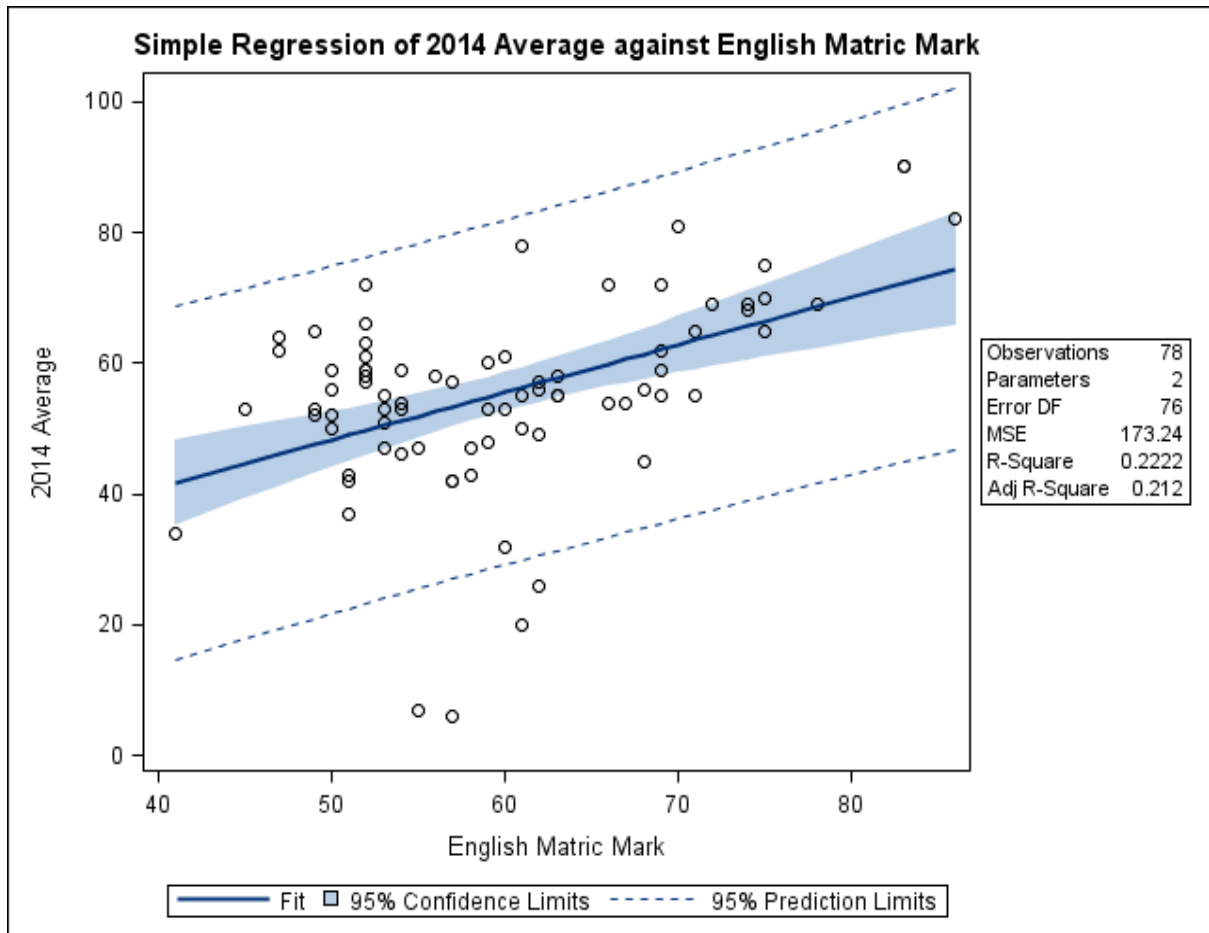
Moreover, the  $t$  statistic for Grade 12 English in this case was 4.66 and was statistically significant ( $p < .0001$ ). This means that on its own, Grade 12 English had a positive relationship with the participants' average performance at the end of 2014. The  $t$  statistic and its  $p$  value for the performance on this examination are captured in the last two columns of **Table 44** below.

**Table 44: The  $t$  statistic for Grade 12 English as predictor of 2014 average (n=78)**

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	T Value	Pr >  t
<b>Intercept</b>	Intercept	1	11.97216	9.45418	1.27	0.2093
<b>Grade 12 English</b>	Grade 12 English score	1	0.72659	0.15593	4.66	<.0001

The results of a linear regression of 2014 average on Grade 12 English are graphically presented in the fit plot in **Figure 8** below.

**Figure 8: The fit plot for Grade 12 English as predictor of 2014 average performance (n=78)**



Both the slope of the regression line and the location of the data points in **Figure 8** above confirm that Grade 12 English had a positive and stronger relationship with the predicted variable when compared to TALL and PTESLAL.

Overall, these results mean that considered alone, Grade 12 English was a better predictor of 2014 average than TALL and PTESLAL. This is also evident in the higher *t* statistic for this assessment than those recorded for TALL and PTESLAL. The slope of the regression line in the fit plot for Grade 12 English and 2014 average further confirms that the former was the strongest predictor of the latter when compared to TALL and PTESLAL. Once again, this is an early indication



that if any of the three will end up being a better predictor of 2014 average in a simultaneous comparison at all, it will be Grade 12 English.

#### **4.4.4 Multiple regression analyses: TALL, PTESLAL, Grade 12 English and 2014 average**

##### **4.4.4.1 TALL and Grade 12 English as predictors of 2014 average**

Following the linear regression analyses of the second data set – TALL, PTESLAL and Grade 12 English as predictors of 2014 average – a multiple regression of 2014 average was carried out on TALL and Grade 12 English on the one hand and PTESLAL and Grade 12 English on the other. The reason for choosing these permutations is similar to the one given with regard to NBT AL, PTESLAL and Grade 12 English as predictors of 2012 average earlier; Grade 12 English is part of the Grade 12 examination, the assessment tool traditionally used as the predictor of academic performance by South African universities. Also, the correlation of this variable with 2014 average so far shows that it is likely to be the best predictor of this average in the end and that the two tests will possibly only add to the predictive information that it will provide.

The first multiple regression of 2014 average carried out on the current data set involved TALL and Grade 12 English as predictors. The  $F$  statistic from this analysis equalled  $F(2,75) = 11.68$ ,  $p = <.0001$ . This means that at least one of the predictors in this model had a positive relationship with the predicted variable. The  $F$  statistic and its  $p$  value for the model are captured in the last two columns of **Table 45** below.

**Table 45: The F statistic for the model with TALL and Grade 12 English as specified predictors of 2014 average (n=78)**

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4020.04816	2010.02408	11.68	<.0001
Error	75	12907	172.09867		
Corrected Total	77	16927			

The R-Square from this analysis equalled 0.2375, meaning that 24% of the variance in the participants' 2014 average performance could be accounted for by this predictor model. The R-Square for the model is captured in the last column of **Table 46** below.

**Table 46: The R-Square for the model with TALL and Grade 12 English as specified predictors of 2014 average (n=78)**

Root MSE	13.11864	R-Square	0.2375
Dependent Mean	55.47436	Adj R-Sq	0.2172
Coeff Var	23.64811		

Lastly, the *t* statistic for TALL was 1.23 and was not statistically significant ( $p=0.223$ ) while that for Grade 12 English was 3.89 and was statistically significant ( $p=0.0002$ ). This means that in relation to Grade 12 English, TALL did not have a positive association with the participants' 2014 average performance and that Grade 12 English did. The *t* statistics and their *p* values for the two assessments are presented in the last two columns of **Table 47** below.

**Table 47: The t statistics for TALL and Grade 12 English as predictors of 2014 average performance (n=78)**

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	T Value	Pr >  t
Intercept	Intercept	1	10.88846	9.46441	1.15	0.2536
TALL	TALL score	1	0.13613	0.11104	1.23	0.2240
Grade 12 English examination	Grade 12 English score	1	0.65056	0.16733	3.89	0.0002

#### 4.4.4.2 PTESLAL and Grade 12 English as predictors of 2014 average

The next multiple regression of 2014 average was carried out on PTESLAL and Grade 12 English. The  $F$  statistic from this analysis was  $F(2,75) = 12.40$ ,  $p = <.0001$ . This means that at least one of the predictors constituting this model had a positive relationship with the predicted variable. The  $F$  statistic for the model is captured in the last two columns of **Table 48** below.

**Table 48: The F statistic for the model with PTESLAL and Grade 12 English as specified predictors of 2014 average (n=78)**

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4205.68501	2102.84251	12.40	<.0001
Error	75	12722	169.62352		
Corrected Total	77	16927			

The R-Square for this model equalled 0.2485, meaning that 25% of the variance in the 2014 average performance could be accounted for by this model. See the last two columns of **Table 49** below for this result.

**Table 49: The R-Square for the model with PTESLAL and Grade 12 English as specified predictors of 2014 average (n=78)**

<b>Root MSE</b>	13.02396	<b>R-Square</b>	0.2485
<b>Dependent Mean</b>	55.47436	<b>Adj R-Sq</b>	0.2284
<b>Coeff Var</b>	23.47744		

Thirdly, the  $t$  statistic for PTESLAL equalled 1.62 and was not statistically significant ( $p=0.1098$ ) while the one for Grade 12 English was 3.20 and was statistically significant ( $p=0.0020$ ). This means that in relation to Grade 12 English examination, PTESLAL did not have a positive relationship with the participants' 2014 average and that Grade 12 English did. The  $t$  statistics and their  $p$  values for the two predictors are captured in the last two columns of **Table 50** below.

**Table 50: The  $t$  statistics for PTESLAL and Grade 12 English as predictors of 2014 average performance (n=78)**

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	T Value	Pr >  t
<b>Intercept</b>	Intercept	1	13.32875	9.39251	1.42	0.1600
<b>PTESLAL</b>	PTESLAL score	1	0.15054	0.09301	1.62	0.1098
<b>Grade 12 English examination</b>	Grade 12 English score	12	0.57606	0.18016	3.20	0.0020

#### 4.4.4.3 TALL, PTESLAL and Grade 12 English as predictors of 2014

The last multiple regression carried out on the second data set – TALL, PTESLAL and Grade 12 English scores as predictors of 2014 average – involving pairs of predictors was followed by one in which all the three predictors were all simultaneously involved. Firstly, the  $F$  statistic from this analysis was  $F(3,74) = 8.19$ ,  $p = <.0001$ . This means that at least one of the predictors in this model had a positive relationship with the predicted variable. This is shown in the last two columns of **Table 51** below.

**Table 51: The F statistic for the model with TALL, PTESLAL and Grade 12 English as specified predictors of 2014 average performance (n=78)**

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
<b>Model</b>	3	4220.81809	1406.93936	8.19	<.0001
<b>Error</b>	74	12707	171.71122		
<b>Corrected Total</b>	77	16927			

Secondly, the R-Square from this analysis equalled 0.2493, meaning that 25% of the variance in the participants' 2014 average performance could be accounted for by this predictor model. This is shown in the last column of **Table 52** below.

**Table 52: The R-Square for the model with TALL, PTESLAL and Grade 12 English as specified predictors of 2014 average (n=78)**

<b>Root MSE</b>	13.10386	<b>R-Square</b>	0.2493
<b>Dependent Mean</b>	55.47436	<b>Adj R-Sq</b>	0.2189
<b>Coeff Var</b>	23.62148		

Lastly, the  $t$  statistic for TALL was 0.30, the one for PTESLAL equalled 1.08 and the one for Grade 12 English was 3.17. The  $t$  statistics for TALL and PTESLAL were not statistically significant ( $p=0.7674$  and  $p=0.2831$  respectively) while the one for Grade 12 English was statistically significant ( $p=0.0022$ ). This indicates that considered simultaneously, Grade 12 English had a positive relationship with the participants' 2014 average performance while TALL and PTESLAL did not. The  $t$  statistics and  $p$  values for the three predictors are captured in the last two columns of **Table 53** below.

**Table 53: The  $t$  statistics for TALL, PTESLAL and Grade 12 English as predictors of 2014 average (n=78)**

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	T Value	Pr >  t
<b>Intercept</b>	Intercept	1	12.79862	9.61738	1.33	0.1873
<b>TALL</b>	TALL score	1	0.04188	0.14107	0.30	0.7674
<b>PTESLAL</b>	PTESLAL score	1	0.12870	0.11902	1.08	0.2831
<b>Grade 12 English</b>	Grade 12 English score	1	0.57450	0.18134	3.17	0.0022

The results of multiple regressions involving TALL and Grade 12 English on the one hand and PTESLAL and Grade 12 English as predictors of 2014 average on the other show that in both these cases, Grade 12 English was the only significant predictor of 2014 average and that the two tests were not. The same was the case in a multiple regression involving the three assessments as predictors of 2014 average. Grade 12 English remained the only significant and therefore better predictor of this average than the two tests. On the flip side, the statistically insignificant  $t$

statistics for TALL and PTESLAL mean that they were not better predictors of 2014 average alongside and in comparison to Grade 12 English. From this analysis, it is also evident that the adjusted R-Square for the best predictor increases from 21,2% in a simple linear regression to 21.72% when paired with TALL in a multiple regression analysis while the same value for this predictor increases to 22.84% when paired with PTESLAL in the same analysis. This means that while their *t* statistics were not significant, the two tests were able to add 1% and 2% respectively to the degree to which Grade 12 English could predict 2012 average. This means that to different degrees, the two tests slightly added to the predictive efficiency of the best predictor and that they therefore had incremental validity (SCU 2015). Furthermore, when considered simultaneously with TALL and PTESLAL, the adjusted R-Square for Grade 12 English increases from 21.2% in a linear regression with 2014 average as the outcome variable to 21.89% in a multiple regression with the two tests as co-predictors of the same average. This means that combined, the two tests added 1% to the predictive efficiency of Grade 12 English and that they still had incremental validity when combined. It is worth noting in this case too that PTESLAL did not possess both predictive and incremental validity in the first data set but that it possessed a small degree of both from the analysis of the second data set. As pointed out earlier, this brings the reliability of this test into question. This is in contrast with Grade 12 English which showed evidence of consistently being the best predictor of academic performance from the analysis of both data sets.

## 4.5 TALL, Grade 12 English and 2014 average

### 4.5.1 Descriptive statistics

The third and last data set to be analysed involved TALL and Grade 12 English as predictors of 2014 average performance. As pointed out at the beginning of this chapter, the data set analysed in the previous section was part of the one to be analysed in this section. The descriptive statistics for the scores obtained by the participants on the three variables involved, namely, TALL and Grade 12 English as predictors of 2014 average are presented in **Table 54** below.

**Table 54: The means and standard deviations of the scores on TALL, Grade 12 English and 2014 average**

Variable	N	Mean	SD	Minimum	Maximum
2014 Average	603	55.84	12.05	3	90
TALL	636	40.62	15.72	9	95
Grade 12 English	483	60.62	8.33	35	86

As can be seen from **Table 54** above, the mean score for this group of participants was the highest on Grade 12 English (M=60.62), still high on their 2014 average (M=55.84) and low on TALL (M=40.62). This shows that TALL was the most challenging of the two predictors for the group of participants involved. Once again, this should be expected in view of what the test purports to measure especially in a context of reportedly low levels of academic literacy among high school leavers entering South African universities in recent years (cf. Boughey 2013; Van Dyk & Van de Poel 2013; Butler 2013; Van Dyk 2015). The small difference between performance on Grade 12 English and 2014 average means that 2014 average was more aligned with Grade 12 English scores than it was with



performance on TALL. This should not be the case, however, because TALL is aimed at measuring the ability to cope with the kind of discourse that students are expected to engage with at university. Furthermore, the standard deviation of the scores was the highest for TALL (SD=15.72), higher for 2014 average (SD=12.05) and the lowest for Grade 12 English (SD=8.33). The possible reason for the differences in the standard deviations of the predictors used in this study was provided earlier.

#### 4.5.2 Intercorrelations: TALL, Grade 12 English and 2014 average

Following the computation of the descriptive statistics on the last data set – TALL and Grade 12 English as predictors of 2014 average – correlations among the three variables were also calculated. The results of this analysis are presented in **Table 55** below.

**Table 55: Intercorrelations for TALL, Grade 12 English and 2014 average**

Variable	2014 average	TALL	Grade 12 English
2014 average	--	.24*	.31*
TALL		--	.35*
Grade 12 English			--

\* $p < .05$

As can be seen in **Table 55** above, the correlation with 2014 average was the highest for Grade 12 English ( $r = .31$ ) and lower for TALL ( $r = .24$ ). Both these correlation coefficients were statistically significant ( $p = < .0001$ ). This means that participants who performed well on any of the three assessments tended to do the same on the others and vice versa. Furthermore, the higher correlation for Grade 12

English and 2014 average is an early indication that if any of the predictor variables will predict academic performance better at all it will be Grade 12 English. Once again, for reasons already given, the expectation is that a test like TALL would correlate more with academic performance than Grade 12 English. It is also evident in **Table 55** that the correlation between TALL and Grade 12 English was  $r = .35$  and that it was statistically significant ( $p = <.0001$ ). This means that TALL stood less chance of predicting the 2014 average better than Grade 12 English. The positive but low degree of correlation between both predictors shows that TALL had a chance of having incremental instead of a higher predictive validity than Grade 12 English. In other words, the degree of collinearity between the two predictors did not pose any threat to TALL's potential to add to the predictive efficiency of Grade 12 English.

#### **4.5.3 Linear regression analysis: TALL, Grade 12 English and 2014 average**

##### **4.5.3.1 TALL as a predictor of 2014 average**

After the correlations among TALL, Grade 12 English and 2014 average were computed, a linear regression of the predicted variable on each of the two predictors was carried out. The first of these analyses involved TALL as a predictor of 2014 average. The  $F$  statistic from this analysis was  $F(1,602) = 38.03$ ,  $p = <.001$ . This indicates that at least one of the predictors in this model had a positive relationship with the predicted variable. The  $F$  statistic and its  $p$  value for the model are reflected in the first row of the last two columns of **Table 56** below.

**Table 56: The F statistic for the model with TALL as the specified predictor of 2014 average (n=604)**

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5194.57753	5194.57753	38.03	<.0001
Error	602	82234	136.60052		
Corrected Total	603	87428			

The R-Square from the analysis equalled 0.0594, meaning that 6% of the variance in the participants' 2014 average performance could be explained by this predictor model. This is captured in the last two columns in **Table 57** below.

**Table 57: The R-Square for the model with TALL as the specified predictor of 2012 average (n=604)**

Root MSE	11.68762	R-Square	0.0594
Dependent Mean	55.82781	Adj R-Sq	0.0579
Coeff Var	20.93512		

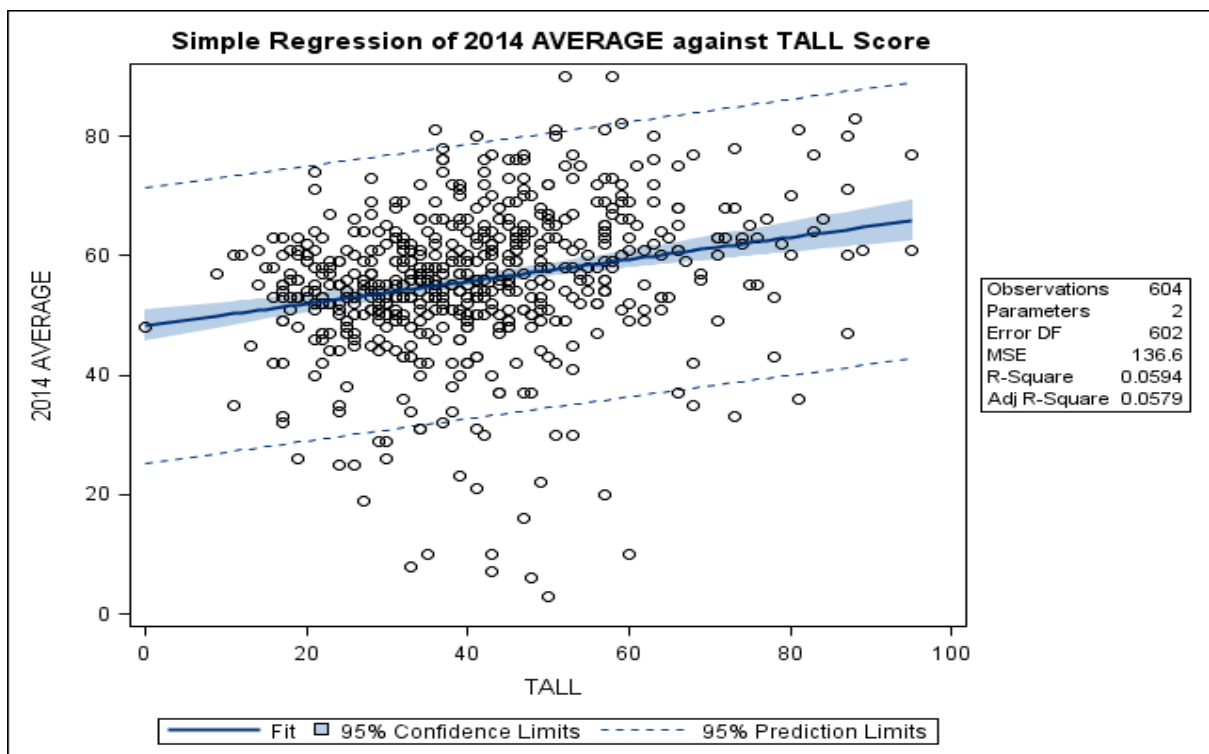
Lastly, the *t* statistic for TALL equalled 6.17 and was statistically significant ( $p < .0001$ ). This indicates that on its own, TALL had a positive association with the outcome variable. The *t* statistic and its *p* value are reflected in the last two columns of **Table 58** below.

**Table 58: The t Statistic for TALL as the predictor of 2014 average performance (n=604)**

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	T Value	Pr >  t
Intercept	Intercept	1	48.26179	1.31587	36.68	<.0001
TALL	TALL score	1	0.18534	0.03005	6.17	<.0001

The results of the linear regression of 2014 average performance on TALL are graphically presented in the fit plot in **Figure 9** below.

**Figure 9: The fit plot for TALL as predictor of 2014 average (n=604)**



It is evident from the regression line and the location of the data points in **Figure 9** above that TALL had a positive but moderate relationship with 2014 average.

Overall, these results mean that on its own, TALL was a significant predictor of 2014 average. The *t* statistic for this test from the analysis attests to this. This

statistic is positive and statistically significant. Also, the fit plot for the result of a linear regression with TALL as the predictor of 2014 average confirms that this test had a linear but not strong relationship with the participants' end of 2014 average performance, a confirmation that the latter was moderately but significantly predicted by the test. Thus, unlike PTESLAL, TALL possessed a statistically significant degree of predictive validity for both a smaller (n=98) and bigger (n=604) 2014 sample, an indication that the test was a reliable measure of academic literacy for both.

#### *4.5.3.2 Grade 12 English as a predictor of 2014 average*

Following the linear regression involving the participants' performance on TALL as the predictor and their 2014 average as the predicted variable, a linear regression of the latter was also carried out on Grade 12 English. The *F* statistic from this analysis was  $F(1,476) = 49.72, p = .0001$ . This means that one of the predictors in this model had a positive relationship with the participants' 2014 average performance. This is shown in the last two columns of **Table 59** below.

**Table 59: The F statistic for the model with Grade 12 English as the specified predictor of 2014 average performance (n=478)**

<b>Analysis of Variance</b>					
<b>Source</b>	<b>DF</b>	<b>Sum of Squares</b>	<b>Mean Square</b>	<b>F Value</b>	<b>Pr &gt; F</b>
<b>Model</b>	1	6625.32974	6625.32974	49.72	<.0001
<b>Error</b>	476	63426	133.24867		
<b>Corrected Total</b>	477	70052			

Secondly, the R-Square from this analysis equalled 0.0946, meaning that 9% of the variance in the participants' 2014 average performance could be explained by the predictor model involved. This is captured in the last two columns of **Table 60** below.

**Table 60: The R-Square for the model with Grade 12 English as the specified predictor of 2014 average performance (n=478)**

<b>Root MSE</b>	11.54334	<b>R-Square</b>	0.0946
<b>Dependent Mean</b>	55.02510	<b>Adj R-Sq</b>	0.0927
<b>Coeff Var</b>	20.97831		

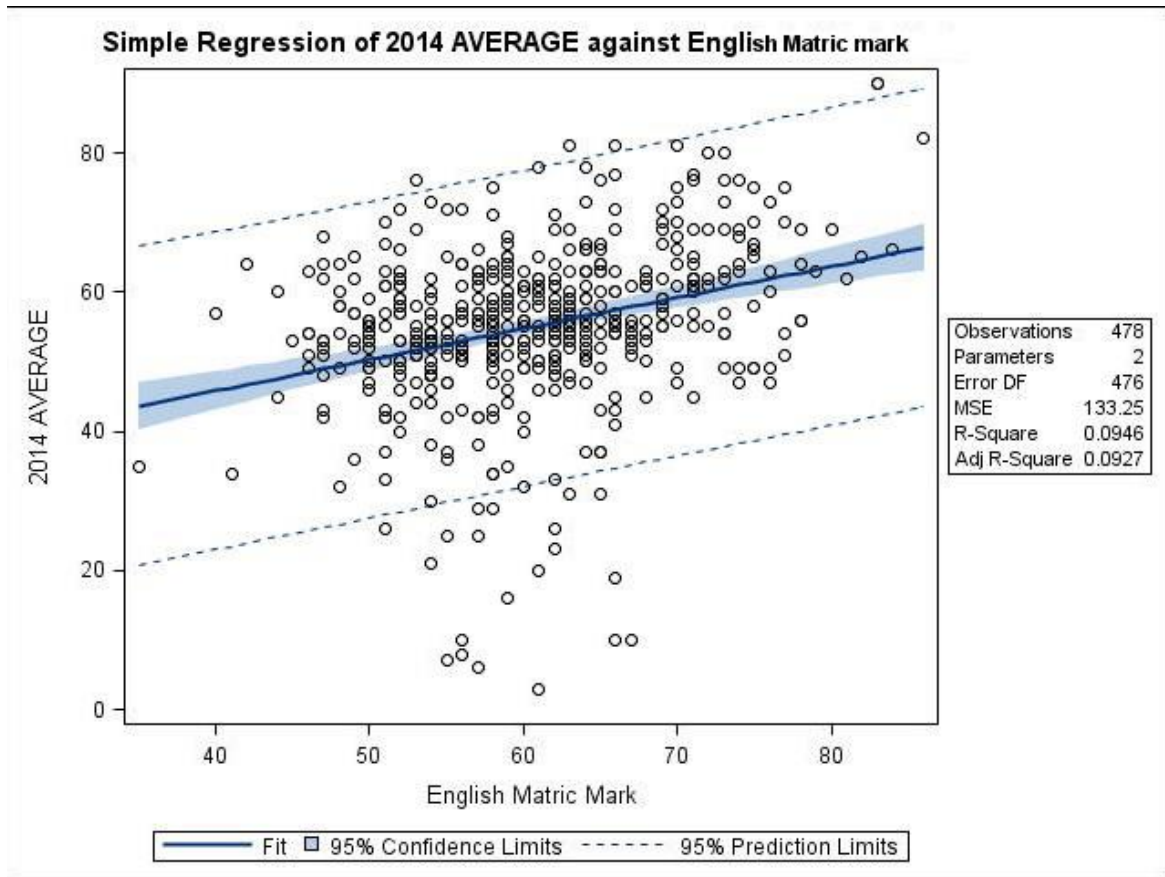
Lastly, the *t* statistic for Grade 12 English amounted to 7.05 and was statistically significant ( $p = <.0001$ ). This means that considered alone, the participants' performance on Grade 12 English had a positive relationship with their 2014 average performance. This is captured in the last two columns of **Table 61** below.

**Table 61: The *t* statistic for Grade 12 English as a predictor of 2014 average performance (n=478)**

<b>Parameter Estimates</b>						
<b>Variable</b>	<b>Label</b>	<b>DF</b>	<b>Parameter Estimate</b>	<b>Standard Error</b>	<b>T Value</b>	<b>Pr &gt;  t </b>
<b>Intercept</b>	Intercept	1	27.98516	3.87089	7.23	<.0001
<b>Grade 12 English</b>	Grade 12 English score	1	0.44642	0.06331	7.05	<.0001

The results of a linear regression involving Grade 12 English as the predictor of 2014 average are graphically captured in the fit plot in **Figure 10** below.

Figure 10: The fit plot for Grade 12 English as the predictor of 2014 average performance (n=478)



The slope of the regression line as well as the location of the data points in **Figure 10** above show that Grade 12 English had a positive and stronger relationship with the outcome variable.

The overall meaning of these results is that on its own, Grade 12 English was a better predictor of 2014 average than TALL. This is evident in the higher and statistically significant *t* statistic reported for the former. Furthermore, the fit plot for the linear regression analysis involving Grade 12 English and 2014 average confirms that the former had a stronger linear relationship with the latter when compared to TALL. This means therefore that at linear regression analysis level, TALL and Grade 12 English were, to different extents, significant predictors of

2014 average performance and that Grade 12 English had a higher predictive validity than TALL.

#### 4.5.3.3 TALL and Grade 12 English as predictors of 2014 average

Following the linear regression of 2014 average on TALL and Grade 12 English respectively, a multiple regression of the former on these predictors was also computed. The  $F$  statistic from this analysis was  $F(2,475) = 26.84$ ,  $p = <.0001$ . This means that at least one of the predictors in this model had a positive relationship with 2014 average performance. This is captured in the last two columns of **Table 62** below.

**Table 62: The F statistic for the model with TALL and Grade 12 English as specified predictors of 2014 average (n=478)**

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
<b>Model</b>	2	7113.90868	3556.95434	26.84	<.0001
<b>Error</b>	475	62938	132.50061		
<b>Corrected Total</b>	477	70052			

Secondly, the R-Square from this analysis equalled 0.1016, meaning that 10% of the variance in the 2014 average performance could be explained by the model.

This is shown in the last two columns of **Table 63** below.



**Table 63: The R-Square for the model with TALL and Grade 12 English as specified predictors of 2014 average performance (n=478)**

<b>Root MSE</b>	11.51089	<b>R-Square</b>	0.1016
<b>Dependent Mean</b>	55.02510	<b>Adj R-Sq</b>	0.0978
<b>Coeff Var</b>	20.91934		

Finally, the  $t$  statistic for TALL equalled 1.92 and was not statistically significant ( $p = 0.0554$ ) while the same statistic for Grade 12 English was 5.95 and was statistically significant ( $p = <.0001$ ). This is reflected in the last two columns of **Table 64** below.

**Table 64: The  $t$  statistic for TALL and Grade 12 English as predictors of 2014 average (n=478)**

<b>Parameter Estimates</b>						
<b>Variable</b>	<b>Label</b>	<b>DF</b>	<b>Parameter Estimate</b>	<b>Standard Error</b>	<b>T Value</b>	<b>Pr &gt;  t </b>
<b>Intercept</b>	Intercept	1	27.82800	3.86088	7.21	<.0001
<b>TALL</b>	TALL score	1	0.07284	0.03793	1.92	0.0554
<b>Grade 12 English examination</b>	Grade 12 English score	1	0.40114	0.06739	5.95	<.0001

The results of a multiple regression of 2014 average on Grade 12 English and TALL mean that Grade 12 English had a higher predictive relationship with 2014 average and that this was probably not a result of chance. This is evident in its higher  $t$  statistic and its low  $p$  value. On the other hand, the  $t$  statistic for TALL was lower than that of Grade 12 English and its  $p$  value was slightly higher than .05, meaning that this could have been a result of chance. This means that Grade

12 English was a better predictor of 2014 average than TALL. Also worth mentioning in connection with these results, however, is that the adjusted R-Square for Grade 12 English as the predictor of 2014 average alone increases from 9.27% to 9.78% when paired with TALL in a simultaneous comparison of the two tests as predictors of the same outcome variable. This means that while TALL was not a better predictor of 2014 average than Grade 12 English in a simultaneous regression, it added 1% to Grade 12 English's ability to predict 2014 average. This means that TALL slightly added to the predictive efficiency of Grade 12 English and that it therefore possessed incremental validity in relation to the latter (SCU 2015).

#### **4.6 Conclusion**

In this chapter, the data collected for analysis have been identified and an explanation has been offered for the way that these data were organized for the purpose of analysis. Furthermore, the statistical procedures used to run several quantitative analyses of these data have been identified, explained and justified. Finally, the results of the analyses carried out on each set of data have been presented and discussed.

The next chapter will focus on the analysis and interpretation of these results.

## **Chapter 5: Interpretation and analysis of the results**

### **5.1 Introduction**

The key focus of this chapter is to interpret and analyse the results of this study that were presented in Chapter Four. It explains why Grade 12 English was the best predictor of first year academic performance in relation to the National Benchmark Test in Academic Literacy (NBT AL), Proficiency Test English Second Language Advanced Level (PTESLAL) and Test of Academic Literacy Levels (TALL) in 2012 and 2014 at the university chosen for the study. The merits and demerits of this result are also discussed. Next, the chapter accounts for PTESLAL's inconsistent possession of incremental validity with regard to the participants' first year academic performance in 2012 and 2014 in relation to the best predictor. It then attempts to account for the failure of NBT AL to predict academic performance for the 2012 participants in relation to PTESLAL and Grade 12 English, and consequently to possess incremental validity in relation to the best predictor. Also, the chapter attempts to account for the consistent ability of TALL to possess incremental validity in relation to the best predictor for the two sets of participants in 2014. Finally, the chapter briefly deals with the results of recent studies in the context of which those of the present study with regard to the predictive validity of tests of academic literacy and Grade 12 English results should be understood. These findings are all dealt with one after another in the remaining sections of this chapter.

## 5.2 The results in relation to Grade 12 English

The results of all the linear and multiple regression analyses carried out on the three data sets used in this study mean that Grade 12 English was the best predictor of first year academic performance in 2012 and 2014 when compared to NBT AL, PTESLAL and TALL. This means that this assessment possessed better predictive validity with regard to the performance of the participants in their various programmes of study at the end of the two years. In essence, the results mean that the participants' performance on Grade 12 English related positively and better with their end of first year academic performance than it did with their performance on NBT AL, PTESLAL and TALL. Research of the kind carried out in this study “often depends on the measurement of observed variables” in the form of test scores which “are used to provide evidence in support of meaningful claims about learners' knowledge, skills, and abilities (KSAs)” (Purpura, Brown & Schoonen 2015: 38). Put differently, the test scores are used “to make theory-driven inferences about how the underlying constructs fluctuate over time and under varying conditions” (Purpura et al. 2015: 38). Given the essential role that test constructs play in the claims made and decisions taken on the basis of test performance “... it is imperative that the theoretical constructs we use in our work reflect collective understandings of the phenomena we wish to measure, even if this means moving beyond traditional conceptualizations of the constructs in our own subfield of applied linguistics” (Purpura et al. 2015: 38). This means that a conceptually clear and defensible definition of a construct to be measured should inform the design and development of the measurement instrument to be used. Van

Dyk (2015: 168) has asserted the importance of clear and defensible construct definition especially in high stakes testing in the following words:

For the purpose of high stakes testing particularly, it is of the utmost importance that there is theoretical justification for the test. Construct validity is therefore an issue, as it refers to the degree an underlying theory (or theories) rationalises the measurement of a specific domain.

The evidence of the predictive validity of Grade 12 English revealed in the present study notwithstanding, it is difficult for one to conclude that the construct underpinning this assessment meets the criteria proposed by Purpura et al. (2015) above. In fact, Umalusi, the Council for Quality Assurance in General and Further Education and Training, has commissioned several reports aimed at understanding the nature of the problems affecting the Grade 12 language examinations. The fourth of these reports was a product of the Home Languages project which drew from the language assessment expertise of the Inter-institutional Centre for Language Development and Assessment (ICELDA), a consortium of four South African universities, namely Stellenbosch, North West, Free State and Pretoria. Among the challenges the ICELDA team of experts was required to investigate was whether the construct underpinning the Grade 12 Home Language assessment was conceptually clearly defined and justifiable. As Weideman, Du Plessis and Steyn (2015: 7) explain, “the initial brief of the Home Languages project was to determine whether the assessment conformed to one of the basic principles of language testing namely, whether the construct being assessed was clear and well defined.” The fact that Umalusi itself specifically commissioned a report for this purpose is an indication of admission on its part that the construct might need to be better defined to ensure that it is theoretically defensible. What is clear from this is

that Umalusi has now started to respond to the critical principle of testing, namely, that “a construct first needs to be defined and then empirical evidence (internal consistency, intra-test validity and inter-test validity, among others) should be presented to demonstrate that a specific test indeed measures what it purports to measure” (Van Dyk 2015: 168). In the section below, I deal with this principle in relation to the Grade 12 language assessments.

### **5.2.1 Construct definition with regard to Grade 12 language assessments**

It was demonstrated in Chapter One that the Curriculum and Assessment Policy Statement (CAPS) that supposedly informs language teaching and assessment for the National Senior Certificate appears to be informed by Cummins’s (1984, 1996, 2009) theory of language ability which distinguishes between the Basic Interpersonal Communicative Skills (BICS) and Cognitive Academic Language Proficiency (CALP), the two differentiated types of language ability required of language users in conversational and academic language use settings respectively. It was also pointed out in that chapter that the CAPS for language teaching and learning at Grade 12 level is consistent with the Bachman and Palmer (1996) view of language ability which promotes a very open and authentic understanding of language ability. It is also clear that this curriculum embraces Patterson and Weideman’s (2013b) view of how the sphere of language use renders it typical to the context in which it operates. To this end, CAPS is clearly an echo of Communicative Language Teaching (CLT) where “a command of language cannot solely be defined with reference to the grammatical control learners have of the language. In any kind of interaction through language, be it political, academic,

ethical, aesthetic or economic, the functional uses and purposes of language have precedence over language structure” (Weideman et al. 2015: 8-9). In this sense, in CAPS, “language is viewed not only as an instrument of social interaction in various spheres of discourse, but also as a general, and functionally definable, communicative tool” (Weideman et al. 2015: 8). Having determined whether the Grade 12 language syllabus is a further manifestation of all these views of language ability and if the three papers constituting the final examinations align with these views, the ICELDA team involved in the Home Languages project found, however, that none of these was actually the case (Weideman et al. 2015: 9). Weideman et al. (2015: 9) summarize this observation in the following words:

... the detailed syllabus contained in *CAPS* is not an entirely consistent working out of CLT, the approach which it claims to subscribe to. What is more, the extensive work suggested in the emphasis that is retained ... on the appreciation of language used for aesthetic purposes and enjoyment of literature, indicates a much larger role for literature study than is conventional in most interpretations of CLT.

Similar to this observation, Van der Walt (2010) has argued that the Grade 12 English First Additional Language curriculum espouses a focus on the development of cognitive academic language but that the learning materials used in the curriculum are too general to promote learner achievement in academic language. It is for this reason that Ayliff (2010) has concluded that the communicative approach promoted by this curriculum places little emphasis on academic language development and that this is the reason for the poor writing ability that first year students bring to universities. Also, as shown earlier in Chapter Three, Paper 1 of the Grade 12 English examination is intended to assess Language in context, Paper 2 focuses on Literature, and Paper 3 purports to assess Writing. Weideman et al.

(2015) conclude that the foci of these papers are not aligned to the language views that evidently inform CAPS and that the curriculum espoused in CAPS itself is not aligned with the assessment proposed in this document. These findings and the concerns they raise have implications for the validity of the Grade 12 English examinations. These implications are identified and discussed in the section that follows.

### **5.2.2 The issue of the validity of the Grade 12 English examination**

In Chapter Two of this study, it was pointed out that validity is the term used for judging whether an assessment instrument measures what it purports to measure. It was also pointed out in that chapter that any claim for the validity of such an instrument can only be made if evidence can be produced and presented to justify how the scores on it are interpreted and used. In the words of Purpura et al. (2015: 39),

...it is critical that prior to using scores to make claims about constructs, the validity of the scores for the intended purpose (s) should be submitted to validity evaluation. In this way, the meaningfulness and appropriateness of the scores for their intended interpretation, and for their use in making decisions, can be adequately justified with relevant evidence.

As a matter of logic, it is not possible for one to make a validity claim for a measure unless the construct of such a measure is clearly articulated. The finding by Weideman et al. (2015) that the construct of the Grade 12 English examination is yet to be clearly defined means that validating the interpretation and use of the scores obtained on this examination has not been possible. This is so because construct validity is the precondition for all other kinds of validity that were also presented and explained in Chapter Two of the present study. It is primarily because



of the need to assess a test taker's standing in relation to or possession of a particular construct that assessments such as the Grade 12 English examinations are developed and used in the first place. If this construct is not clearly defined, it is very likely that the measure used to assess it will not do so with an acceptable degree of accuracy. The reported lack of clarity in the definition of the construct informing Grade 12 English therefore means that the construct validity of this examination is not known and that the basis for its predictive validity for first year academic performance evident in this study raises questions. Purpura et al. (2015: 43) have explained the essentiality of construct definition to construct validity in the following words:

Evidence of construct validity is meant to justify that the measure, as designed, is a meaningful representation of the underlying ... construct being assessed, given what is known about the construct. To obtain this evidence, the researcher first defines the construct theoretically by identifying what needs to be measured.

Implied in the unknown and probable lack of construct validity for Grade 12 English is the issue of its reliability. I will now turn to the discussion of this implication in the section below.

### **5.2.3 The issue of the reliability of Grade 12 English assessment**

As pointed out in Chapter Two, measurement reliability refers to the ability of an instrument to measure what it is intended to measure consistently. It was also pointed out in the same chapter that reliability is one of the preconditions for validity. A valid test is, by default, also reliable. In other words, no claim can be made that a test is valid if it is not reliable. Thus, the evident lack of validity by Grade 12 English referred to earlier implies that it is probably also lacking in

reliability. Indeed, among the recommendations by a 2007 report commissioned by Umalusi on the quality of the Grade 12 examination was the need to consider the use of technology to improve the reliability of this assessment (Du Plessis, Steyn & Weideman 2016).

In Chapter Two, reference was made to two statistical procedures used for determining test score reliability, namely, Cronbach's alpha and Greatest Lower Bound. Both these statistics require the use of software to determine the reliability of measures underpinned by unidimensional and multidimensional constructs respectively. The recommendation by the 2007 Umalusi report therefore means that no initiative had been taken to use technology to determine the reliability of the Grade 12 examinations. In the three data sets analysed in the present study, Grade 12 English appeared to be consistent in its ability to predict first year academic performance better than the other predictors used. Given the relationship between validity and reliability, this also implies that the assessment was reliable in its measurement of the construct it purports to measure. In the context of the lack of clarity in the definition of its construct, unknown construct validity and reported shortcoming with regard to the use of technology to determine its technical consistency, however, one wonders what the basis for its consistent predictive validity was in the three data sets analysed in the present study.

Failure to compute the reliability of an assessment poses threats to the evaluation or generalizability claims that can be made on the basis of its scores (Purpura et al. 2015:64). This means therefore that the finding in the present study that Grade 12 English possesses better predictive validity than NBT AL, PTESLAL and TALL in

that respect may lack credibility and that this is in the main a function of the absence of research evidence to support its psychometric soundness. The one explanation for its predictive force may lie in its being closely associated with the whole of the Grade 12 marks. This is an explanation that has, in fact, been suggested in other studies (Myburgh 2015). The strength of the prediction is therefore closely related to its being a proxy for the full range of closely clustered, associated Grade 12 marks. That explanation may well need further exploration.

### **5.3 The results in relation to PTESLAL**

The results of the analysis of the first and second data sets revealed that in the first instance, PTESLAL did not possess predictive and incremental validity and that it possessed these types of validity from the analysis of the second set of data. As was pointed out in Chapter Four, the implication of this is that the test was not consistent in the way it treated the two sample groups used and that, by extension, this brings its validity under question. Other than the possibly poor psychometric qualities inherent to the test, a related possible explanation for this inconsistency is that performance on the test is test and sample dependent. The latter is the essential characteristic of the tests developed based on the models of Classical Test Theory (CTT). In such tests,

examinee test scores and corresponding true scores will always depend on the selection of assessment tasks from the domain of assessment tasks over which their ability scores are defined. Examinees will have lower true scores on difficult tests and higher true scores on easier tests ... (Hambleton & Jones 1993: 38)

This means that the item statistics that are conventionally the focus of Classical Item Analysis, namely, item difficulty, item discrimination and the standard error of

measurement are also test and sample dependent and therefore lack generalizability. In the words of Bachman (2004: 139), “the item statistics that we obtain are dependent on the particular group, or sample, of test takers who take the test...” and as a result, “... it is very difficult to compare items whose item statistics are based on the performance of different groups of test takers, and to compare test takers whose scores are obtained from different tests.” This sounds like an explanation for the lack of invariance in the performance on PTESLAL that was revealed by this study. This possibly renders the one instance where PTESLAL showed evidence of possessing predictive and incremental validity with regard to the participants’ first year performance untrustworthy. A test that possesses predictive validity should be able to do so consistently regardless of the sample of test takers taking it and that of the test items used. Evidence should be available, however, for such a test’s reliability and construct validity. As pointed out earlier in this chapter, construct validity is the precondition for the other two traditional types of validity, namely, content and criterion-related validity.

#### **5.4 The results in relation to NBT AL**

The results of the analysis carried out on the first data set also mean that NBT AL was not able to predict first year academic performance when considered alone and when simultaneously considered with PTESLAL and Grade 12 English as predictors of this performance. As pointed out in Chapter One, NBT AL is one of only two tests specifically developed to measure academic literacy among first time applicants to South African universities. This means that it should be expected that performance on this test significantly mirror academic performance, especially at

first year level. This should in particular be the case, because the construct underpinning this test was the outcome of a wide consultation of panels of academics considered to be familiar with the demands of university education that first year students typically have to negotiate in the medium of instruction (Cliff 2015). It is an outcome of this consultation that Cliff and Yeld (2006: 19) have described as the construct of academic literacy for the design and development of NBT AL that focuses on

students' capacities to engage successfully with the demands of academic study in the medium of instruction of the particular study environment. In this sense, success is constituted of the interplay between the language (medium of instruction) and the academic demands (typical tasks required in higher education) placed upon students.

If this is indeed what a test of academic literacy should measure and what is required of first year students to succeed at university, the results of the present study mean that for the particular group of participants used, NBT AL failed to show evidence of a meaningful relationship with their first year performance at university. Two reasons are possible for this.

Firstly, except for a recent study by Cliff (2015) in which the acceptable reliability indices as well as evidence of the construct validity for this test are reported, in general, peer-reviewed research evidence of the psychometric utility of the test has been very difficult to come by in the public domain. By nature, test development is a never-ending process of research whose aim should be to ensure that evidence for the psychometric quality, relevance, and ultimate credibility of a test of the NBT AL's calibre is publicly available. As Van Dyk (2015: 164) explains, this kind of research "is in a post-modern society more important than ever due to the fact that

one wishes to be transparent, fair and justified at all times when taking decisions” (Van Dyk 2015: 164). If made available in the public domain, the advantage of this kind of research for a test such as NBT AL is that it can attract constructive criticism from peers in language and educational testing. Logically, this criticism has great potential to help initiate efforts towards the improvement of the test when necessary. The very nature of the project in which NBT AL is developed means that the research of the kind that one envisages for the test does possibly exist. Van Rooy and Coetzee van Rooy (2015: 34) observe that “while the NBT has not yet enjoyed the same amount of scholarly investigation” as TALL, for example, “it is clear that similar care has been taken in its development”. As pointed out above, however, keeping this research away from public scrutiny militates against possible growth in the validity of the test.

Secondly, on the face of it, the construct of NBT AL does look like it is informed by what students should be able to do in the medium of instruction for them to succeed academically. Furthermore, as shown in Chapter Three, the space allocated to the operational subdomains of this test and the varied levels of cognitive demand at which the items tapping these subdomains are pitched make systematic sense. It seems, however, that failure by the test to generate evidence of a predictive relationship with first year academic performance might be a function of the task types used to measure this construct. As demonstrated in Chapter Three, traditional multiple choice items are solely used to tap test taker levels of ability on these subdomains. An example of this type of item measuring ‘cohesion’, one of the sub-constructs assessed in the NBT AL, is presented below:

1. **In paragraph 2, we read: “This is a craving for harmful products that we are better off without.” In this sentence, the word “This” mainly refers to**
  - a. enviable attention
  - b. excessive consumption
  - c. unnecessary wants
  - d. increasing demands

(CETAP 2016: 10)

The restricted discourse focus of this item bears testimony to the injustice that the traditional multiple choice item can do to a rich construct such as the one underpinning this test, as opposed to a test task that requires the test taker to restore a scrambled text or restore sentences that have been systematically removed from a text to assess the same sub-construct, for example. The latter task types have more potential to capture test taker ability to engage with cohesion at the level of the discourse than the traditional multiple choice item type used in the example above. This way of assessing cohesion mirrors the way academic texts typically require students to make connections between parts of a text in order for them to achieve a more complete understanding of such a text.

This is not to say that multiple choice items are completely of no utility to academic literacy testing. Indeed, Miller et al. (2009: 202) have argued that this item type “can effectively measure various types of knowledge and complex learning outcomes” and that “it is free from some of the common shortcomings characteristic of the other item types”. Miller et al. (2009: 203) raise a concern about this item type, however, which is a probable explanation for a test like NBT AL to fail to

offer evidence of its relationship with academic performance, the criterion with which it should closely relate:

The problems presented to students are verbal problems, free from the many irrelevant factors present in natural situations. In addition, the applications students are asked to make are verbal applications, free from the personal commitment necessary for application in natural situations. In short, ... the multiple choice item, ... measures whether the student knows or understands what to do when confronted with a problem situation, but it cannot determine how the student actually will perform in that situation.

This limitation means that the multiple-choice item carries great potential to limit the predictive kind of criterion-related validity that a test like NBT AL should have. In the words of Yeld (2001: 248), criterion-related validity refers to how a test “relates to phenomena external to the test, for example other tests, to future performance in the ‘real’ world, and to educational and other systems in civil society”. As pointed out in Chapter One, NBT AL is a criterion-referenced test that was introduced against the background of the need to measure the academic readiness of undergraduate students. As also shown in Chapter Four, however, performance on this test by the group of participants used in this study failed to produce evidence of a meaningful predictive relationship with this criterion. It is probably not a good idea, therefore, that a medium to high stakes test of the NBT AL’s calibre has been solely reliant on multiple choice items to assess a critical criterion such as readiness for university education.

This makes more sense when one recognizes that the evolution of language teaching and assessment has always been influenced by a contemporary understanding of what the constituents of language ability are. As pointed out elsewhere in this study, current conceptualizations of language ability mainly rest on the



understanding that language is a means of communication and that the ability to use it cannot solely be the function of the knowledge of how a language is structured. As Weideman (2014: 5) explains, “communication implies interaction amongst two or more individuals, and this interaction may even be displaced (non-simultaneous) and remote, depending on the communication medium”. This dictates that the methods employed both for teaching and assessing this ability should also be aligned with this view. Weideman (2014: 5) rightly argues further that “interaction with academic texts is what is most commonly and justifiably thought to constitute the appropriate source that provides material for tests of academic literacy.” The emphasis of the interactive nature of communication as the basis for language ability means that multiple choice items might not be the most efficient item type for assessing academic literacy. This seems particularly true when one acknowledges that multiple choice testing is as old as the traditional discrete-point language teaching methodologies that have now given way to CLT.

## **5.5 The results in relation to TALL**

Finally, the results of the linear and multiple regression analyses of the second and third data sets involving TALL as one of the predictors of 2014 average mean that this test was able to predict the participants’ end of first year academic performance and that it could add slightly to the predictive efficiency of Grade 12 English, the best predictor of this performance on the two occasions. The ultimate meaning of this is that for the participants involved, TALL possessed both predictive and incremental validity. This was the case in the analyses of the two 2014 subsets of data where TALL was involved as one of the predictors. This means that performance on the

TALL used for this study was evidently not sample dependent. It was pointed out in Chapter Two of this study that Van der Walt and Steyn (2007, 2008) used a computer program called FACETS to run a Rasch Analysis of TALL data and confirmed the validity of this test with regard to its level of difficulty and appropriate fit to the model. Rasch is a one parameter model of Item Response Theory (IRT), “a general statistical theory about examinee item and test performance and how performance relates to the abilities that are measured by the items in the test” (Hambleton & Jones 1993: 40). The key characteristic of IRT models is that, unlike CTT models, they are used to develop tests in which test performance is test and sample independent (Bachman 2004). This is probably the reason for the consistent manner in which TALL was able to predict performance in the analysis of the two subsets of data referred to earlier. Also, the task types used for assessing academic literacy in the test are an adaptation of traditional multiple choice items whose use rested on a restrictive understanding of language ability and which would fall short of efficiently assessing the construct of academic literacy underpinning this test (cf. Weideman 2014). An example of this adaptation is evident in the item type used to measure, ‘grammar and text relations’, a sub-construct assessed in TALL:

**In the following, you have to indicate the possible place where a word may have been deleted, and which word belongs there**

Charles Goodyear (1800-1860) invented the vulcanization of rubber when he was experimenting by heating a mixture of rubber and sulphur. The Goodyear story is one of either pure luck or careful research, but both are debatable. Goodyear insisted that it was **i** and **ii**, though **iii** many **iv** contemporaneous **i** accounts **ii** indicate **iii** the **iv**

**1. Where has the word been deleted?**

- A. At position (i)
- B. At position (ii)**
- C. At position (iii)
- D. At position (iv)

**2. Which word has been left out here?**

- A. indeed
- B. very
- C. former**
- D. historically

**3. Where has the word been deleted?**

- A. At position (i)
- B. At position (ii)
- C. At position (iii)
- D. At position (iv)**

**4. Which word has been left out here?**

- A. historical
- B. latter**
- C. now
- D. incontrovertibly

(Weideman 2014: 10)

This item type bears testimony to the innovation, creativity and flexibility that the open view of academic language ability that underpins a test of academic literacy such as TALL call for. Designed in this creative and innovative way, this item exhibits evidence of greater potential to do justice to the multi-faceted construct of academic literacy that informs TALL. In the words of Weideman (2014:10), the item is “recognizably cloze procedure, but the adaptation is such that it overcomes the logistical constraints associated with marking the answers by hand, and adds more dimensions to what may otherwise be another humdrum testing technique: that of testing not only textuality, but potentially also grammatical relations as well as (in some cases) communicative function”

On the basis of the results of his study that was referred to earlier in this chapter, Van Dyk (2015: 183) has warned, however, that TAG, the Afrikaans version of TALL, and by equivalence, the English version itself

should be used for the purpose it was designed for: for informing placement. As a single predictor TAG performs weak at best. In combination with other predictors, TAG makes no discernible contribution. Therefore there can only be one conclusion: TAG (and similar tests) should not be used for access decisions.

## **5.6 Recent studies on the predictive validity of academic literacy tests and Grade 12 results**

The results of the present study with regard to TALL, NBT AL and Grade 12 English should be interpreted in light of those of recent ones by Myburgh (2015), Van Dyk, Van de Poel and Van der Slik (2013), Van Rooy and Coetzee Van Rooy (2015), and Van Dyk (2015). In a study focusing on the predictive ability of two tests of academic literacy in relation to the English Home Language examination, Myburgh (2015) found that the latter was a better predictor of academic performance than the former. In this sense, the results of Myburgh's study support those of the current one. Myburgh (2015) proposes three main reasons for the relative predictive force of the English Home Language examination in her study. The first of these relates to the advantage that this examination had over the other two assessments as a result of the longer duration of its administration as well as its longer history of existence. Myburgh (2015: 94) rightly argues that "to compare a 300 mark examination, combined with the accumulation of 10 years of previous assessments, with one or two 60 mark tests" that have existed for a short period of time is too ambitious. With regard to the first of these advantages, Myburgh (2015: 94) contends that "it is well known that a longer test measures more reliably than a short test." Secondly, high schools conventionally prepare learners in advance for the assessments they must take and this is typically not the case with academic literacy tests that are currently used in South Africa (Myburgh 2015). Lastly, it is likely that the average performance of learners on all their subjects at high school is homogenous and that the results of one therefore tends to move in the direction of

their overall average performance (Myburgh 2015), which was the outcome variable in both Myburgh's and the present study.

So, while Grade 12 English assessment may lack an articulated construct, aligned with the curriculum, as we have noted above, its measurement is still potentially the accumulation of enough information about ability to exceed the predictive qualities of shorter assessments.

In another study focusing on the predictive validity of perceived preparedness of academic reading ability, TALL and TAG, gender, first language and race, and Grade 12 results, Van Dyk et al. (2013) found, among others, that Grade 12 results had the strongest impact on first year academic performance when compared to TALL and TAG. Van Dyk et al. (2013: 361) subsequently observe, however, that the effect of academic literacy on this performance "can also be indirect since scores on TALL and TAG are expected to be influenced by students' scores on secondary school performance (Grade 12) results as well."

In yet another study focusing on the predictive validity of matric average results, matric language marks, scores on academic literacy tests and academic literacy courses, Van Rooy and Coetzee Van Rooy (2015) also found, among others, that Grade 12 results were better predictors of academic success than both NBT AL and TALL. Van Rooy and Coetzee Van Rooy found, however, that these results were useful predictors of academic success only for the scores above 65% and not for those below this mark. This was also a finding in the study by Van Dyk (2015) involving Grade 12 results as a predictor of academic performance. Van Dyk (2015: 181) concludes therefore that "the matric results ... remain the single best

predictor of academic performance. Note, however, the difference between the two quartiles when considering the matric results: academic performance of students in the top quartile can be predicted much stronger than those in the bottom quartile”. Van Rooy and Coetzee van Rooy (2015: 42) summarise these findings in the following words: “matriculation results are better predictors of students with high marks, who were in the lowest risk category from the beginning, whereas for students with lower school marks, the predictive value of the marks is not high, and it is therefore not a very helpful measure if used to determine admission at all”. Also, Van Rooy and Coetzee Van Rooy (2015: 43) found that the manner in which the matric scores were able to predict success for their participants was inconsistent: “Some students who achieved a matric average of below 65% perform well at university and some of them do not succeed academically at all.” In the view of Van Rooy and Coetzee Van Rooy (2015), these are problematic results because they imply that the matric results that are below the third quartile are not useful predictors of academic performance.

The study by Van Rooy and Coetzee Van Rooy (2015) also revealed that while NBT AL and TALL were not better predictors of academic performance than Grade 12 results, academic literacy courses predicted academic success better than these results. Unlike the Grade 12 results which they found to be good predictors of academic success for the scores in the upper quartile only, the courses showed evidence of being able to do this across all the quartiles. As Van Rooy and Coetzee Van Rooy (2015: 42) explain, “unlike matric marks, there is no differential effect across the marks spectrum for academic literacy modules: the correlation in the top

halves and bottom halves are quite similar ...” for all the four modules they investigated. The predictive validity of these courses means that their content is aligned to the kind of discourse competence that is required for students to succeed in their studies (Van Rooy & Coetzee Van Rooy 2015; Van Dyk 2015). Logically, these modules should also be based on the constructs that underpin the tests of academic literacy that are conventionally used to identify students for academic literacy instruction. This means that academic literacy teaching cannot be divorced from academic literacy assessment. In the words of Myburgh (2015: 95), “if the results of academic literacy interventions are better predictors, one would, however, still need a measure of academic literacy to decide who should take such courses.”

Indeed, the results of Van Rooy and Coetzee Van Rooy’s study (2015) have shown that Grade 12 English results do not correlate well with NBT AL and TALL scores, an indication that such results cannot justifiably be used for channelling students into academic literacy courses. This is true even in the face of the finding by Van Rooy and Coetzee Van Rooy (2015) that academic literacy tests on their own are not better predictors of academic performance than these modules. The probable reason for the difference in the predictive efficiency of the two artefacts is that academic literacy courses are typically offered for a full year while academic literacy tests are at most a three hour long affair. This means that scores on academic literacy modules are a larger sample of the ability to negotiate the demands of university education in the medium of instruction than performance on academic literacy tests. It is for this reason that Myburgh (2015: 95) wonders “whether tests, no matter how well constructed and developed, would perhaps not

always lack the richness and depth which programmes of longer duration [such as academic literacy modules] offer in terms of accuracy and reliability”.

## 5.7 Conclusion

This chapter was an attempt to interpret and analyse the results of this study. It acknowledges that Grade 12 English is the best predictor of academic success when compared to the rest of the predictive measures used. Subsequent to this, it presents existing research findings that raise questions about the predictive efficiency of this assessment. The findings are that Grade 12 English is not aligned with the curriculum that informs it, its construct is yet to be clearly articulated, and that no research evidence exists for its psychometric soundness. Furthermore, the chapter attempts to account for the failure of PTESLAL to maintain consistent incremental validity throughout its involvement as a predictor variable in the study. The chapter ascribes this to the test’s probable lack of psychometric efficiency and to the observation that the test was probably developed through models of test development which are susceptible to this inconsistency. It also attempts to account for the evidently weak relationship between NBT AL and first year academic performance. It argues that the scarcity of peer-reviewed studies on this test deprives it of the opportunity for exposure to rigorous examination and continuous refinement, and that its task types might not be the best for assessing a language ability naturally as multifaceted as academic literacy. Furthermore, the chapter accounts for the evidently consistent incremental validity exhibited by test taker performance on TALL in the study. The argument is made that the psychometric soundness of this test has been publicly established. Finally, the chapter draws on



recent predictive validity studies within which it suggests the findings of the current study should be understood. In the main, it draws attention to the recently established predictive superiority of academic literacy modules and the implied utility of academic literacy tests for placing students in such modules.

The next chapter summarizes the key findings of the study, deals with its limitations and makes recommendations for future research, before I turn, in the final chapter, to a consideration of the implications of this study for validity theory in language assessment.

## **Chapter 6: Conclusions, recommendations and suggestions for future research**

### **6.1 Introduction**

In the past two to three decades, more and more high school leavers have been gaining access to institutions of higher learning throughout the world. In South Africa, this has been accompanied by concerns that these candidates do not possess an adequate ability to cope with the demands of academic education in the medium of instruction. Traditionally, Grade 12 results are used as the basis for allowing or denying these students access to universities and, by implication, assessing their ability to demonstrate evidence of possessing the discourse competence required of them to succeed at university. In recent years, however, the ability of these results to provide valid information about the academic readiness of first time applicants for admission to universities has been questioned. As a result, universities have sought additional ways of predicting the ability of their students to succeed academically. Some of these methods have involved using academic language tests that are deemed to provide this information, including those of academic literacy that have now been in use for about a decade.

### **6.2 Overview of the current investigation**

This study was initiated to investigate the incremental validity of four language assessments used to measure academic readiness among first time entrants to universities in South Africa. These measures included the National Benchmark Test in Academic Literacy (NBT AL), the Placement Test in English Second Language Advanced Level (PTESLAL), the Test of Academic Literacy Levels (TALL) and the Grade 12 English examinations. Except those for TALL, scores on these

assessments were already available from the South African university where the study was conducted. TALL was administered to a sample of first year students at this university, and the results were subsequently provided by the test owner. The ability of each of these assessments to predict the participants' end of first year performance on its own was first investigated. This was determined by computing correlational and linear regression analyses of the participants' scores on these assessments and their average scores at the end of their first year of study in 2012 and 2014. Secondly, the ability of these assessments to add to the predictive ability of the best predictor of these scores was also investigated. The statistical procedure used to determine this was the multiple regression analysis.

My hypothesis for the study was that of all the four predictors, TALL was the one that was likely to possess incremental validity. The main reason for this anticipation was that none of the other three assessments had the various aspects of their quality as empirically and transparently studied as those of TALL. The results of the statistical analyses carried out for the study revealed that Grade 12 English was a better predictor of first year academic performance than the other predictor assessments involved. Also, it turned out that the NBT AL had neither predictive nor incremental validity for the participants when compared with PTESLAL and Grade 12 English at the end of 2012. The results also revealed that PTESLAL also showed evidence of neither predictive nor incremental validity for the same participants at the end of the same year. In addition, these results revealed, however, that this test had both predictive validity on its own and incremental validity for a different group of participants when compared with TALL and Grade

12 English at the end of 2014. Finally, the results showed that TALL possessed both predictive validity on its own for the participants referred to above and incremental validity alongside PTESLAL and Grade 12 English at the end of 2014. Also, TALL showed evidence of possessing predictive validity on its own for this group of participants as well as for a larger group of the 2014 cohort of participants, and incremental validity in relation to Grade 12 English.

### **6.3 Recommendations**

My first recommendation from this study relates to Grade 12 English, which is that this assessment should continue to be used as part of Grade 12 results for making admission decisions by universities in South Africa. This recommendation is made on the basis of the finding of the present study as well as those by Myburgh (2015), Van Dyk et al. (2013), Van Rooy and Coetzee van Rooy (2015) and Van Dyk (2015) that Grade 12 English examinations in particular, and Grade 12 results in general, predict academic performance better than other assessments used for this purpose by South African universities. These results should, however, not be the only means employed for making such decisions. The shortcomings of Grade 12 English assessments that were identified by Weideman, Du Plessis and Steyn (2015) and Du Plessis, Steyn and Weideman (2016) and that were dealt with in Chapter Five provide strong support for my recommendation that performance on this assessment in particular and Grade 12 results in general should not be the only source of access decision making by universities.

The need for supplementary information to that provided by these results gains further impetus from an Umalusi (2012) report which revealed that performance on

the Grade 12 home languages examinations had not been comparable across the home languages taught at school over the years. In the observation of Du Plessis and Du Plessis (2015: 217), for example, “learners who offer English and Afrikaans at HL level score lower than those who offer other languages at this level”. Weideman (2016: 5) articulates this challenge further in his point that “a Tshivenda candidate would have a one out of 500 chance to fail, while candidates in other languages would have to face up to 60 times greater possibility.” The explanation for this partly rests in the revelation by the content analysis of the November 2012 Afrikaans, English and Sesotho language papers by Du Plessis and Du Plessis (2015: 219) that the construct supposedly measured by these examinations lacked clarity and that the scoring of these papers lacked scalar equivalence in the way marks are allocated. Lastly the lack of equivalence in the Grade 12 examination system has also been evident between performance on language and non-language subject examinations. In the observation of Weideman (2016: 5), “when the home languages results are compared to those of other subjects, there is no doubt that their averages are exceptionally high.” The variation in performance on the Grade 12 examinations highlighted above indicates that the results from these examinations are unfair and lacking in validity and that they should therefore, as I recommend above, be used alongside other sources of information for making student admission decisions.

The second recommendation I would like to make is that tests of academic literacy be used as additional sources of information to the Grade 12 English results to assess the readiness of university applicants to cope with the demands of academic

education in the medium of instruction. As indicated in Chapter Five of this study, Van Rooy and Coetzee Van Rooy (2015) found that academic literacy instructional interventions predicted academic performance better than Grade 12 results in general. These interventions and their assessment counterparts should logically be informed by the same construct of academic literacy. This makes tests of academic literacy the most pertinent source of assessment for the placement of first year students in such interventions. This is lent further credence by the consideration that the two tests of academic literacy investigated in this study are, as shown in Chapter One, informed by constructs of academic literacy that are more focused on the ability to cope with the discourse demands of university education when compared to that of Grade 12 English examinations, which as pointed out in the same chapter, focus on assessing a differentiated language ability.

The necessity for using tests of academic literacy as additional sources of information regarding academic readiness becomes more pertinent in view of findings by several studies emanating from the Alternative Admissions Research Project (AARP) of the University of Cape Town (UCT). The first of these studies was by Visser and Hanslo (2005), which focused on the predictive validity of a test of academic literacy known as the Placement Test in English for Educational Purposes (PTEEP), the predecessor of the NBT AL and a component of the battery of tests developed by this project to generate alternative information regarding the academic readiness of students applying for admission to UCT. Visser and Hanslo (2005) used scores on the PTEEP to track the survival and drop-out rates of

students from both former Model C or ex-House of Assembly (ex-HOA) and former Department of Education and Training (ex-DET) schools at UCT from 1995 to 2002. The former type of schools included those that were well resourced and mainly for Whites while the latter type comprised those that were under-resourced and were mainly attended by Blacks. In the words of Visser and Hanslo (2005: 1163), the most powerful advantage of the survival methodology they used in their study “is that by constructing hazard models of students’ careers, one can investigate not only whether particular groups (e.g. stratified by Race and Gender) drop out, but also when they are most likely to do so”. As Visser and Hanslo (2005: 1163) further argue, the survival methodology enables one to respond to the following questions: “Are students more at risk of leaving during particular stages of their careers? Does the profile of risk differ among groups? To what extent do assessment instruments predict the risk of dropping out?” A further advantage of this methodology is that it also allows the researcher to include students who do not necessarily drop out for academic reasons but also those who might discontinue their studies at a university for other reasons such as financial difficulties or transferring to another university (Visser & Hanslo 2005).

The results of this study were valuable in revealing that PTEEP top performers from both school backgrounds tended to survive longer at UCT as opposed to PTEEP bottom performers from the two school backgrounds whose attrition rate was high (Visser & Hanslo 2005). This means that the test was a better predictor of academic performance for both these groups than the school leaving examination. Secondly, the results also showed that the probability that ex-DET students would

drop out was higher than that for the ex-HOA group. This means that the PTEEP was “able to provide useful additional information regarding risk of exclusion especially in the ex-DET group of students, where top PTEEP performers clearly display a lower likelihood of being excluded compared to the bottom PTEEP performers” (Visser & Hanslo 2005: 1174). Thus, the ultimate value of the results of this study lies in its ability to demonstrate that unlike school leaving results, the PTEEP was able to provide differential predictive information for students from historically advantaged and disadvantaged backgrounds. This finding is subsequently confirmed in a study by Cliff, Yeld and Hanslo (2003: 9) which found that performance on the PTEEP seemed “to make important contributions towards explaining variation and predicting performance at the end of first year (along with those factors relevant to the group as a whole and home language and school leaving English)” for Black students, and more than it did for White students. It is against this background of the distinction that this test could make between students with different demographics that Cliff and Hanslo (2005: 8) suggest that if tests of academic readiness “are to be regarded as providing alternate or complementary information to the school-leaving examination, differential levels of performance need to be considered in making selection decisions and in assessing readiness, and therefore the curriculum needs, of students from different educational backgrounds.” In a word: at different levels of performance on a test of academic literacy (in this case: top and bottom performers), such tests may well be able to yield more insightful results and stronger relations to academic performance. This is a point that deserves further investigation.



The need for tests of academic literacy to be used as additional sources of information on the academic readiness of first time entrants to universities in South Africa garners similar support from yet another study (Cliff & Hanslo 2005) also emanating from the AARP at UCT. The focus of this study was to investigate the predictive utility of a battery of tests of academic readiness developed by this project namely, the PTEEP, the Mathematics Achievement (MACH) test, the Mathematics Comprehension (MCOM) test and the Scientific Reasoning Test (SRT) in comparison with the conventional school leaving results. The key finding from this study was that unlike school leaving results, these tests “have the capacity to produce variation in performance (1) for a total pool of writers; (2) by educationally disadvantaged background as a special subset; and (3) by programme group to which applicants apply” (Cliff & Hanslo 2005: 12). Furthermore, the results revealed that scores on these tests tended to correlate better with performance on Higher Grade than they did with scores on Standard Grade subjects (Cliff & Hanslo 2005). In the view of Cliff and Hanslo (2005: 13), this suggests that the AARP tests investigated in their study “are more strongly associated with conceptually more demanding school subjects, which may imply that the tests are indeed aligned with the levels of conceptual demand that are likely to be placed on students in Higher Education”. The homogeneity of the groups of students in my study may, therefore, have constituted a barrier for tests of academic literacy, as well as other measures, to show their full utility. Certainly, in subsequent studies, it would be useful to consider more heterogeneous populations.

The need for additional academic literacy assessment to be used to augment the information provided by school leaving results has recently also been reinforced by yet another study (Fleisch, Schoer & Cliff 2015) also from the AARP of UCT. The focus of this study was to compare performance on the National Benchmark Test in Academic Literacy (NBT AL) as well as Grade 12 English First Additional (FAL) and Home Language (HL) examination results of a cohort of first year students entering the Bachelor of Education Degree programme at the University of Witwatersrand (Wits) in 2014. Firstly, the results showed that English FAL candidates tended to underperform on NBT AL when compared to English HL students. This was the case even though these students' average scores on these language examinations as well as their Admission Point Scores (APS) on the overall school leaving examination were similar. Secondly, the results of this study also revealed that high school leavers with English FAL scores tended to perform poorly as compared to those who completed English HL at school in all the subdomains of the NBT AL. The difference in performance between the two groups on four competencies assessed in the test, namely, cohesion, essential/non-essential, text genre, and vocabulary was around half a standard deviation while that in grammar, inferencing, metaphorical expressions and relations in discourse was almost one full standard deviation (Fleisch, Schoer & Cliff 2015: 170). Overall, the students' performance on the NBT AL showed that "neither APS nor English scores by themselves allow universities to identify students that need support" (Fleisch, Schoer & Cliff 2015: 169). This makes a test of academic literacy such as the NBT AL an appropriate instrument for identifying future difficulties to cope with the discourse demands of academic education. Fleisch,

Schoer and Cliff (2015: 169) have argued for the utility of a nuanced measure of this kind for the following reasons:

(a) the test is targeted at an assessment of entry-level students' capacity to cope with the demands of higher education study, ... (b) NBT AL allows for more meaningful comparisons between students that come from different educational backgrounds ... (c) NBT AL ... allows for targeted intervention programmes, which cannot be developed from results of a school leaving examinations, which do not reflect these nuances.

The results of the three studies dealt with above are valuable in the context of the observation by Cliff, Yeld and Hanslo (2003: 1-2) that the ability of a student to succeed at university rests considerably on that student's background in a very broad sense: "factors influencing success are a complex blend of cognitive, affective, motivational, dispositional, socio-cultural, economic and institutional variables." The results of the three studies have shown therefore that tests of academic literacy are potentially able to classify and provide differential predictive information about students as a function of, among others, their broader background. So far, no research evidence exists to show that school leaving results have been able to do this. It is for this reason that Cliff, Yeld and Hanslo (2003: 2) have observed that

In a country such as South Africa, for instance, school-leaving certification has had a particularly unreliable relationship with Higher Education academic performance especially in cases where this certification intersects with factors such as mother tongue versus medium of instruction, inadequate school backgrounds and demographic variables such as race and socio-economic status.

Visser and Hanslo (2005) conclude on this basis therefore that Higher Education is likely to identify students who are at risk more successfully on the basis of their performance on a test of academic literacy such as the PTEEP as opposed to when this is done on the basis of school leaving results alone. Cliff, Yeld and Hanslo

(2003: 2) also conclude on the same basis that using school results as the only basis for making admission decisions “cannot be done without serious possibility of excluding some talented students who have not had adequate opportunity to demonstrate their potential for Higher Education study on the basis of school results alone”.

As shown by the results of the present study, TALL seems to have better ability than the NBT AL to provide the additional information that could help minimise the possible and unwarranted exclusion of first time entrants to universities on the basis of school leaving results only. This suggests that if it was necessary to choose between using any of the two tests as a source of information additional to that provided by Grade 12 English in particular, and Grade 12 results in general, TALL would be the best choice. This seems to be the case particularly in view of the finding by Van Rooy and Coetzee Van Rooy (2015: 38) that while performance on the two tests seemed to be similar and that they therefore probably measured the same construct, TALL showed evidence of a larger inter-quartile range and standard deviation. As Van Rooy and Coetzee Van Rooy (2015: 38) further argue, this makes TALL “the more valuable measure ... because it distributes the scores over a larger range, and thus enables better partitioning into groups than the NBT scores”.

The third recommendation I would like to make based on the results of the present study relates to PTESLAL. The use of this test for access by the university where this study was conducted should probably be revisited. Not only did the study reveal that the test’s predictive relationship with first year academic performance

was questionable, it also signalled that the test was not a reliable measure for the two samples of participants used. The decision to use the test for admission by this university was probably taken by policy makers who are, as Van Dyk (2015: 162) has observed, often neither experts in the fields of testing and assessment nor in applied linguistics in general. This means that the information provided by the test might not be a valid source of decision making with regard to admitting students. The long term result of this is that the access decisions made on the basis of performance on this test might have negative implications for graduation rates at this institution, not to mention the potential for placing individual students at a disadvantage.

Several recommendations also need to be made on the basis of the results of this study with regard to NBT AL. The first of these is that the National Benchmark Tests Project (NBTP) should become more transparent about the national testing service it provides. Weideman (2006: 82) has defined the transparency of a test as the degree to which its designer makes information available about its “content and workings”. Rambiritch and Weideman (2016: 4) observe that “while testing experts have stressed the need for an open dialogue between test developers and test takers, for test takers to be able to ask questions about the tests and for test developers to take responsibility for their designs, this has not always happened in practice.”

Worth mentioning in this regard is that the testing service provided by the NBTP has been transparent in several but two critical ways. The first of the latter is that no sample tests of this project are currently accessible to the test takers for them to

acquaint themselves with such tests before taking them. In the words of Rambiritch and Weideman (2015: 16) “providing students with a sample of the test is one way of ensuring transparency. Very often what is most daunting about taking a test is the fact that the test taker does not know what to expect”. Secondly, as also pointed out in Chapter Five, evidence of transparency in the form of peer reviewed publications on the National Benchmark Tests (NBTs) in general has been very hard to come by.

It is necessary therefore that the NBTP considers subjecting both the tests and the research carried out on them to public scrutiny in all possible ways. One way of doing this would be to make the data sheets for these tests available to all interested bona fide researchers in much the same way as the data and results for TALL, TAG, TALPS (Test of Academic Literacy for Postgraduate Students) and other tests are made available by the Inter-institutional Centre for Language Development and Assessment (ICELDA). As pointed out in Chapter Five, the kind of scrutiny offered by writing for peers in the same field in particular can only benefit the quality of these tests and contribute towards their improvement. As Rambiritch and Weideman (2016: 17) argue, a test of the calibre of NBT AL “will constantly need refinement, and this is best done if it is also evaluated by others working in the same field”. This is particularly true in the current higher education atmosphere where poor understanding of the value of the NBTs by users has, in some cases, subjected them to unwarranted scepticisms and criticism. Being transparent about the content and workings of the NBTs should logically promote accountability on the part of the NBTP and this will, in turn, promote a better

understanding of the value of the tests. In educational testing, accountability means that test developers make “information about their tests available to those most affected” and “take responsibility for their designs” (Rambiritch & Weideman 2016: 6).

My second recommendation with regard to NBT AL is that the NBTP should increase efforts to collaborate in research projects with all universities that are participating in the project especially with regard to the predictive validity of this test for students at those universities. As pointed out in Chapter One, academic performance is the criterion informing the constructs of all the tests developed under the auspices of this project. It is important therefore that the degree of relationship between these tests and academic performance is established across the South African higher education landscape. This should enable the NBTP to know whether the tests are of any teaching and learning relevance and utility for these institutions. As also pointed out in Chapter One, tests always exist for a purpose and it is important that this purpose is theoretically and empirically justifiable, more so for the expected calibre of NBT AL.

Thirdly, I recommend that, in its current form, the item type used in NBT AL should be revisited. As indicated in Chapter Five, traditional multiple choice items have solely been used to assess test taker levels of academic literacy in this test. While it is not my intention to downplay the important role that this item type has played in the history of language testing in particular, I do need to point out, however, that techniques for language testing have always evolved alongside changes in views of language ability and the resultant ways of language teaching.

At the time when language ability was defined in terms of the ‘skills’ of reading, writing, listening and speaking, for example, the traditional multiple choice item was effectively used to assess these ‘skills’. As shown in Chapter One, however, language ability is now understood to be broader than a mere mastery of the ‘skills’ of reading, writing, listening and speaking. This means that while the multiple choice item can potentially still be of utility to academic literacy testing, some degree of innovation is necessary in the way this item type is now developed and used to accommodate this broader view of language ability. Also, in their current state in NBT AL, multiple choice items are mainly designed to assess reading and reasoning misconceptions (cf. Cliff 2015). One needs to point out, however, that evidence of acceptable levels of academic literacy might not solely reside in one’s ability to show that they are immune from these misconceptions. As can be seen in the definition of the construct of the NBT AL itself, academic literacy involves more than just expected or correct reading and reasoning conceptions.

My recommendation that the way the multiple choice item is employed in NBT AL be reconsidered garners further support from the suggestion by Weideman, Patterson and Pot (2016: 1) “for modifications and additions to the design of current task types in tests of academic literacy” which will “allow theoretically defensible ... design[s] of the tests and ... be useful to those responsible for developing further versions of these tests ...” The larger context of this suggestion is the observation by Weideman et al. (2016: 10) that the constructs of academic literacy underpinning both TALL and the NBT AL have “not been further investigated in close to a decade of use”. What is more, one of the principles of



language assessments identified by Weideman (2014) is that such tests should be differentiated. That means that they may profitably and productively employ more task types (or subtests) than the current NBT AL test. In fact, one may speculate that the higher reliability indices associated with the various versions of TALL, in comparison to those of other tests, may perhaps derive from their employment of a larger variety of subtests and task types. That may also be a useful further research question.

The final recommendation I would like to make relates to the difficulty level of the two tests of academic literacy namely, NBT AL and TALL, for the participants in the present study. All the mean scores computed for these tests were the lowest when compared to those for the other variables. This is an indication that these tests were the most difficult. This also suggests that the difficulty level of these tests was not as aligned with those of the other variables, including the outcome variable. It appears from this that these tests might be more appropriately challenging for students at traditional academic universities and not necessarily for those at universities of technology such as the one from which the data for this study were collected. If this is the case, it underlines the need for models of Item Response Theory (IRT) to be more frequently employed in the development of these tests, provided, of course, that data for sufficiently large numbers of test takers are available. As pointed out in Chapter Two of this study, one such model known as *Rasch* was once used to validate TALL. More use of models of this kind should therefore be considered for the development of this test.

It is not publicly known whether the development of NBT AL has involved any use of any IRT model to date. Given the large numbers involved, there is no reason why IRT should not be used in the NBT AL. It is worth recommending that the application of this theory to the development of this test also be considered, or expanded if already in use.

IRT enables the test developer to identify items that are appropriately difficult for the test taker at a particular level of the ability scale. It also makes it possible for one to identify the items that discriminate well at a particular point on this scale. From this information, the test developer is able to develop an item bank for all ability levels from which an informed selection of items can be made for the purpose of test assembly.

#### **6.4 Limitations of the study**

The first limitation of this study is that the sample size especially of the second data set i.e. TALL, PTESLAL, Grade 12 English and 2014 average scores, was smaller than one would have preferred. The reason for this was that the number of applicants to CUT who do not meet straight admission requirements on the basis of their Grade 12 results and who are therefore required to take PTESLAL is hardly ever large, the same or equal every year. In the case of the second data set, the sample size was further diminished by a few participants who had scores for this test but did not have one for either TALL or Grade 12 English, or both. This meant that random sampling, which would allow for the generalizability of the results of the study, was compromised.

The second limitation of the study was that one was unable to obtain data for all the predictor assessments for the two years. The results of the study would be more informative and insightful if the analysis was carried out based on parallel data from all these assessments for both these years. This shortcoming leaves the study open to the possibility that the samples used, especially in the two analyses where the two tests of academic literacy were involved, were different. One should hasten to add, however, that from the time concerns started to be raised about the low levels of academic literacy preparedness of the pool of applicants arriving at South African universities, neither have these levels convincingly been reported to have improved nor to have deteriorated. The difference in the predictive ability of these tests revealed by this study is therefore likely to have been a function of the quality of the tests themselves.

The third limitation of this study relates to the current time scale of academic literacy testing in South Africa. Although it is conventionally acceptable that tests taken early in a student's career ease their ability to predict as that student progresses through their years of study, this does need reinvestigation. As pointed out elsewhere in this thesis, academic literacy testing came into being against the background of the evident mismatch between what high school education equips learners with and what these learners are expected to be able to do on entry to higher education. As also shown by this study and others, high school results continue, albeit not satisfactorily, to be the best predictor of student performance at university. This makes tests of academic readiness ideal sources of additional information about academic readiness after students are admitted. In other words,

these tests should be seen to be fit only for use as placement rather than access decision making mechanisms. The tendency to administer particularly the NBTs in the year preceding admission has created ambivalence about what these tests can and cannot do. This is evident in the fact that some universities use these tests for access rather than for placement. The NBT performance report for the 2012 and 2013 intake cycle admits, for example, that “a total of 49 institutions, organizations and bursary awarding bodies currently participate in the NBTs. Of these, 11 institutions did not submit requests for NBT scores for the 2013 intake cycle, whereas 38 institutions and organization used NBT scores for admission and placement purposes” (NBTP 2013: 12). The tendency to use the NBTs for access is perhaps possibly further promoted by one of the documented purposes for the introduction of the NBTP, which is, according to Griesel (2006:4), “to provide a service to HE institutions requiring additional information in the admission and placement of students”.

The last limitation of this study is that while it is appropriate to focus on measuring academic language ability for assessing readiness for university education, this can never be isolated from other potential predictors of academic performance, to which I will return in the section below. In fact, language ability has been found to predict a small proportion of the ability to succeed at university. As McNamara (2004) has observed, validation studies involving academic language tests in the past were able to record regression coefficients in the region of .30 at most. As McNamara (2004: 769) argues, this

indicates that differences in language test scores account for only about 10 percent of the variance in scores in academic subjects, suggesting that language

plays a definite but limited role in the academic success of students in such settings – hard work, organization, and intelligence seem to be more powerful factors in predicting success.

Similarly, Kobrin, Patterson, Shaw, Mattern and Barbuti (2008) have observed that the maximum ability of language tests to predict academic performance is 10 percent. Van Dyk (2015: 174) argues therefore that “language should thus, and with good reason, be expected to explain a limited, but significant percentage of academic performance.” This means that tests of academic literacy such as those that were investigated in this study cannot solely be used for making access decisions. It is in view of this that Cliff, Yeld and Hanslo (2003: 5) have concluded that in the current “context of diversity of student intake and educational provision” all factors that can potentially influence future academic performance should be considered.

## **6.5 Suggestions for further research**

The first suggestion for future research is that studies of the incremental validity of the assessments investigated in this study should be carried out using bigger samples of data that should afford those who conduct such studies an opportunity to use sampling procedures that can generate results on the basis of which the external incremental validity of these assessments can be more definitely established. Such investigations should involve the use of scores that are concurrently generated for one academic year so that the possibility of any difference in the samples used is controlled for. Furthermore, such studies might well be linked to those in other domains on student preparedness that have been found to impact overall success, and which I deal with in the section below. This is very important in the context of

the on-going search by South African universities for a measure that can predict student success with some acceptable degree of validity. This is also critical for dealing with the widespread concern about the low graduation rates across the South African higher education landscape.

The feasibility of the incremental validity studies of the kind suggested above is evident in a recent one by Kobrin, Camara and Milewski (2004) whose focus was the incremental validity of the Scholastics Achievement Tests (SAT) I and II for students from different ethnic backgrounds in California and the whole of the United States. The predictor already in use in the case of this study was the High School Grade Point Average (HSGPA) and the outcome variable was First Year Grade Point Average (FGPA). Firstly, Kobrin et al (2004: 272) found that all three variables had predictive validity to different degrees and that when combined, the three tests possessed the highest predictive validity for most ethnic groups. As Kobrin et al. (2004: 272) explain, “the predictive validity using all three measures was usually higher across ethnic groups than the predictive validity of only two measures”. The second finding of this study was that SAT I added to the predictive power of HSGPA and SAT II for most ethnic groups. This suggests that “the SAT I offers an important increase in predictive validity over and above HSGPA and the three SAT II tests”. Thirdly, the researchers also found that SAT II predicted differently for different ethnic groups when considered alone. In this regard, Kobrin et al. (2004: 272) observe that “the validity coefficients for American Indian and Hispanic students are lower than those for Asian-American, black, white and ‘other’ students.” Kobrin et al. (2004: 272) conclude on the basis of all these results

that “it is better from a purely predictive validity standpoint to consider all three of these measures when making admission decisions, although in some cases a second test may not have a practical effect in admission.” This recommendation is an echo of one of those made in the previous chapter with regard to Grade 12 results and tests of academic literacy that are used by South African universities for access and placement decision making.

## **6.6 The low graduation output by South African universities**

Twenty one years into the democratic dispensation, South African universities still experience the challenge of high student drop-out and low graduation rates. As the Council on Higher Education (CHE) (2013: 15) has observed, the country’s graduate output has been “found to have major shortcomings in terms of overall numbers, equity and the proportion of the student body that succeeds”. The key source of this challenge is the widespread lack of readiness among applicants to these universities to cope with the demands of academic education. This has been labelled the “articulation gap”, the discontinuity or mismatch between the competencies that high schools leavers achieve and the academic demands they are required to meet for them to succeed at university (CHE 2013: 17). The literature on this topic has, in a very broad sense, attributed the gap to a complex interplay of political, socio-economic, emotional and academic factors. As Cliff and Hanslo (2009: 266) explain, the ability to succeed academically depends on “the quality of schooling of individuals or cohorts; the population group to which an individual belongs; the socio-economic status of individuals or groups; motivational and dispositional orientations of students, their approaches to learning; and so on”.

Cliff, Yeld and Hanslo (2003: 1-2) have similarly argued that the insight that research on the academic readiness of first entrants to the world of higher education has yielded is “that factors influencing [academic] success are a blend of cognitive, affective, motivational, socio-cultural, economic and institutional variables”. CHE (2013) has also identified these factors as being material, affective and academic in nature.

In particular, the socio-economic and academic sources of the articulation gap identified above links it directly to South Africa’s political history of apartheid. In the first place, the unequal and racially skewed distribution of financial resources that was promoted by the apartheid regime means that the majority of historically disadvantaged students come from poor families that are short of the finances necessary to enable them to access and ultimately succeed in higher education (CHE 2013). Indeed, the participation rates for White and Indian students in tertiary education in South Africa are still as high as those for developing countries while the rate at which African and coloured students have access to this education is still very low (CHE 2013). Similarly, while the high drop-out and consequent low completion rates remain a challenge that cuts across race, these are less the case for Whites and Indians than they are for Blacks and Coloureds. For the latter groups, South African higher education continues to be a low-participation and high-attrition system (CHE 2013: 52). Surely, this is not only politically and socially unacceptable, but also economically: the downstream effects of the wastage in any part of the education system on the country’s economic well-being must be addressed, for the sake of everyone.



While CHE (2015: 55) acknowledges that no research has been carried out on the impact of socio-economic history on the academic performance of university students from historically disadvantaged backgrounds, it does argue, however, that anecdotal evidence suggests “that many students either do not enter higher education, or drop out without completing their studies, because of lack of access to finance” (CHE 2013: 55). Moreover, the impact of material factors on graduation rates among these students has, however, been evident in the growing pressure that their need for financial aid has placed on the National Student Financial Aid Scheme (NSFAS) (CHE 2013). This has been the case, even though from the time this scheme was established in 1994, the state budget for it has grown tremendously, with, for example, a R5 billion allocation made in 2012 (CHE 2013). Logically, the funding difficulties facing these students are likely to impact negatively on their chances of access to and completing university education in time or even completing it at all. It is in the context of this situation, that CHE (2013: 56) has therefore suggested that “if NSFAS is to ... contribute to improving graduate output, the focus should not only be on increasing the total funding available but also on the effectiveness and adequacy of the funding for facilitating learning.” The way that these financial pressures came to the boil at the end of 2015 and the beginning of 2016 during the mass student protests of the #FeesMustFall campaign is a sharp reminder of how acute the challenges are of financing higher education.

Secondly, the apartheid regime promoted a racially segregated education system where public schools that were mainly attended by Blacks were under-resourced as

opposed to Model C schools which were mainly reserved for Whites and were heavily resourced (Visser & Hanslo 2005). This state of affairs still very much prevails in post-apartheid South Africa. Wealthy parents, most of whom are Whites, largely remain the ones who can afford to take their children to well-resourced English medium schools while the majority of Blacks parents cannot. This means that White students continue to be in a better position to access good quality education and that Black students are not. This situation has resulted in the academic under-preparedness of the majority of students from historically disadvantaged backgrounds entering institutions of higher learning in recent years. The under-preparedness “takes different forms in different subject areas but the common feature in all settings is that what the students know and can do – attainments that were good enough to gain them entry into higher education - do not match the expectations of the institution” (CHE 2013: 57). In a generic academic literacy sense, this means that the students “have not been adequately prepared for, nor can they be expected to successfully negotiate the demands of conventional language, learning and thinking required of them, particularly in the absence of curriculum and learning support” (Cliff, Yeld & Hanslo 2003: 4). Academic preparedness is the essence of ultimate success at university. As CHE (2013: 57) puts it, “formal learning depends on whether students can and do respond positively to the educational process in higher education”.

The high drop-out and consequent low completion rates are further compounded by affective factors that Cliff and Hanslo (2009: 266) have described as “motivational and dispositional orientations of students” and “their approaches to learning”, and

which CHE (2013) has identified as additional impediments to successful academic performance at South African universities. In support of the foregoing, CHE (2007: 38-39) has argued that in “South Africa, Academic Development experience has indicated that benefits of well-designed educational interventions can be neutralized by lack of motivation, anxiety about personal or financial circumstances, or alienation from the institution.” It is in line with this view that CHE (2013: 57) has argued for the necessity for interventional efforts to deal with student under-preparedness to “extend beyond formal curriculum into the provision of psychological and social support, and of opportunities for students to engage actively with their institutions and environment in a variety of ways.”

## **6.7 Conclusion**

This chapter begins by providing a summary of the focus and results of this study. It then moves on to deal with the limitations of the study and make suggestions for future studies on the basis of the results of the current study. Finally, the chapter focuses on a discussion of the challenge of low completion rates that is faced by South African universities and the factors responsible for this.

The next chapter will focus on a discussion of the implications of the results of this study for current theories of validity, and what those implications mean for the future design and development of academic language tests and courses.

## **Chapter 7: Implications of the study**

### **7.1 Introduction**

In order to set the scene and create the context for the present study, current theories of test validity were explored in Chapter Two. In the present chapter, these theories are briefly outlined, and the implications of the results of this study for the theories are dealt with. Subsequently, the implications of these results and those of the literature reviewed in the study, in particular for the validity of tests of academic literacy, are discussed. Finally, the chapter deals with the implications of the results of the study for the validity of courses of academic literacy.

### **7.2 Theories of test validity**

#### **7.2.1 The traditional view of test validity**

Traditionally, validity has been defined as the degree to which a test measures what it purports to measure. Viewed from this perspective, validity is a function of the ability of a test to produce objective results on the basis of which inferences about test taker ability can be made. In other words, from the traditional perspective, a test is valid if it measures what it is intended to measure and evidence for this resides in the objective scores that such a test can produce. This means that the validity of test scores depends on the validity of the test that generates those scores in the first place and that no claim can be made therefore, about the validity of one at the exclusion of the other. As indicated in Chapter Two of this study, this view has received support from scholars like Davies and Elder (2005: 279) who have argued that

... through acquiring over time, and through repeated validation arguments, an adequate reputation, any test must eventually present a principled choice to

those wishing to use it, and that choice can be attributed to nothing else than its known validity.

As further shown in that chapter, this view has also drawn support from Borsboom, Mellenbergh and Van Heerden (2004: 279) who have similarly argued that it is rational for one to argue that a test that has been used many times and for the same purpose can rightly be judged to possess validity. In the words of Borsboom et al. (2004: 279), this makes it possible for one to “speak of the validity of that particular test – as a characteristic of it”.

Secondly, the traditional view of validity makes a distinction between three types of validity namely, content, construct and criterion-related validity.

As explained in Chapter Two of this study, the first of these, content validity, refers to the extent to which the type of tasks used in a test are a representative reflection of those that test takers will be required to perform in a real life situation. In language testing, the real life situation referred to above is now commonly known as the Target Language Use (TLU) domain (Bachman & Palmer 1996). Practically, a test possesses content validity if its specifications and task types are judged to align with the ability the test taker is expected to demonstrate in a particular TLU domain.

Construct validity, in turn, refers to the theoretical defensibility and justification of the ability a test is intended to measure. This means that tests are developed on the basis of constructs which must be defended and justified with reference to a theory that underpins the ability they purport to measure.

Lastly, criterion-related validity refers to the ability of a test to show evidence of association with other criteria that are judged to be informed by constructs that are related to the one underpinning the test being validated. Two types of this kind of validity are known as concurrent and predictive validity. The former relates to the association a test has with another one that is administered around the same time, while the latter involves the relationship between performance on a test and some criterion to be administered in future.

### **7.2.2 Messick's view of test validity**

The traditional view of validity presented in the section above is on the opposite end of another perspective, wherein validity is regarded not merely as a quality of the scores that a test produces and of the test itself, but resides rather in the interpretation (inferences) of scores. This view was originated by Messick (1980, 1989) and is in the view of Weideman (2012: 1), customarily held up as the culmination of the meaning of validity. Messick (1980: 1023) has defined validity as “an overall evaluative judgement of the adequacy and appropriateness of inferences drawn from test scores”. Messick (1980: 1013-1014) expresses this view lucidly in his argument that “questions of validity are questions of what may properly be inferred from a test score; validity refers to the appropriateness of inferences from test scores or other forms of assessment.” In these statements, we can already note two issues: first, Messick makes validity dependent on interpretation; second, that the terms ‘properly’ and ‘appropriateness’ are introduced. These are two key terms in any reading of Messick and we should note that ‘properly’ is used for the outlawed term ‘validity’, but means the same:

‘validly’, ‘adequately’ or ‘legitimately’, ‘with equal force’. This view has received support from others (e.g. Kane 1992; Bachman & Palmer 1996). Consistent with Messick’s (1980, 1989) concept of validity, Kane (1992: 527) has defined validity as “the interpretation assigned to test scores rather than with the scores or the test”. Bachman and Palmer (1996: 21) have similarly defined validity as “the meaningfulness and appropriateness of the interpretation that we make on the basis of test scores”. All of these have been elaborations – though with slight modifications – on what is now the current orthodoxy in validity theory.

Furthermore, the traditional distinction made between the types of validity outlined above is also some way removed from what Messick (1980, 1989) and others (e.g. Bachman & Palmer 1996) regard as validity. The conceptual difference is evident, first, in Messick’s further view that the essence of test validity is construct validity, which he, as we have noted, defines as “an overall evaluative judgement of the adequacy and appropriateness of inferences drawn from test scores” (1980: 1023). Messick (1981: 9) argues for his construct-driven concept of validity in the following terms:

Since construct validity is the evidential basis of test interpretation and since the imputed meaning of test scores is critical for appraising the potential social consequences of proposed test use, it would seem to follow that construct validity is as basic from an applied point of view as it is from a scientific one. Thus in education and psychology, not just scientific measurement but all measurement should be construct-referenced.

The other two traditional classifications of validity, namely content and criterion-related validity, are in the view of Messick (1980: 1989) merely sources of evidence for his overarching idea of construct validity. Messick (1980: 1014) argues that the unwanted consequence of compartmentalizing validity into the three

types is that “test users focus on one or another of the types of validity, as though any one would do, rather than on the specific inferences they intend to make from scores” and that “there is an implication that once evidence of one type of validity is forthcoming, one is relieved of responsibility for further enquiry.” Messick (1980: 1014) argues therefore that conceptual clarity is possible if definitions of content validity focus on describing their intent and character, such as “content relevance and content coverage rather than content validity”. Similarly, Messick (1989) rules out criterion-related validity as a type of validity because in his view, it

relies on selected parts of the test’s external structure. The interest is not in the pattern of relationships of the test scores with other measures generally, but rather is more narrowly pointed towards selected relationships with measures that are criterial for a particular applied purpose in a specific applied setting. (Messick 1980: 17)

What Messick has sought, in other words, was a unifying view of validity. It is achieved, in his case, by promoting construct validity to prime position. It is in view of this that Kane (2006: 21) has observed that in his unitary approach to validity, Messick relegates content validity “to a subsidiary role in supporting the relevance of the test tasks to the construct of interest,” and that “he treated the criterion model as an ancillary methodology for validating secondary measures of a construct against its primary measure”. Clearly, Messick’s further argument is that validity cannot be distinguished into different types. As Rambiritch (2012b: 114) observes, however, while these can be compartmentalised as different kinds of evidence, the points associated with each of these as kinds of validity remain important ones, and their distinctiveness is blurred by calling them all ‘validity’.



In their own way, Bachman and Palmer (1996: 17) also adopt a ‘unitary’ approach to test validation by arguing for their overarching concept of ‘test usefulness’ as the “most important consideration in designing and developing language tests.” Weideman (2012: 3) observes, however, that Bachman and Palmer’s (1996) approach is a fall back on Messick’s concept of validity in that like him, they define construct validity as “the meaningfulness and appropriateness of the interpretations that we make on the basis of test scores” (Bachman & Palmer 1996: 21). Weideman (2012: 3) further identifies a similar unitary approach in Kunnan’s (2000: 1) emphasis of test fairness as the primary consideration in test design and development. Weideman (2012: 2) argues that the three authors merely “re-interpret the ‘unitary’ condition of validity in slightly different ways”, by emphasising either usefulness or promoting fairness to the single most important quality of a test.

The last dimension of Messick’s theory of validity worth mentioning is that validation is an ongoing process which never ends. This view has received support from Bachman and Palmer (1996: 23) in their argument that validation is an “ongoing process” and that “we should not give the impression that a given interpretation is ‘valid’ or ‘has been validated’”. Similarly, Kane (2011: 4) is of the view that his argument-based framework for validity “is quite flexible in the sense that it does not specify any particular kind of interpretation or use for assessment scores, and invites assessment developers and users to specify their proposed interpretations and uses”. Where this flexibility leaves a test’s objective ability to

produce objective results, that can then be subjectively interpreted, is a question that is (momentarily) left unanswered.

### **7.3 Implications of the results of the study for theories of test validity**

The results of the present study have logical implications for the two perspectives of validity presented above. As pointed out several times in this thesis, the focus of this study was the incremental validity of four assessments of the academic language readiness of first time entrants into higher education in South Africa, namely the Grade 12 English examinations, the National Benchmark Test in Academic Literacy (NBT AL), the Test of Academic Literacy Levels (TALL) and the Placement Test English Second Language Advanced Level (PTESLAL).

As was revealed by the results of the study in Chapter Four, Grade 12 English appeared to relate better with the outcome variable, namely academic performance at first year level, than the other three predictor assessments investigated. It was also evident in that chapter that TALL possessed incremental validity in relation to Grade 12 English and that the other two assessments did not show any significant relationship with the outcome variable in the same way.

The differential relationship between performance on the predictor assessments and the outcome variable was evidence of the different ability of the former to predict the latter. In other words, while all these assessments are used for predicting the academic performance of students at first year level, they were able to do so differently as a function of the different ways in which they are designed to do this. This was the case even for TALL and the NBT AL, the two tests of academic

literacy currently used by universities to assess academic literacy and whose constructs, as shown in Chapter One, have similar origins. This was also the case for performance on TALL and PTESLAL, which showed evidence of multicollinearity in one of the data sets that were analysed. This is a source of evidence that the incremental validity (or lack thereof) of these predictor assessments rested first and foremost on their different designs and not necessarily on how the scores they generated were interpreted and used. In other words, the evidence for the extent to which the assessments were able to show that the scores they generate relate to academic performance or not was a result of their technical ability to produce results that can or cannot attest to this ability.

This means therefore that a test must first be valid - in the first, traditional interpretation of its validity as an objective quality - for it to produce valid results. This is a distinction that is still present in Messick's views, but then the use of the term validity, which has been abandoned, is merely taken up by terms such as "measuring properly/adequately". On the flip side, this also shows that a test that is lacking in validity cannot produce results that are valid. On the basis of this, one can argue that the view of validity that emphasizes it as a quality solely of the interpretation and use of test scores does not hold. It is for this reason that Davies and Elder (2005: 797) have argued that "it is not just a trick of semantics, therefore, to say that one test is more valid than the other for a specific purpose". Similarly, regardless of their attachment of validity to test score interpretation and use, Bachman and Palmer (1996) do, in fact, associate the test design requirement of reliability, a precondition for validity, with the test itself and not necessarily with

the scores on such a test or their interpretation. This is evident in their definition of reliability as “consistency in measurement” (Bachman and Palmer 1996: 19). In the same manner, they continue to speak of the utility or usefulness of a test as its overall measure of quality. Surely, there is a contradiction here, if all other qualities or characteristics of a test – its usefulness, reliability, practicality and so forth – are legitimate features of a test, but not its validity.

In view of the foregoing, Weideman (2009, 2012) has warned against confusing the objective capacity of a test to measure what it purports to measure with the subjective interpretation of the scores obtained on that test. Weideman (2009, 2012) argues that no amount of interpretation can add any degree of reliability or validity to a measure which is inherently lacking in these qualities. In the words of Weideman (2012: 4), associating validity solely with how test scores are interpreted and used “runs the risk of downplaying the quality of the instrument. No amount of interpretation can improve the measurement result (score) obtained from an inadequate instrument that gives a faulty and untrustworthy reading”. Weideman (2009: 10) further observes that the fruitlessness of the effort to ascribe validity to test scores is evident in the return of its association with a test in a different guise in statements such as a “test is ... a valid measure of the construct” (McNamara & Roever 2006: 109) and “items measuring only the skill or ability under investigation” (McNamara & Roever 2006: 81). As Weideman (2009: 10) further argues, this has also manifested itself in circumlocutions such as a “test accomplishing its intended purpose” (Messick 1980: 1025) and “tests purported to

tap aspects of a trait” (Messick 1989: 48, 50, 51, 73). It is in view of this that

Weideman (2012: 6) writes:

It seems to me that some of the critique of validity theory merely wants to say: if a test does what it is supposed to do, why would it not be valid? Surely a test that accomplishes its intended purpose has the desired effect i.e. yields the intended measurement? ... To say that a test is valid is therefore identical to saying that it has certain technical or instrumental power or force, that its results could become evidence or causes of certain desired (intended or purported) effects.

From the results of the present study, and the context of the assessment, it is also evident that the idea of separating out the three types of validity as was traditionally the case is more meaningful and defensible than the ‘unitary’ approach of Messick (1980, 1989) to validity which has, as pointed out above, also been pursued by others in different guises. The implication therefore is that, given the kinds of interpretation that would be given to such test results in the assessment context of the study, it is theoretically defensible to pursue the question of which interpretation would incrementally enhance the appropriate use of the results.

As it became evident in Chapter Four, Grade 12 English results turned out to be the best predictor of first year academic performance for all the groups of students involved in this study. In other words, Grade 12 English scores showed evidence of possessing the traditional predictive kind of validity better than the other three predictors. This was the case even though, as was shown in Chapter Five, that when Weideman, Du Plessis and Steyn (2015) investigated the construct validity of this assessment, they found that it was questionable because the construct of the assessment was not yet clearly defined. Furthermore, as also shown in that chapter, the assessment was not aligned with the views of language ability that inform the curriculum which Grade 12 English examinations aim to assess, a further indication

of the problematic construct validity of this assessment. The value of the study by Weideman et al. (2015), however, rests in their separating out and investigating a test design principle namely, construct validity, without elevating it over or subsuming it under others. The study by Weideman et al. (2015) enabled them to reveal the possible construct validity shortcomings of the Grade 12 English examination, while the present one has revealed that the same examination possesses better predictive validity than the other three predictors used.

This underlines the need for distinguishing clearly between all the concepts that inform test design and to investigate the quality of each, if necessary, on its own. Subsuming the traditional content and criterion-related types of validity under construct validity as Messick (1980, 1989) does runs the risk of relegating them to the margins of test design and development and in the process disregards the equally important role they may be able to play in the overall validation of a test. In addition, it would be unreasonable, therefore, to argue that by virtue of solely having demonstrated evidence of predictive validity in this study, Grade 12 English can automatically be accorded equal validity from a construct perspective. Instead, this kind of validity would have to be investigated separately in studies like that of Weideman et al. (2015), in order for this judgment to be made. This underlines the importance of dealing with all aspects of test quality as related but distinct parts of a coherent whole and raises questions about the notion of validity as a ‘unitary’ concept in which construct validity encapsulates all other qualities of a test. I shall return below to a potentially more sophisticated modification of this perspective, in the principle of test design that requires the developers of a test to make a

systematic argument that brings each disparate quality into a coherent whole, while distinguishing among such qualities.

The same argument can be made on the basis of the results of this study with regard to the PTESLAL. As pointed out in Chapter One, PTESLAL is a test of English Proficiency used for access at the university where this study was conducted. As a test of English proficiency, PTESLAL has been declared to possess content validity by its developer, the Human Sciences Research Council (HSRC 1991: 19). If this is indeed the case, one cannot necessarily conclude that the test possesses construct validity also. Such a conclusion will have to be a result of an investigation of this kind of validity as a distinct aspect of the overall validity of this test. The results of this study have shown that notwithstanding judgement by its developers that it has content validity, it was inconclusive in the present study, whether the test possessed incremental validity for first year students at that university. The reason for this, as pointed out in Chapter Five, is that in the two of the data sets analysed in this study in which scores on this test were involved, on the second occasion of these analyses, PTESLAL possessed incremental validity and on the first occasion, it did not. This finding would not be possible if this study focused on construct or content validity, because predictive validity is an aspect of test quality that is distinct from its content and construct validity. In other words, PTESLAL's ability to predict academic performance does not automatically equate to its content or construct validity. This suggests therefore that part of test validity may be a culmination of efforts to investigate the content and criterion-related validity of a test without subsuming them under construct validity in a 'unitary' approach. In other words,

the fact that the content of PTESLAL has been validated (HSRC 1991) but that its predictive validity was evidently inconclusive in the present study contradicts the ‘unitary’ conception of validity where content validity like the criterion-related type, is downplayed as a quality of tests, and one that should not be considered on its own, but only as subsidiary to construct validity. Kane’s (2006: 19) definition of content validation, for example, underscores its distinctness from other types of validity:

a content domain is outlined in the form of a test plan or blueprint, which may involve several dimensions (e.g., content per se, cognitive level, item type), with different numbers of items assigned to each cell in the plan. The items are not samples from the domain; they are created to match the test specifications, and to the extent that they do, they may be considered to be representative of the content domain described by the plan.

This is a description unique to content validation and which cannot be used interchangeably with either construct or criterion-related validation. Once again, this underlines how distinguishing between these concepts can help clarify their relationship and the distinct role that they should play in overall test validity.

It is possible, in addition, to pursue this argument further on the basis of the results of the present study with regard to the NBT AL. These results showed that in the analysis in which scores on this test were used, no evidence of incremental validity for these scores could be generated. This was the case even though convincing arguments (cf. Cliff & Yeld 2006; Cliff 2015) have been made for the construct and content validity of this test. As demonstrated in Chapter One of this study, the construct of this test draws heavily on applied linguistic theories that promote a broad, integrated and contextual view of language ability. As also pointed out in Chapter Two, the validity of this construct has recently been attested to by a study



(Cliff 2015) focusing on the teaching and learning implications of performance on the test. As was further pointed out elsewhere in this thesis, the content of this test derives from a broad consultation of expertise in the higher education sector on what the content of the test should be. It does not follow, in other words, that the incremental validity shortcoming of this test as revealed by the results of this study is attributable to its content or construct validity. This shows, once again, that the content, construct, and criterion-related validity of a test are related but distinct qualities of tests, each of which merits research attention in its own right. Thus, getting to the bottom of the poor incremental predictive ability of the NBT AL revealed by this study will require that all aspects of the validity of this test are investigated in a way that accords each of such aspects attention on its own and does not lump them all together under construct validity.

Finally, an argument to justify the distinction traditionally made between construct, content and criterion-related validity can also be made on the basis of the results of this study with regard to TALL. As shown in Chapter One of this study, TALL, like the NBT AL, is based on a construct of academic literacy which is also informed by applied linguistic views that promote a broad, open and an integrated understanding of language as a means of communication. This construct as well as the content of the test is also the outcome of a broad consultation of academics on exactly what should be measured in a test of academic literacy of this kind. A case for the content and construct validity of this test can therefore also be made. Furthermore, the construct validity of TALL has also been shown to hold through statistical studies that involved inter-correlational and factor analyses of

performance on it (cf. Weideman 2009; Van der Walt & Steyn 2008). As shown in Chapter Four of this study, TALL was the only one of the four predictor assessments investigated that showed evidence of possessing incremental validity. This result was generated by a study that focused on incremental validity only, the results of which confirm that this is a test quality that should and can be investigated separately from its content and construct validity. Acknowledging that these are distinct features of test design and investigating them as such offer the benefit of helping one ensure that each of them contributes to the overall quality of a test.

At the same time, not only has Messick (1989: 9) argued that “to speak of validity as a unified concept is not to imply that validity cannot be differentiated into facets”, he is also of the view that “the distinction introduced may seem fuzzy because the facets of validity are not only intertwined but overlapping”. Rambiritch (2012b: 118-119) observes, however, that the use of the terms ‘unify’ and ‘unifying’ is in fact, tantamount to saying that these concepts are the same:

If the concept of validity is a unified one, then it would make sense to see everything under that concept as being or meaning the same. If content, criterion and face validity are unified or the same as not varying it would potentially make sense to use any one type to validate the test – they are after all, uniform or in Messick’s words ‘unified’.

As distinct elements or components of a test, an alternative view may be to see the different kinds of validity that have been traditionally distinguished as components of a test that need to be thoroughly and systematically examined. I shall return below to how that may be accomplished.

### 7.3.1 Weideman's framework for applied linguistic designs

The foregoing argument becomes forceful in the context of the framework for the “responsible agenda for applied linguistics” that Weideman (2006, 2007) proposes. Weideman (2014: 1) defines applied linguistics as a “discipline of design: It solves language problems by suggesting a plan, or blueprint, to handle them.” With this definition as the starting point, Weideman (2009) articulates a framework of design principles for the three main applied linguistic artefacts, namely language courses, language tests and language policies (Weideman 2009). Weideman (2014) argues that his framework is applicable across all these artefacts and that any one of these (a course, a test or policy) has two terminal functions which he calls the qualifying and founding functions. He argues that the qualifying function of a plan presented in the form of a language course, language test or language policy resides in the technical aspect of its design. Furthermore, the analytical function of this plan has its foundation in the theoretical mode of experience (Weideman 2007). In the words of Weideman (2011: 102), “the technical design mode leads and qualifies the design of a solution to a language related problem, while the analytical dimension provides the foundational basis of the intervention.”

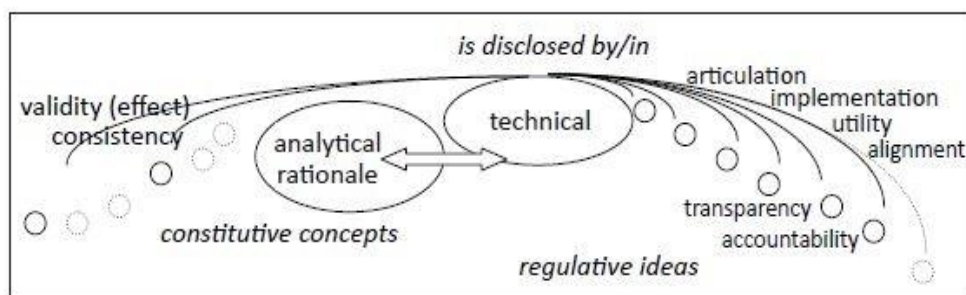
In language testing, the technical design referred to above leads or guides the measurement instrument, while the analytical function resides in the construct and test specifications that serve to provide a theoretical justification for the technical instrument. While the relationship between the two functions is reciprocal, the founding, analytical aspect provides a rationale for the technical but does not control it (Myburgh 2015; Weideman 2007). In the words of Rambiritch (2012b:

110) “the theory provides a rationale for the design but does not control it” and “the design therefore takes precedence, not the theory”. This is also aptly captured by Weideman (2009: 15) in the following words:

The relation between the leading, technical function of a test and its founding, analytical function is reciprocal. That is, in the design of an applied linguistic instrument, the technical imagination of the designer leads the whole endeavour, but at some point in the design process the development of the artefact must open itself up to critical modification and even correction by analytical and theoretical considerations and rational argument.

Weideman (2009) further argues that the technical design aspect of applied linguistic designs yields what he refers to as constitutive or necessary, and regulative or sufficient conditions that govern this design. In language testing in particular, the constitutive conditions include concepts such as reliability, validity and theoretical justification, while the regulative conditions include, among others, transparency, accountability and fairness (Weideman 2009). In **Figure 11** below the interaction between the analytical and technical as well as the constitutive and regulative conditions that govern the technical design of a plan proposed for a language problem are graphically captured.

**Figure 11: The constitutive and regulative conditions of applied linguistic designs**



(Weideman 2011: 110)

This framework, Weideman (2012: 10) argues, is beneficial in that it not only relates a test “to its intrinsic conventional conditions (reliability, construct and other

forms of validity, and so on), but also of specifying the so-called ‘social dimensions’ of tests (accessibility, accountability and fairness) as *inherent* requirements for responsible test design, not as add-ons.” Rambiritch (2012b: 110) has argued in support of the same framework on the grounds that it “serves to articulate coherently and systematically issues of responsibility and integrity – as well as to make allowance for other dimensions, such as the social and the ethical”.

The point being made in this observation is that each of these conditions is distinct and therefore contributes distinctly to test design. They can therefore not be subsumed under one ‘unitary’ concept of validity. Concepts such as technical adequacy, appropriateness, technical meaningfulness, interpretation of scores, utility, relevance, and public defensibility “must be conceptually distinguishable as constitutive technical concepts or regulative, concept-transcending ideas. And if they are distinguishable, that means they are conceptually distinct” (Weideman 2012: 8). This underlines the need for one to be critical of views that subsume concepts under others as Messick (1980, 1989) does with validity, because conceptual clarity is both essential and a necessary condition for test design. In the words of Weideman (2009: 11), “the requirement for conceptual acuity for the sake of an improved designed instrument is not served if we conflate concepts.” As pointed out earlier in this chapter, conflating these concepts means that the one that subsumes others receives more attention at the expense of those that are marginalized or regarded as mere evidence for the overarching one. This happens at the expense of the test involved and its potential improvement in quality, because, as has been argued so far, each of the constitutive and regulative

requirements has a distinctly important contribution to make towards the overall psychometric and social qualities of a test.

#### 7.4 Reinterpretations of Messick’s view of validity

The concerns raised above about Messick’s (1980, 1989) unitary concept of validity have led to the need for the reinterpretation of this concept in general and his so called “Facets of validity” in particular. These facets are captured in **Figure 12** below.

**Figure 12: Messick’s “Facets of validity”**

	Test interpretation	Test use
Evidential basis	Construct validity	Construct validity + Relevance/Utility
Consequential basis	Value implication	Social Consequence

(Messick 1980: 1023)

The need for the reinterpretation of Messick’s theory resides particularly in the general observation that it has been difficult to operationalize and that it fails to address social concerns in language testing (Weideman 2009, 2012), and the use of scores. Kane (2011: 7) similarly observes that “the uniform model based on construct validity is elegant and conceptually rich and suggestive, but it is not easy to implement effectively, because it does not provide a place to start, guidance on how to proceed, or criteria for gauging progress and deciding when to stop.” The obvious result of this is a complication of the test design process especially for non-testing specialists who have, in recent years, become increasingly involved in the selection and use of tests. Rambiritch (2012b: 115) argues, for example, that

the present day emphasis on the importance of testing means that many professionals in a variety of fields have to play the role of test developer – such as the language teacher who wants to design a test to test the writing levels of her class, but has no formal training in designing tests. Her first step then would be to consult the literature available on the designing of language tests – leaving her with the daunting task of unravelling Messick’s concept of validity.

#### 7.4.1 McNamara and Roever’s reinterpretation of Messick’s view of test validity

McNamara and Roever (2006: 427) have provided their own reinterpretation of Messick’s matrix of validity in an attempt to make it “more manageable” and accessible for the purpose of practical test design. The reason for this exercise was, in McNamara and Roever’s (2006: 13) words, to clarify “the way in which Messick’s theory takes theoretical account of the aspects of the social dimension of assessment.” Having examined Messick’s “Facets of validity” presented in **Figure 12** above, McNamara and Roever (2006: 13) observe that “aspects of the social context of testing are more overtly present in the model, in the bottom two cells of the matrix”. To them, the concern that this raises is the “relationship of the fairness orientated dimensions of the top line of the matrix to the more overtly social dimensions of the bottom line, a question it could be argued that Messick never resolved and remains a fundamental issue facing our field” (McNamara & Roever 2006: 13). This is, in the view of Rambiritch (2012), an important point to raise in the context of Messick’s (1980, 1989) argument for an integrative and unified view of construct validity. As Rambiritch (2012b: 117) observes, a closer look at this matrix forces one

to admit that there is no close integration or unifying of different concepts. While Messick’s matrix asks us to consider questions about the social dimension of language testing, these questions have been relegated to the

bottom row of the matrix. The empirical and social still exist, but may in such a view continue to operate as separate entities in the field of testing.

In a word, Messick’s own conceptualization does not provide enough of an explanation of how the ‘facets’ of validity may be systematically integrated into a single, unifying, systematic argument. Having identified the shortcomings of Messick’s facets of validity, McNamara and Roever (2006: 14) have reinterpreted these facets in the form of the matrix captured in **Figure 13** below.

**Figure 13: McNamara and Roever’s reinterpretation of Messick’s matrix of validity**

	What test scores are assumed to mean	When tests are actually used
Using evidence in support of claims: test fairness	What reasoning and empirical evidence support the claims we wish to make about candidates based on their test performance?	Are these interpretations meaningful, useful and fair in particular context?
The overt social context of testing	What social and cultural values and assumptions underlie test constructs and the sense we make of test scores?	What happens in our education systems and the larger social context when we use tests?

(McNamara & Roever 2006: 14)

#### 7.4.2 Weideman’s reinterpretation of Messick’s view of test validity

In the observation of Weideman (2012), however, McNamara and Roever’s (2006) reinterpretation does not wholly resolve the conceptual impasse left by Messick’s (1980, 1989) view of validity. He therefore offers a third reading and reinterpretation of this view in the context of the framework for the design of applied linguistic instruments that he proposes. In so doing, Weideman (2012: 7) incorporates Messick’s defining concepts of ‘adequacy’ and ‘appropriateness’ to his facets of validity in an attempt to deal with the problem of a conflation of the concepts of test design that was referred to earlier. Weideman’s (2012) reinterpretation yields another matrix which is presented in **Figure 14** below.



**Figure 14: The relationship of a selection of fundamental considerations in language testing**

	adequacy of ...	appropriateness of ...
inferences made from test scores	depends on multiple sources of empirical evidence	relates to impact consideration/consequences of tests
the design decisions derived from the interpretation of empirical evidence	is reflected in the usefulness/utility or (domain) relevance of the test	will enhance and anticipate the social justification and political defensibility of using the test

Weideman (2012: 6)

On the basis of the matrix in **Figure 14** above, Weideman (2012: 7) observes that four claims can be made about the requirement for the design and development of language tests:

- The technical adequacy of inferences made from test scores depends on multiple sources of empirical evidence.
- The appropriateness of inferences made from test scores relates to the detrimental or beneficial impact or consequences that the use of a test will have.
- The adequacy of the design decisions derived from the interpretation of empirical evidence about the test is reflected in the usefulness, utility, or relevance to actual language use in the domain being tested.
- The appropriateness of the design decisions derived from the interpretation of empirical evidence about the test will either undermine or enhance the social justification for using the test, and its public or political defensibility.

From these and on the basis of the framework he proposes for applied linguistic designs, Weideman (2012: 9) articulates a list of guidelines for the design of tests that can either be classified as constitutive or regulative conditions or principles for test design:

- Systematically integrate multiple sets of evidence in arguing for validity of the test or course design.
- Specify clearly and to the users of the design, and where possible to the public, the appropriately limited scope of the instrument or the intervention, and exercise humility in doing so.
- Ensure that the measurements obtained and the instructional opportunities envisaged are adequately consistent.

- Ensure effective measurement or instruction by using defensibly adequate instruments or material.
- Have an appropriately and adequately differentiated course or test.
- Make the course or the test intuitively appealing and acceptable.
- Mount a theoretical defence of what is taught and tested in the most current terms.
- Make sure that the test yields interpretable and meaningful results, and that the course is intelligible and clear in all respects.
- Make not only the course or the test, but information about them, accessible to as many as are affected by them.
- Present the course and obtain the test results efficiently and ensure that both are useful.
- Mutually align the test with the instruction that will either follow or precede it, and both test and instruction as closely as possible with the learning.
- Be prepared to give an account to the users as well as to the public of how the test has been used, or what the course is likely to accomplish.
- Value the integrity of the test and the course; make no compromises of quality that will undermine their status as instruments that are fair to everyone, and that have been designed with care and love.
- Spare no effort to make the course and the test appropriately trustworthy and reputable.

It was in accordance with these conditions and their possible classification as either constitutive and regulative requirements for test design that Van Dyk (2010) conducted a study that focused solely on the constitutive requirements of language testing, namely the reliability, validity and construct defensibility of the TAG (Toets van Akademiese Geletterheidsvlakke), the Afrikaans version of TALL. It is also for the same reason that Rambiritch (2012a) completed a study focusing solely on the regulative conditions of language test design, namely accessibility, transparency and accountability with regard to the Test of Academic Literacy for Postgraduate Students (TALPS). Finally, Van der Walt and Steyn (2008) conducted a study that focused on validating the TAG by considering both the constitutive and regulative conditions separately as proposed in Weideman's (2009) framework. Van der Walt and Steyn (2008: 194) begin by acknowledging that validation is a process that involves collecting evidence for construct validity,

content validity, criterion related-validity and reliability. Van der Walt and Steyn (2008: 194) also argue that this process encompasses considering social aspects of testing that include test taker feedback, test consequences, test ethics, social responsibility, washback and the impact of test scores. In their study, Van der Walt and Steyn (2008) accord equal and separate attention to all these concepts and neither elevate any above the rest nor adopt a 'unitary' approach to this process, even though they do attempt to bring them together into one argument that benefits from multiple components. The same approach is used in the Van Dyk (2010) and Rambiritch (2012a) studies referred to earlier. What these studies do attempt, however, is to fulfil the criterion to "systematically integrate multiple sets of evidence ..." (Weideman 2012: 7) in arguing for the adequacy of the design of these tests.

The results of the present study also have implications for another aspect of Messick's (1989: 13) theory of validity, which is the claim that validation is an ongoing process which never ends. As shown by these results, on the three occasions that the predictive validity of Grade 12 English results was investigated, it was evident that this assessment possessed, albeit not satisfactorily, the ability to predict first year academic performance for the three groups of participants whose scores were used in 2012 and 2014. As was also evident from these results, TALL was consistently able to add to this validity for the two groups of the 2014 participants that were involved. This shows that it is possible for an assessment to be valid across separate administrations and different samples without necessarily having to be validated continuously. It is in view of this that Weideman (2009:

118) asks if it is “inconceivable that the process of producing evidence will confirm that, to the best of the test designer’s knowledge, the test has the desired effect, i.e. it yields certain objective scores or measurements?” Rambiritch (2012b: 118) raises similar questions on this issue:

Is it not possible that there could be an end to the process of validation? Should there not be a valid test at the end of the process?” ... Is it not acceptable to ask whether an instrument that has undergone a process of validation may be shown to be a valid test? Does the validation not demonstrate that it does what it was designed to do?

The questions raised above gain justification in the fact that currently, models of test design exist that are driven by the relationship between test taker levels of the ability measured by a test and the characteristics of test items. Such models include those that are informed by Item Response Theory (IRT). Cohen and Swerdlik (2010: 158) have described IRT procedures as those “that provide a way to model the probability that a person with X ability will be able to perform at a level of Y”. To this end, IRT models start from the premise that performance on a test item depends on the test taker’s level of the ability measured by an item and the characteristics of that item (Bachman 2004: 141). The item characteristics referred to above include difficulty, discrimination and pseudo chance or guessing. Depending on how many of these characteristics an IRT model focuses on to validate a test, such a model can either be a one, two or three parameter type (Bachman 2004). The very premise on which IRT models rests means that unlike those of Classical Test Theory (CTT), performance statistics on the tests designed on the basis of such models are neither test nor sample dependent. Bachman (2004: 139) explains how CTT models function in the following words:

On the one hand, the item statistics that we obtain are dependent on the particular group, or sample of test takers who take the test, while on the other hand, the test scores the test takers receive are dependent upon a particular set, or sample, of items that make up the test.

This means that the same test can produce different performance statistics for different groups of test takers and that alternate forms of the same test can produce different performance statistics for the same group of test takers. The argument for validation as an endless process therefore appears to hold for tests developed on the basis of CTT models. It means, in other words, that if every administration of the same test can yield different statistics as a function of the differences in the samples of test takers and the particular version of the test used, no generalizations about these statistics can be made with confidence. Such a test would, as Messick argues, need to be endlessly validated.

In contrast, the way in which test taker ability levels are estimated in IRT is “fundamentally different from traditional approaches such as classical test theory” (Gregory 2007: 111). The key difference between the two theories is that performance on the tests developed through IRT models is invariant. This means that a test taker’s level of a measured construct can be estimated from their responses to any set of items whose Item Response Functions (IRFs) are known (Bachman 2004; Cohen & Swerdlik 2010). An IRF is a mathematical equation that describes the relation between a test taker’s level of the ability being measured and the probability that they will respond in the keyed direction (Gregory 2007). In this sense, IRT enables the test developer to assemble item banks from which they can select items that match with different ability levels across a test taker population. This means that “items with appropriate difficulty levels corresponding to the trait

level of the examinee will be administered” (Gregory 2007: 111). The greatest advantage of IRT, however, is that it makes it possible for alternative versions of the same test to be linked. Reise, Ainsworth and Haviland (2005: 98) describe this advantage in the following terms:

In large scale educational assessment, item-parameter invariance facilitates the linking of scales from different measures (i.e., placing scores on a single, common scale), across students in different grade levels (e.g., third through sixth grade in the same school) and within a grade level (e.g., fourth graders in different schools).

This means that using IRT models in test development potentially facilitates the validation process to such an extent that it is not necessarily a process that never ends.

The argument here has focussed on one being able to examine a specific kind of validity – how useful a predictor of overall performance certain language tests may be – rather than having to do a full-blown validation of every aspect or component of the assessment, and on justifying such a limitation in focus theoretically. The justification offered here rests on the argument that, provided that the use of a test is clearly understood (as is the case with the tests examined in this study), and provided, further, that the context of its use is specified, a more limited examination of dimensions of the test, and the way that its results may be legitimately interpreted, are in order. Without such understanding and specification, a more limited approach to what is called ‘validation’ may well not be appropriate. What is more, this more limited focus does and should not detract from the expectation that, the more high stakes the assessment is, the greater the responsibility of the designers and users of language tests is to ensure that the tests are subjected to stringent scrutiny. That stringent scrutiny of the overall quality of a test is what is

usually attempted in a full-blown validation study, which indeed may be revisited and refined, and in that sense is never fully complete. But the rigours of test use and administration require that we do not always have the luxury to undertake such comprehensive validations. Should one decide to undertake such a more comprehensive validation process, the pioneering work in South Africa on what it may entail for these kinds of tests of academic literacy has been done, with the most prominent examples being Van der Walt and Steyn (2008), Rambiritch (2012a), Van Dyk (2010), and Myburgh (2015). In such studies, the systematic bringing together of evidence from multiple sources, and their integration into a single argument for the sake of demonstrating the qualities of a test, is the essence of the process.

However, validation as a comprehensive process should apply not only to test design, but also more widely to other applied linguistic artefacts. I shall refer below to language course and language policy design. Where all three work together, it enhances the chances and opportunities that the interventions will truly benefit those affected. In that sense, then, it is not merely about validation, but about overall responsible design.

## **7.5 Implications of the literature for the validity of tests of academic literacy**

The literature reviewed in this study also has a number of implications for the validity of tests of academic literacy in particular.

The first of these is that in the context of higher education in South Africa, such tests should demonstrate evidence of being able to measure academic literacy

differentially for students from different socio-economic and school backgrounds. As indicated in the previous chapter, studies on the predictive ability of tests of this kind at the University of Cape Town have revealed that they are able to provide a better picture of the possible survival and drop-out rates of undergraduate students from diverse backgrounds. While no study focusing specifically on this issue has been carried out on TALL, Van Rooy and Coetzee van Rooy (2015: 38) found that scores on this test have a “bigger interquartile range and standard deviation” and that this was an indication that the test “distributes the scores over a larger range, and thus enables better partitioning into groups”. As indicated in Chapter Six, a recent study by CHE (2013) has shown that in South Africa students from historically disadvantaged backgrounds are the most affected by under-preparedness for academic education and by consequent high drop-out and low completion rates. Among others, this study found, for example, that “very small proportions of African, Coloured and Indian students graduate in regulation time” and that “completion rates are especially low in Engineering and Science degrees” (CHE 2013: 43). This means that if the country is to improve graduation rates across all these races, the validity of the assessment measures used for both access and placement of students with different demographics will need to be refined to accommodate this difference. It might be necessary, in other words, for tests of academic language readiness to be judged to possess validity if they are able to produce evidence of predicting academic performance as a function of differences in socio-economic and school backgrounds as well.



The second implication of these analyses for these tests is that they must demonstrate evidence of measuring academic language as a unique ability to cope with the discourse demands of academic education in the language of teaching and learning. As indicated elsewhere in this thesis, the articulation gap between high school and university finds ultimate expression in the academic under-achievement, high drop-out rates and poor completion rates among those affected (CHE 2013). It is generally accepted that this gap is a result of, among others, university students' low levels of academic literacy. In a general sense, academic literacy has been defined as students' ability to

- make meaning from what they read;
- understand and interpret conceptual and metaphorical language;
- identify and track academic argument;
- follow discourse structure in text;
- make inferences about and extrapolate from what they read;
- demonstrate familiarity with and understanding of the conventions of visual and modal literacies, such as reading and interpreting graphs, pictures, flow-charts and diagrams;
- cope with basic numeracy

(Cliff, Yeld & Hanslo 2003; cf. Van Dyk and Weideman 2004).

CHE (2013) observes that difficulties to cope with these demands are faced by all students entering higher education throughout the world in general and in South Africa in particular. For students from poor academic backgrounds, however, this is further complicated by the inadequate ways in which high schools train them to approach texts, which in turn limits their epistemological access to knowledge (CHE 2013). The implication of this for the validity of tests of academic language readiness is that they must be able to show evidence of assessing general academic literacy as a factor in the overall readiness of students to succeed at university. In

other words, such assessments should be seen to possess validity if they are able to assess incoming students to cope with the discourse demands of university education in a general sense.

Another implication for tests of academic literacy in particular and those of academic readiness in general is the necessity for them to be validated from the perspective of test takers. This kind of validation is typically carried out by obtaining feedback about a test from the test takers. Bachman (2004: 276) raises two questions in support of the need for tests to be validated from this perspective:

To what extent are the processes that test takers use to answer a task typical of the processes that language users would employ in responding to similar tasks in the TLU domain? Are these processes included in our construct definition?

Questionnaires, interviews and other qualitative instruments such as verbal protocols can be used to obtain this feedback. As indicated in Chapter Two, Van der Walt and Steyn (2008) employed a post-test questionnaire to investigate the test takers' perceptions of the transparency of TALL. So did Butler (2009) and Rambiritch (2012a) for the Test of Academic Literacy for Postgraduate Students (TALPS), and Du Plessis (2012) for its second version. The value of the feedback generated from this exercise is that it enables the developers of the test to assess its validity from the perspective of the test takers. This implies that a test should, among others, be judged to be valid as a function of how the test taker perceives it.

The results of the present study with regard to the two tests of academic literacy investigated also have validity implications that relate to the levels of test difficulty for diploma and degree candidates at South African universities. As shown in Chapter Four of this study, the participants' performance on the two tests was

evidently the lowest, an indication that the tests were more cognitively demanding for them than the other assessments that were investigated. As pointed out in Chapter One, universities of technology such as the one where this study was conducted typically offer diploma programmes for which the admission requirements are conventionally lower than those at traditional academic universities where mainly degree programmes are offered. As also pointed out in Chapter Six, it is possible therefore that these tests might have been more difficult for the participants in the present study than they would have been for students admitted for degree studies. In Chapter Two, we observed that test difficulty is a factor in test validity because a test that is too easy or too difficult for a particular group of students impedes the possibility of making valid inferences about their levels of the ability being tested. As also pointed out in that chapter, Chapelle and Brindley (2002: 277) have rightly argued that when test “difficulty is interpreted in view of the construct that an item of a test is intended to measure, it can be used as one part of a validity argument”. The effort by the National Benchmark Tests Project (NBTP) to address the issue of validity as a function of test difficulty is evident in the different performance levels it has set for degree as opposed to diploma study. In **Table 65** below, these performance levels and how they should be interpreted are presented.

**Table 65: The Benchmarks for the National Benchmark Tests**

Proficient	100	<p>Test performance suggests that future academic performance will not be adversely affected (students may pass or fail at university, but this is highly unlikely to be attributable to strengths or weaknesses in the domains tested). If admitted, students may be placed into regular programmes of study.</p> <p>Degree: AL [64%]; QL [70%] MAT [68%]</p> <p>Diploma/Certificate: AL [64%]; QL [63%] MAT [65%]</p>
Intermediate		<p>The challenges identified are such that it is predicted that academic progress will be adversely affected. If admitted, students' educational needs should be met as deemed appropriate by the institution (e.g. extended or augmented programmes, special skills provision).</p> <p>Degree: AL [38%]; QL [38%]; MAT [35%]</p> <p>Diploma/Certificate: AL [31%]; QL [34%] MAT [35%]</p>
Basic	0	<p>Test performance reveals serious learning challenges: it is predicted that students will not cope with degree-level study without extensive and long-term support, perhaps best provided through bridging programmes (i.e. non-credit preparatory courses, special skills provision) or FET provision. Institutions admitting students performing at this level would need to provide such support themselves.</p>

(NBTP 2015b: 17-18)

For the three domains of the NBTs, namely Academic Literacy, Quantitative Literacy and Mathematics, the performance levels presented in **Table 65** above are an outcome of standard setting processes. “Standard setting is the methodology used to define *levels* of achievement or proficiency and the *cutscores* corresponding to those levels” (Bejar 2008: 1). For the NBTs, the benchmarks or cut scores are an outcome of judgments by panels of academics from all over the country and are revisited by such panels once every three years (NBTP 2015). The importance of establishing the validity of these benchmarks cannot be overemphasized. The

National Council on Measurement in Education (2010: 15) captures the importance of the validity of the standard setting process as well as the benchmarks it sets in the following words:

standard setting is more appropriately conceived of as a measurement process... Because standard setting is a measurement process, standard setting results should be evaluated using the same expectations and theoretical frameworks used to evaluate other measurement processes in education such as student measurement.

It is necessary therefore that the predictive validity of the NBT benchmarks for both degree and diploma programmes is established. The implication of this is that these tests can only be valid to the extent that these benchmarks are able to assist with the successful placement of students with minimal error. In the words of Bejar (2008: 1), unless the benchmarks “are appropriately set, the results of the assessment could come into question”.

The cut scores for TALL are set differently from those of the NBT AL. They neither classify test taker performance explicitly according to levels of proficiency, nor do they distinguish between degree and diploma candidates. Instead, test takers are classified according to their levels of risk to deal with the academic literacy demands of their studies and numbers ranging from 1 to 5 are allocated to these levels (Weideman 2011). This is depicted in **Table 66** below.

**Table 66: Levels of risk associated with scores on TALL**

Risk level	Interpretation
1	Very high risk
2	High risk/clear risk
3	Borderline (moderate risk)
4	Less risk
5	Little to no risk

(Weideman 2011: 107)

Initially, historical data from a test previously used by the University of Pretoria were used to determine the cut scores that are presented in **Table 66** above (Weideman 2011). These data were based on the fact that the earlier test was norm-referenced and that its scores were calibrated against performance by Grade 10 learners (Weideman 2011: 106). Performance on this test had shown that “over a number of years, those measuring at a level of language ability associated with that of Grade 10 learners or lower grades had stayed consistent at between 27% and 33%”. Determining the cut scores was therefore a result of, among others, the “experience already gained, developed further, and meticulously recorded in subsequent years” (Weideman 2011: 107). This notwithstanding, it remains important that the validity of these cut scores is established especially from the point of view of the possible differentiated ability levels of students pursuing degree as opposed to diploma studies. As was pointed out with regard to the NBT AL above, to the extent that these cut scores can predict test taker performance with minimal error, TALL itself can be judged to be valid.

## 7.6 Implications of the results of the study for course validity

The implications of this study for test validity also have relevance for the validity of academic language curricula. As indicated in Chapter One, the language assessments investigated in this study have been used for making access and placement decisions and this links them directly to the language curricula aimed at addressing the articulation gap that such assessments can potentially reveal. On the South African higher education landscape, language curricula of this kind have often been part of what are now commonly known as Extended Curriculum Programmes. These programmes aim to “provide additional curriculum time for foundational learning to enable students to develop sound academic and social foundations for succeeding in higher education. Extended programmes thus constitute a curriculum intervention designed specifically to address the articulation gap ...” (CHE 2013:18). Quantitative and qualitative studies have revealed that extended programmes have possibly been effective in reducing the articulation gap and have consequently helped in the improvement of completion rates among students from academically disadvantaged backgrounds (CHE 2013). The effectiveness of these programmes has at the same time, however, been negatively affected by “their marginal status in the sector, which has negatively affected their design, staffing and reach” (CHE 2013: 18). It is the design of the academic language courses within these programmes in particular for which the results of the present study have implications.

As pointed out earlier in this chapter, language tests, language curricula and language policy are the three prominent artefacts within the broader field of applied

linguistics (Weideman 2014). Applied linguistics has, in the words of Weideman (2014: 2), become a distinct discipline “through three prominent sub-disciplines that concern themselves with language designs and plans. These three, that deal with designed solutions for apparently intractable language problems, are language management, language instruction and language assessment.” Weideman (2009, 2012, 2014) argues that the framework he proposes for applied linguistic designs referred to above should apply to each of these three artefacts. It is in the context of this argument that Weideman (2014: 6) asks:

Can the design of one kind of applied linguistic artefact not perhaps be beneficially employed to inform that of another? Would comparisons of these designs not perhaps have reciprocal benefits? ... How much reciprocity is there in the realms of language testing, language course design, and language policy?

These questions point to the importance of validity to the design of all these three artefacts. Viewed from the perspective of the present study and that of the framework that Weideman (2014: 6) articulates, this means that the concept of validity should also be applied to a language course “so that we explicitly check whether the design of a course has been done as responsibly and carefully as a test”. It also means that if this process is carried out within the framework of applied linguistics that Weideman (2009) proposes, the design may conform to the constitutive and regulative conditions of applied linguistic designs that were referred to above. As indicated earlier in this chapter, these conditions should be conceptually distinct and preferably not be conflated. Weideman (2014: 7) pursues this view in the following words:

If the argument that design principles are common across different kinds of applied linguistic designs (language *courses*, language *tests*, language *plans*) is correct, this means that conceptually one should focus on the relationship between the two critical (foundational and qualifying)



functions, considering especially the principles that emanate from the technical function of designing, shaping, forming or planning.

In the context of the present study, this implies that, like their assessment counterparts, academic language courses should be valid from the point of view of the three traditional classifications of validity, namely, construct, content and criterion-related types. It also means that these types of validity can be investigated as distinct aspects of course validity. In academic language course design, construct validity should refer to the theoretical defensibility of the construct underpinning the course. Such a course would logically be designed on the basis of the construct of an academic language test used to determine the language needs of the targeted students. In other words, the construct of the two artefacts needs to be aligned if construct validity is to be attained in course development. As pointed out in Chapter One of this study, constructs of academic literacy are used both for test and course design (Patterson and Weideman 2013a: 107). The content validity of a language course refers to the extent to which the tasks designed for such a course are aligned with the construct that underpins it and are adequately representative of those that the test taker will be expected to perform in the TLU domain. To this end, courses of academic literacy should enable students to engage with the discourse tasks that are typical of those that they will need to perform efficiently in order to succeed at university study. Lastly, a language curriculum will possess criterion related validity of the predictive kind if evidence can be generated to show that performance in it relates predictively with future academic performance. It is this last expectation that features strongly in the minds of administrators who select tests

for either placement or access, and measure the effectiveness of courses that aim at developing language ability.

In line with the findings of the present study, construct, content and criterion-related validity all have an important contribution to the overall validity of a course and merit individual investigation that should not subsume them under one ‘unitary’ concept. A study of an academic literacy course at a South African university by Sebolai (2014) demonstrates the practicality and value of investigating each aspect of a course as a distinct contributor to its overall validity. Among others, this study focused on the conceptual design of this course (construct) and its task types (content). Following this investigation and using the construct of TALL as the basis, Sebolai and Huff (2015) report on a curriculum renewal process aimed at improving the construct and content validity of this course. Similarly, Weideman (2007) demonstrates ways to operationalize both the construct and content of TALL for the purpose of classroom academic literacy instruction. Lastly, Van Rooy and Coetzee van Rooy (2015) have investigated the predictive ability of, among others, courses of academic literacy and found that they demonstrated better predictive validity than the other predictive assessments that were used. All these studies attest to the practical value of separating out and investigating the constitutive and regulative conditions of course design in their own right and as separately important factors in the overall validity of a course. The value of investigating a course of academic language readiness from all these perspectives rests in the potential to improve the overall validity of such a course and enhance its impact on those who

enrol for it. To the extent that such a course is valid, it will have positive consequences for those taking it.

## **7.7 Conclusion**

This study is a contribution to the current debate on the role of language assessment in the success of university students. It builds on previous studies on the predictive validity of language assessments at a time when the importance of the search for valid measures of the ability to succeed at university cannot be overemphasised. Consistent with those of previous studies, the results of the present study confirm that language ability plays a limited but crucial role in predicting the ability of university students to succeed in their first year of study. These results also show that while different tests have been used to determine language readiness for academic study at South African universities, some of these tests can possess better predictive validity than others and that some can possess better incremental validity than others. In a way, this lends support to the traditional view that validity is a property of a test, as opposed to the view wherein validity is understood to be a function of how test scores are interpreted and used. In other words, the results show that tests do possess validity as a function of how efficiently they are developed to measure what they purport to measure and that the process of test validation can, therefore, at least be provisionally completed as long as the purpose and context of the validation are clearly specified. Furthermore, the results lend support to the traditional view of validity in which three main types of validity, namely construct, content and criterion-related validity are recognized. This is on the other end of the unitary approach to validity where construct validity subsumes

the other two types of validity. The limitations of this study notwithstanding, it is to be hoped that it will be a continuation point for future research on the role of language ability in the ultimate success of students at university and the exploration of the most feasible ways of validating tests of this ability.

# References

- Alderson, J.C., Clapham, C. & Wall, D. 2005. *Language test construction and evaluation*. Cambridge: Cambridge University Press.
- Alidou H., Aliou, B., Brock-Utne, B., Diallo, Y.S., Heugh, K. & Wolff, E. 2006. Optimizing learning and education in Africa – the language factor. A stock-taking research on mother tongue and bilingual education in sub-Saharan Africa. Paris: Association for the Development of Education in Africa (ADEA). Available: [http://www.Adeanet.org/biennial-2006/document/B3\\_IMTBLE\\_en.pdf](http://www.Adeanet.org/biennial-2006/document/B3_IMTBLE_en.pdf).
- Ayliff, D. 2010. “Why can’t Johnny write? He *sounds* okay!” Attending to form in English second language teaching. *Perspectives in Education*, 34(4): 455-467.
- Bachman, L.F. 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L.F. 2004. *Statistical analysis for language assessment*. Cambridge: Cambridge University Press.
- Bachman, L.F. & Palmer, A.S. 1996. *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.
- Barry, D. 2002. Language equity and assessment in South African education. *Journal for Language Teaching*, 36(1&2): 105 – 117.
- Bejar, I.I. 2008. Standard setting: What is it? Why is it important? *Listening. Learning. Leading*. Princeton, New Jersey: Educational Testing Service.
- Blanton, L.L. 1994. Discourse, artefacts and the Ozarks: Understanding academic literacy. *Journal of Second Language Writing*, 3(1): 1-16.
- Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H. & Krathwohl, D.R. 1956. *Taxonomy of educational objectives: The classification of educational goals*. Handbook I: Cognitive domain. New York: David McKay Company.
- Boughey, B. 2013. What are we thinking of? A critical overview of approaches to developing academic literacy in South African higher education. *Journal for Language Teaching*, 47(2): 25-42.
- Bourdieu, P. & Passeron J.C. 1990. *Reproduction in education, society and culture*. Newbury Park: Sage Publications.
- Borsboom, D., Mellenbergh, G.J. & Van Heerden J. 2004. The concept of validity. *Psychological Review*, 111(4): 1061 – 1071.
- Butler, H.G. 2006. A framework for course design in academic writing for tertiary education. Unpublished PhD thesis. Pretoria: University of Pretoria.
- Butler, H.G. 2009. The design of a postgraduate test of academic literacy: Accommodating student and supervisor expectations. In: Geldenhuys, J. (Ed.). *Assessing and developing academic literacy*, special issue of *Southern African linguistics and applied language studies*, 27(3): 291-300.

- Butler, H.G. 2013. Discipline-specific versus generic academic literacy intervention for university education: An issue of impact? *Journal for Language Teaching*, 47(2): 71-88.
- Cattell, R.B. 1946. *Description and measurement of personality*. New York: World Book Company.
- CETAP (Centre for Educational Testing for Access and Placement). 2016. The National Benchmark Tests: Academic and Quantitative Literacy (AQL)Test. Available: [http://www.nbt.ac.za/sites/default/files/NBT\\_AL\\_Teachers.pdf](http://www.nbt.ac.za/sites/default/files/NBT_AL_Teachers.pdf).
- Chapelle, C.A. & Brindley G. 2002. Assessment. In Schmitt, N. (Ed.), *An introduction to applied linguistics*. London: Arnold, pp. 267-288.
- Cliff, A.F. 2015. The National Benchmark Test in Academic Literacy: How might it be used to support teaching in higher education? *Language Matters*, 46 (1): 3-21.
- Cliff, A.F. & Hanslo, M. 2005. The use of alternate assessments as contributors to processes for selecting applicants to Health Sciences. Paper prepared for the Europe Conference of the Association for Medical Education in Amsterdam, Netherlands.
- Cliff, A.F. & Hanslo, M. 2009. The design and use of 'alternate' assessments of academic literacy as selection mechanisms in higher education. *Southern African Linguistics and Applied Language Studies*, 27(3): 265-276.
- Cliff, A.F. & Yeld, N. 2006. Test domains and constructs: Academic literacy. In Griesel, H. (Ed.). *Access and entry level benchmarks: The national benchmark tests project*. Pretoria: Higher Education South Africa, pp 19-27.
- Cliff, A.F., Yeld, N. & Hanslo, M. 2003. Assessing the academic literacy skills of entry-level students, using the Placement Test in English for Educational Purposes (PTEEP). Bi-annual conference of the European Association for Research in Learning and Instruction (EARLI), Padova, Italy.
- Cohen, R.J. & Swerdlik M.E. 2010. *Psychological testing and assessment*. New York: McGraw-Hill.
- Council on Higher Education. 2007. *Higher Education monitor 6: A case for improving teaching and learning in South African Higher Education*. Pretoria: Council on Higher Education.
- Council on Higher Education. 2013. *A proposal for undergraduate curriculum reform in South Africa: The case a flexible curriculum structure*. Pretoria. Council on Higher Education.
- Cummins, J. 1984. *Bilingualism and special education: Issues in assessment and pedagogy*. Clevedon, England: Multilingual Matters.
- Cummins, J. 1996. *Negotiating identities: Education for empowerment in a diverse society*. Ontario, California: California Association for Bilingual Education.
- Cummins, J. 2009. Fundamental psychological and sociological principles underlying educational success for linguistic minority students. In Skutnab-Kangas, T., Phillipson, R., Mohanty, A. K. & Panda, M. (Eds.), *Social justice through multilingual education*. Bristol: Multilingual Matters, pp. 19-35.
- Cummins, J. & Swain, M. 1986. *Bilingualism in education*. New York: Longman.

- Davies, A. 1990. *Principles of language testing*. Cambridge: Basil Blackwell.
- Davies, A. & Elder, C. 2005. Validity and validation in language testing. In Hinkel, E. (Ed.), *Handbook of research in second language teaching and learning*, Mahwah, New Jersey: Lawrence Erlbaum Associates, pp. 795-813.
- Department of Education. 1997. Education white paper 3: A programme for the transformation of higher education. Pretoria: *Government Gazette* No.18207, 15 August.
- Department of Basic Education. 2011. *Curriculum and assessment policy statement: Grades 10-12 English Home Language*. Pretoria: Department of Basic Education.
- Dörnyei, Z. 2007. *Research methods in applied linguistics*. Oxford: Oxford University Press.
- Du Plessis, C.L. 2012. The design, refinement and reception of a test of academic literacy for postgraduate students. Unpublished Masters dissertation. Bloemfontein: University of the Free State.
- Du Plessis, C.L. & Du Plessis, L.T. 2015. Dealing with disparities: The teaching and assessment of official languages at first language level in the grade 12 school-leaving phase in South Africa. *Language, Culture and Curriculum*, 28(3): 209-205.
- Du Plessis, C.L., Steyn, S. & Weideman, A. 2016. The assessment of home languages in the South African National Senior Certificate examinations – ensuring fairness and increased credibility. *LitNet Akademies*, 13(1): 425-443.
- Erford, B.T. 2013. *Assessment for counselors*. Belmont, California: Brooks/Cole Cengage Learning.
- Fleisch, B., Schoer V. and Cliff A.F. 2015. When signals are lost in aggregation: A comparison of language marks and competencies of first year university students. *South African Journal of Higher Education*, 29(5): 156-178.
- Fraenkel, J. & Wallen, N. 2003. *How to design and evaluate research in education* (5<sup>th</sup> ed.). New York: McGraw-Hill.
- Gee, J.P. 1990. *Social linguistics and literacies: Ideology in discourses*. London: Falmer Press.
- Gee, J.P. 1996. *Social linguistics and literacies: Ideology in discourses* (2<sup>nd</sup> ed.). London: Taylor and Francis.
- Gregory, R.J. 2007. *Psychological testing: History, principles and applications*. New York: Pearson.
- Griesel, H. 2006. The context of the National benchmark Tests project. In Griesel, H. (Ed.), *Access and entry level benchmarks: The national benchmark tests project*, Pretoria: Higher Education South Africa, pp 1-6.
- Hambleton, R.K. & Jones, R.W. 1993. Comparison of Classical Test Theory and Item Response Theory and their application to test development. *Educational Measurement: Issues and Practice*, 38-47.
- Haynes, N.S. & Lench, H.C. 2003. Incremental validity of new clinical assessment measures. *Psychological Assessment*, 15(4): 456-466.

- Hunsley, J. & Meyer, G. J. 2003. The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment*, 15(4): 446-455.
- Human Sciences Research Council. 1991. *Manual for proficiency test English second language advanced level*. Pretoria: Human Sciences Research Council.
- Jensen, A.R. 1980. *Bias in mental testing*. New York: Free Press.
- Johnson, R.B. & Christensen, L. 2004. *Education research: Quantitative, qualitative, and mixed approaches*, 2<sup>nd</sup> Edition. Boston: Allyn and Bacon.
- Kane, M.T. 1992. An argument-based approach to validity. *Psychological Bulletin*, 112 (3): 527-535.
- Kane, M.T. 2006. Content-related validity evidence in test development. In Downing, S. M. & Haladyna, T. M. (Eds.) *Handbook of test development*. New York: Routledge. 131-153.
- Kane, M.T. 2011. Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing*, 29(1): 3-17.
- Kelly, T.L. 1927. *Interpretation of educational measurement*. New York: Macmillan.
- Kobrin, J. L, Camara, J. W. & Milewski, G. B. 2004. The utility of the SAT I and SAT II for admissions decisions in California and the nation. In Zwick, R. (Ed). *Rethinking the SAT: The future of standardized testing in university admissions*. New York: RoutledgeFalmer, pp. 251-276.
- Kobrin, J.L. Patterson, B.F., Shaw, E.J., Mattern, K.D. & Barbuti, S.M. 2008. *Validity of the SAT for predicting first year college grade point average*. New York: College Board.
- Kumaravadivelu, B. 2003. *Beyond methods: Macrostrategies for language teaching*. London: Yale University Press.
- Kunnan, A.J. 2000. Fairness and justice for all. In Kunnan, A.J. (Ed.). *Fairness and validation in language assessment: Selected papers from the 19<sup>th</sup> Language Testing Research Colloquium*, Orlando, Florida. Cambridge: University of Cambridge Local Examinations Syndicate, pp. 1-14.
- Kurpius, S.E.R. & Stafford, M.E. 2006. *Testing and measurement: A user-friendly guide*. California: Sage Publications.
- Lado, R. 1961. *Language testing: The construction and use of foreign language tests*. New York: McGraw-Hill.
- Le, P.L., du Plessis, C.L. & Weideman, A. 2011. Test and context: The use of the Test of Academic Literacy Levels (TALL) at a tertiary institution in Vietnam. *Journal for Language Teaching*, 45(2): 115-131.
- Lynch, B.K. 2003. *Language assessment and program evaluation*. Edinburgh: Edinburgh University Press.
- Mackey, A. & Gass, S.M. 2005. *Second language research: Methodology and design*. New York: Routledge.
- McNamara, T.F. 1996. *Measuring second language performance*. London: Longman.



- McNamara, T.F. 2004. Language testing. In Davies, A. & Elder, C. (Eds.). *The handbook of applied linguistics*. Malden: Blackwell Publishing, pp. 763-783.
- McNamara, T.F. & Roever, C. 2006. *Language testing: The social dimension*. Language Learning Monograph Series. Language Learning Research Club, University of Michigan: Blackwell Publishing.
- Mdepa, W. & Tshiwula, L. 2012. Student diversity in South African higher education. *Widening participation and lifelong learning*, Special Issue, 13: 19 -33.
- Messick, S. 1980. Test validity and the ethics of assessment. *American Psychologist*, 35: 1012-1027.
- Messick, S. 1989. Validity. In Linn, R.L. (Ed.). *Educational measurement*. Third edition. New York: American Council of Education/Collier Macmillan, pp. 13-103.
- Miller, M.D., Linn, R.L. & Gronlund, N.E. 2009. *Measurement and assessment in teaching*. Upper Saddle River, New Jersey: Pearson Education.
- Ministry of Education. 2001. *National plan for higher education*. Pretoria: Department of Education.
- Montgomery, D.C., Peck, E.A. & Vining, G.G. 2006. *Introduction to linear regression analysis*. Hoboken, New Jersey: Wiley.
- Morgan, G.A., Leech, L., Gloeckner, G.W. & Barret, K.C. 2011. *IBM SPSS for introductory statistics: Use and interpretation*. New York: Routledge.
- Myburgh, J. 2015. The assessment of academic literacy at pre-university level: A comparison of the utility of academic literacy tests and Grade 10 Home Language results. Unpublished MA dissertation. Bloemfontein: University of the Free State.
- National Benchmark Tests Project. 2013. National Benchmark Tests Results – National Report N(1). Unpublished report. Cape Town: Higher Education South Africa.
- National Benchmark Tests Project. 2015a. Standard Setting Workshop. Unpublished Information Pack. Cape Town: Higher Education South Africa.
- National Benchmark Tests Project. 2015b. NBTP National Report: 2015 intake cycle – CETAP report number 1/2015. Unpublished report. Cape Town: Higher Education South Africa.
- National Council on Measurement in Education. 2010. Standard-setting methods as measurement processes. *Educational Measurement: Issues and Practice*, 29(1): 14-24.
- Patterson, R. & Weideman, A. 2013a. The typicality of academic discourse and its relevance for constructs of academic literacy. *Journal for Language Teaching*, 47(1): 107-123.
- Patterson, R. & Weideman, A. 2013b. The refinement of a construct for tests of academic literacy. *Journal for Language Teaching*, 47(1): 124-151.
- Pennycook, A. 1999. Introduction: Critical approaches to TESOL. *TESOL Quarterly*, 33(3): 329 – 348.
- Purpura, J.E., Brown, J.D. & Schoonen, R. 2015. Improving the validity of quantitative measures in applied linguistics research. *Language Learning*, 65(1): 37-75.

- Rambiritch, A. 2012a. Transparency, accessibility and accountability as regulative conditions for a postgraduate test of academic literacy. Unpublished PhD thesis. Bloemfontein: University of the Free State.
- Rambiritch, A. 2012b. Challenging Messick: Proposing a theoretical framework for understanding fundamental concepts in language testing. *Journal for Language Teaching*, 46(2): 108-121.
- Rambiritch, A. & Weideman, A. 2016. Telling the story of a test: The Test of Academic Literacy for Postgraduate Students (TALPS). In Read, J. (Ed.). *Post-admission language assessment in universities: International perspectives*. Forthcoming from Springer.
- Reise, S., Ainsworth, A. & Haviland, M. 2005. Item response theory: Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science*, 14: 95-101.
- Richards, J.C. 2001. *Curriculum development in language teaching*. Cambridge: Cambridge University Press.
- Rounds, P. 1996. The classroom-based researcher as fieldworker: Strangers in a strange land. In Schachter, J. & Gass, S. (Eds.), *Second language classroom research: Issues and opportunities*. Mahwah, New Jersey: Lawrence Erlbaum Associates, pp. 45-59.
- Salkind, N.J. 2006. *Exploring research*. Upper Saddle River, New Jersey: Pearson Prentice Hall.
- Salkind, N.J. 2011. *Statistics for people who think they hate statistics*. Los Angeles: Sage.
- Sebolai, K. 2014. Evaluating academic literacy teaching at a South African university: A case study of an academic literacy programme. *Journal for Language Teaching*, 48(1): 47-65.
- Sebolai, K. & Huff, L. 2015. Academic literacy curriculum renewal at a South African university: A case study. *Journal for Language Teaching*, 49(1): 333-351.
- Smit, C. Boraine, H. & Owen, R. 2006. Statistical analysis of AARP results. Unpublished concept document. Pretoria. University of Pretoria.
- Statistical Consultation Unit (SCU). 2015. Bloemfontein. University of the Free State.
- Stoynoff, S. & Chapelle, C.A. 2005. *ESOL tests and testing*. Alexandria, Virginia: TESOL.
- Umalusi (Council for Quality Assurance in General and Further Education and Training). 2012. *The standards of the National Senior Certificate Home Languages examinations: A comparison of South African official languages*. Pretoria: Umalusi.
- Van der Slik, F. & Weideman, A. 2005. The refinement of a test of academic literacy. *Per Linguam*, 21(1): 23-35.
- Van der Slik., F. 2008. Gender bias and gender differences in tests of academic literacy. *Southern African Linguistics and Applied Language Studies* Special issue: Assessing and developing academic literacy, 27(3): 277-290.
- Van der Slik, F. & Weideman, A. 2009. Revisiting test stability: Further evidence relating to the measurement of difference in performance on a test of academic

- literacy. *Southern African Linguistics and Applied Language Studies*, 27(3): 253-263.
- Van der Slik, F. & Weideman, A. 2010. Examining bias in a test of academic literacy: Does the Test of Academic Literacy Levels (TALL) treat students from English and African language backgrounds differently? *Journal for Language Teaching*, 44(2): 106-118.
- Van der Walt, C. 2010. Of shoes-and ships-and sealing-wax: A dynamic systems approach to language curriculum orientation. *Southern African Linguistics and Applied Language Studies*, 28(4): 323-327.
- Van der Walt, J.L. & Steyn, H.S. jnr. 2007. Pragmatic validation of a test of academic literacy at tertiary level. *Ensovoort*, 11(2): 138-153.
- Van der Walt, J.L. & Steyn, F. 2008. The validation of language tests. *Stellenbosch Papers in Linguistics*, 38, 191-204.
- Van Dyk, T. & Weideman, A. 2004. Switching constructs: On the selection of an appropriate blueprint for academic literacy assessment. *Journal for Language Teaching*, 38 (1): 1-13.
- Van, Dyk, T. 2005. Towards providing effective academic literacy intervention. *Per Linguam*, 21(2): 38 – 51.
- Van Dyk, T. 2010. Konstitutiewe voorwaardes vir die ontwerp van 'n toets van akademiese geletterdheid [TAG]. Unpublished PhD thesis. Bloemfontein: University of the Free State.
- Van Dyk, T. 2015. Tried and tested. *Tijdschrift voor Taalbeheersing*, 37(2): 159-186.
- Van Dyk T. & Van de Poel, K. 2013. Towards a responsible agenda for academic literacy development: Considerations that will benefit students and society. *Journal for Language Teaching*, 47(2): 43-70.
- Van Dyk, T., Van de Poel, K. & Van der Slik, F. 2013. Reading ability and academic acculturation: The case of South African students entering higher education. *Stellenbosch Papers in Linguistics Plus*, 42, 353-369.
- Van Els, T., Bongaerts, T., Extra, G., Van Os, C. & Janssen-van Dieten, A. 1984. *Applied linguistics and the learning and teaching of foreign languages*. London: Edward Arnold.
- Van Rensburg, C. & Weideman, A. 2002. Language proficiency: Current strategies, future remedies. *Journal for Language Teaching*, 36(1&2): 162-164.
- Van Rooy, B. & Coetzee-Van Rooy, S. 2015. The language issue and academic performance at a South African university. *Southern African Linguistics and Applied Language Studies*, 33(1): 31-46.
- Van Wyk, A. & Yeld, N. 2013. Academic literacy and language development. In Kandiko, C. B. & Weyers, M. (Eds.). *The global student experience: An international comparative study*. New York: Routledge, pp. 62-77.
- Visser, A.J. & Hanslo M. 2005. Approaches to predictive studies: Possibilities and challenges. *South African Journal of Higher Education*, 19(6): 1160-1176.
- Weideman, A. 2003. Assessing and developing academic literacy. *Per Linguam*, 19 (1&2): 55-65.

- Weideman, A. 2006. Transparency and accountability in applied linguistics. *Southern African Linguistics and Applied Language Studies*, 24(1): 71-86.
- Weideman, A. 2007. A responsible agenda for applied linguistics: Confessions of a philosopher. *Per Linguam*, 23(2): 29-53.
- Weideman, A. 2007. *Academic literacy: Prepare to learn*. Pretoria: Van Schaik.
- Weideman, A. 2009. Constitutive and regulative conditions for the assessment of academic literacy. *Southern African Linguistics and Applied Language Studies*, 27: 1-26.
- Weideman, A. 2011. Academic literacy tests: Design, development, piloting and refinement. *Journal for Language Teaching*, 45(2): 100-113.
- Weideman, A. 2012. Validation and validity beyond Messick. *Per Linguam*, 28 (2): 1-14.
- Weideman, A. 2013. Academic literacy interventions: What are we not yet doing, or not yet doing right? *Journal for Language Teaching*, 47 (2): 11-23.
- Weideman, A. 2014. Innovation and reciprocity in applied linguistics. *Literator*, 35 (1): 1-10.
- Weideman, A. 2016. The refinement of the idea of consequential validity within an alternative framework for responsible test design. Chapter submitted for Allan J. & Artiles A. (Eds.) 2017. *Assessment inequalities: Routledge World Yearbook of Education*.
- Weideman, A., Du Plessis C. & Steyn, S. 2015. Diversity, variation and fairness: Equivalence in national level language assessments. Paper presented at the 4<sup>th</sup> International Conference on Language, Education and Diversity, Auckland, New Zealand, 23-26 November 2015.
- Weideman, A., Patterson, R. & Pot, A. 2016. Construct refinement in tests of academic literacy. In Read, J. (Ed.). *Post-admission language assessment in universities: International perspectives*. Forthcoming from Spinger.
- Weir, C.J. 1993. *Understanding and developing language tests*. New York: Prentice Hall.
- Whiston, S.C. 2013. *Principles and applications of assessment in counseling*. Pacific Grove, California: Brooks/Cole Cengage Learning.
- Yeld, N. 2001. Assessment, equity and language learning: Key issues for higher education selection in South Africa. Unpublished PhD thesis. Cape Town: University of Cape Town.