

**The assessment of academic literacy at  
pre-university level: A comparison of  
the utility of academic literacy tests  
and Grade 10 Home Language results**

**Jo-Mari Myburgh**

**The assessment of academic literacy at pre-  
university level: A comparison of the utility of  
academic literacy tests and Grade 10 Home  
Language results**

Jo-Mari Myburgh

A dissertation submitted to meet the requirements for the degree Magister Artium (Linguistics) in the Faculty of the Humanities (Department of Language Practice and Linguistics) of the University of the Free State.

February 2015

Supervisor: Professor A.J. Weideman

## Acknowledgements

This study would not at all have been possible without the constant encouragement of my supervisor, Professor Albert Weideman, who not only opened my mind to the captivating world of applied linguistics, but my heart as well. Additionally, I would like to express my deepest gratitude to the National Research Foundation (NRF), which funded my studies for this past year. I would also like to thank my friends and family - for their unconditional love and unwavering support. And finally, I would like to acknowledge countless blessings I have received – all of which have helped me to finish this study with the passion it deserves.

“I have dreamt in my life, dreams that have stayed with me ever after, and changed my ideas; they have gone through and through me, like wine through water, and altered the colour of my mind”

From *Wuthering Heights* by Emily Brontë

## **Declaration**

I herewith declare that this thesis, which is being submitted to meet the requirements for the qualification Magister Artium (Linguistics) in the Faculty of the Humanities (Department of Language Practice and Linguistics) of the University of the Free State, is my own independent work and that I have not previously submitted the same work for a qualification at another university. I agree to cede all rights of copy to the University of the Free State.

## Table of contents

<b>Acknowledgements</b>	i
<b>Declaration</b>	ii
<b>Chapter 1</b>	
<b>Introduction: The importance of academic literacy testing for first-time students at universities in South Africa</b>	
1.1 Background to the problem	1
1.2 Rationale for the study	3
1.3 Research aims	7
1.4 Research procedure	8
1.5 Overview	9
1.6 Value of the research	10
<b>Chapter 2</b>	
<b>The selection and assessment instrument</b>	
2.1 Introduction	13
2.2 Language testing as a sub-discipline of applied linguistics	14
2.3 Phases of language testing and teaching	18
2.4 Language test development phases echo those of other applied linguistic artefacts	23
2.5 Validity and related principles traditionally identified as conditions for test design	26
2.6 Key principles for the design of a language test	29
2.7 Other principles	33
2.8 Conclusion	38
<b>Chapter 3</b>	
<b>The test construct and its operationalisation: Design principles and phases</b>	
3.1 Introduction	39

3.2	The evolution of the construct	40
3.3	Principles for test design and selection	43
3.4	The operationalization of the current construct: Design phases	46
3.5	Test purpose and construct definition	48
3.6	Specifications and task types	50
3.7	Conclusion	54

## **Chapter 4**

### **Research method**

4.1	Introduction	56
4.2	Home language assessment processes	56
4.3	The specific nature of the additional tests	59
4.3.1	TALA	62
4.3.2	The second test	63
4.4	The target group	70
4.5	Procedure	71
4.6	The regression analysis and choice of variables	72
4.7	The claims	73
4.8	Conclusion	75

## **Chapter 5**

### **Analyses and interpretation**

5.1	Introduction	76
5.2	Iteman 3.6 analysis	76
5.2.1	TALA	77
5.2.2	The second test	79
5.3	Iteman 4.3 analysis	79
5.3.1	TALA	80
5.3.2	The second test	82
5.4	TiaPlus analysis	83
5.4.1	TALA	84
5.4.2	The second test	87

5.5	Regression and related analyses	88
5.6	Discussion	93
5.7	Answering the claims	96
5.8	Conclusion	99

## **Chapter 6**

### **Refinement of the second test**

6.1	Introduction	100
6.2	Why refine the second test?	100
6.3	Potential refinements to test items	103
6.3.1	Parameters for a productive item	103
6.3.2	Refinements of individual items	104
6.4	Conclusion	109

## **Chapter 7**

### **Conclusions and recommendations for future research**

7.1	Introduction	110
7.2	Summary	110
7.3	Recommendations	111
7.4	Limitations of the study	114
7.5	Further investigations	116
7.6	Conclusion	117

<b>Bibliography</b>	119
---------------------	-----

## **Annexures**

<b>Annexure A</b>	The second test	126
<b>Annexure B</b>	Iteman 3.6 analysis of TALA	138
<b>Annexure C</b>	Iteman 3.6 analysis of the second test	149
<b>Annexure D</b>	Iteman 4.3 analysis of TALA	160
<b>Annexure E</b>	Iteman 4.3 analysis of the second test	197
<b>Annexure F</b>	TiaPlus analysis of TALA	235
<b>Annexure G</b>	TiaPlus analysis of the second test	252
<b>Annexure H</b>	Correlational analysis	269
<b>Annexure I</b>	Regression analysis	272
<b>Annexure J</b>	ANCOVA analysis (1)	276
<b>Annexure K</b>	ANCOVA analysis (2)	278
<b>Annexure L</b>	ANCOVA analysis (3)	280



## List of tables

Table 2.1	Levels of applied linguistic artefacts	14
Table 2.2	Constitutive and regulative moments in applied linguistic designs	16
Table 2.3	Phases in language assessment	18
Table 2.4	Two perspectives on language	22
Table 2.5	Traditions of applied linguistics	24
Table 3.1	Bachman and Palmer's construct	41
Table 3.2	Specifications and task types	52
Table 4.1	A summary of the test papers for final phase examinations	58
Table 4.2	TALA test specifications	66
Table 5.1	Scale statistics, TALA	77
Table 5.2	Scale statistics, second test	79
Table 5.3	Iteman 4.3 reliability report of TALA's re-pilot	80
Table 5.4	Iteman 4.3 summary statistics of TALA's re-pilot	81
Table 5.5	Iteman 4.3 reliability report of the second test	82
Table 5.6	Iteman 4.3 summary statistics of the second test	82
Table 5.7	Subtest intercorrelations of TALA	84
Table 5.8	DIF statistics for TALA	85
Table 5.9	Misclassifications for TALA	86
Table 5.10	Subtest intercorrelations of the second test	87
Table 5.11	Misclassifications for the second test	88
Table 5.12	Correlation analysis results	89
Table 5.13	TALA's performance during its first pilot	94
Table 6.1	TALA's reliability indices for its first pilot	100
Table 6.2	Flesch-Kincaid levels of texts in both TALA and the second test	101
Table 6.3	Summary of items which did not perform satisfactorily as indicated by Iteman 4.3	104

## List of figures

Figure 2.1	Terminal and other functions of an applied linguistic design	15
Figure 2.2	Bachman and Palmer's model of test usefulness	27
Figure 3.1	The test design cycle	47
Figure 5.1	Covariance analysis between average excluding English and test 2	90
Figure 5.2	Covariance analysis between average excluding English and test 1	91
Figure 5.3	Covariance analysis between average excluding English and English	91

## **Abstract**

The definition of academic literacy utilised for this study proposes that the distinction-making activity accompanying academic discourse constitutes what makes academic discourse unique, which at the same time also discloses that academic discourse is a distinctive language with its own conditions, different from other lingual spheres, as opposed to earlier definitions which often took a closed view of language, regarding it as consisting of sound, form and meaning. A construct deriving from such a specific definition of academic discourse therefore acknowledges the shift in focus of language instruction and assessment brought on by the communicative approach. An academic literacy test designed to establish the academic literacy levels of prospective tertiary education students should therefore be aligned with this construct. For this study, two academic literacy tests were administered to two groups of Grade 10 students in order to determine how accurately these tests would disclose the students' levels of ability to handle language for learning. The students' school marks were then compared to the marks received for the academic literacy tests. Although the school language marks predicted the general academic performance of the test population more accurately than the proposed academic literacy tests, the second test used came close to predicting these levels almost as accurately as the school marks. Read in conjunction with a number of other current studies, this result, however, still emphasises the significance of and need for well-designed, construct-based and correctly pitched (as regards level) academic literacy tests.





# Chapter 1

## Introduction: The importance of academic literacy testing for first-time students at universities in South Africa

### 1.1 Background to the problem

We now have more students attending universities and other forms of tertiary education institutions than ever before. The number of students applying to South African universities has grown immensely during the last two to three decades (Cliff, Yeld & Hanslo 2003:1). According to SouthAfrica.info (2014), Higher Education enrolments have increased by 41% since 1993 while the Department of Basic Education (2005:8) also records a steady increase of student enrolments since 2000. In its 25 April 2014 edition, *Rapport* reported that the number of black students who completed their tertiary education had increased by 300% since 1991 (Jeffery 2014). This shift from a type of elite education system to an education system which supports larger numbers of students was both predicted and welcomed by the National Commission on Higher Education (NCHE) in 2001 (Department of Basic Education 2001). Whilst in essence this is a good thing which many see as contributing towards “enhanced skills development for students, improved job and career opportunities, improvements in society, the economy and communities, and a commitment to realising the principles of life-long learning” (Cliff, Yeld & Hanslo 2003:1), it also brings with it its own challenges; for we know that to be able to perform successfully at university, a student needs to be able to handle the kind of language used there, which is academic discourse. In a number of studies undertaken since the mid-1990s, it has become clear, however, that the ability of new entrants in Higher Education to handle academic discourse may not be at an adequate level (Van Rensburg & Weideman 2002:152). Two key factors come to mind when one considers this problem. Firstly, we need to ask whether the school curriculum places enough emphasis on the importance of teaching academic

discourse in order to prepare learners for the demands of Higher Education, and secondly we need to ask whether academic discourse is subsequently being assessed in a valid and responsible way.

According to the current Curriculum and Assessment Policy Statement (CAPS) which provides the guidelines through which teachers are expected to plan their lessons and year programme, students are expected to engage with language and texts that function within the following material lingual spheres (Weideman 2009:39; Weideman 2011:60; Patterson & Weideman 2013a:109) or types of discourse (Department of Basic Education 2011):

- social (including inter-personal communication and the handling of information)
- economic/professional (including the world of work and commerce)
- academic (including academic and scientific language and advanced language ability)
- aesthetic (including the appreciation of literature and art)
- ethical (including an appreciation of the values embedded in language use)
- and political (including the critical discernment of power relations in discourse)

From this list we can conclude that academic discourse, as one element of a differentiated ability to use language, is in fact included in the curriculum requirements as stated by CAPS. However, Du Plessis, Steyn and Weideman (2014:6) question whether the construct provided by the curriculum and its subsequent testing are aligned. They note, for example, that within CAPS (Department of Basic Education 2011:9) it is mentioned that students need to be “able to use a sufficiently high standard of language in order to be able to gain access to ‘further or Higher Education or the world of work’” (Weideman, Du Plessis & Steyn 2014:5). It is necessary to ask, however, if a “high standard of language” and academic discourse can be regarded as the same type of discourse. Such a lack of clarity can easily lead to a misalignment between the aims of the curriculum and the subsequent assessment of students’ attempts at the realisation of these aims. Additionally, a misalignment of instruction and assessment can affect the reliability and validity of school language results.

A further concern pertains to whether the tests used by schools to assess the ability to handle the language of the various material lingual spheres (Weideman 2009:39) or different discourse types (including academic discourse or a high standard of language) demonstrate a responsible assessment of the language ability of our students (Patterson & Weideman 2013a:109). The chief concern lies in the valid or invalid assessment of students, because to “prevent biased examination tasks all learners should have access to the same outside knowledge” (Weideman, Du Plessis & Steyn 2014:13), which in the South African context is clearly not the case. Many inequalities still exist amongst communities and schools, and unfortunately all of our students do not receive the same privileges, resources and assistance when it comes to education.

## **1.2 Rationale for the study**

Since there are doubts about the ability of new entrants into Higher Education to handle the demands of academic discourse, universities have instituted a number of different support mechanisms in order to provide a solution to the problem of levels of academic literacy that are too low. Such support mechanisms conventionally take the shape of general academic support programmes or specific academic literacy courses. Who should be placed on such interventions, however? In South Africa, two approaches are currently being utilised to determine whether students hold an adequate level of academic literacy, in order to place them in programmes which are designed to assist them as necessary in engaging effectively with the written and spoken texts that are part of academic discourse.

The first option which is prevalent in determining whether students are capable of handling academic discourse includes making use of post-entry tests. These tests are “administered to students after they have been admitted to a tertiary institution, with a view to identifying those who are likely to struggle to meet the language demands of their degree programme and who should be encouraged or required to



enhance their academic language skills” (Read 2012:1). The University of Cape Town did pioneering work in this regard in South Africa, by developing an Alternative Admissions Research Project (AARP) which designed a test by the name of PTEEP (Placement Test in English for Educational Purposes) (Cliff, Yeld & Hanslo 2003:4). PTEEP was superceded by the NBTs (National Benchmark Tests) which were commissioned by HESA (Higher Education South Africa) as a purportedly standardised option to test academic literacy for South African students who would like to further their education beyond secondary school (National Benchmark Tests Project 2013). Other valuable tests of this type also exist. The Inter-Institutional Centre for Language Development and Assessment (ICELDA 2014) offers a range of these types of tests, which includes TALL (Test of Academic Literacy Levels), TAG (Toets van Akademiese Geletterdheidsvlakke) and its postgraduate counterpart, TALPS (Test of Academic Literacy for Postgraduate Students). A good proportion of South African universities and many other tertiary educational institutions across the country make use of such post-entry tests (National Benchmark Tests Project 2013). But, as is evident form the NBT website (2013), not all prospective South African students write these tests, which brings us to the next approach.

Some universities make use of a second approach with regards to the determination of academic literacy levels, which is the utilisation of a student’s school language results. Specifically, one institutional partner in the ICELDA (Institutional Centre for Language Development and Assessment) consortium has begun to utilise this approach. Based on the mark students obtained for English First Additional Language or English Home Language in their final year at school, the Faculty of Humanities at the University of Pretoria uses students’ school English mark to determine whether they should enrol for academic literacy modules (University of Pretoria 2014). This second option is, however, likely to be inadequate for a number of reasons, as Cliff, Yeld and Hanslo (2003:2) emphasise:

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.

Exit-level examinations at school should be regarded as high-stakes tests, since the results generated by the tests are used to deny or grant students access into universities and also into the workplace (Weideman, Du Plessis & Steyn 2014:2). In this sense, using school results as the only measure of university readiness might exclude a wide variety of potentially able students, as they have not been given an adequate opportunity of demonstrating their true academic potential (Cliff, Yeld & Hanslo 2003:2). In turn, this yields the possibility of other problems, such as students feeling demotivated and let down by the education system, as well as parents doubting the significance of obtaining an education. In the case of the present study, however, the question is not so much about academic potential in general, but about identifying in a reliable and valid way what level of academic literacy a student possesses and, by extension, what level of language support, in the form of a language course, a student would need. Of course, academic literacy cannot be equated with academic support in general. This study is concerned with instruments that can identify levels of academic literacy. If the only instrument used to gauge this is a high stakes measure administered directly before entry into university, then the determination of how to place a prospective entrant appropriately may get confused with issues of access – the high stakes decision on whether or not to allow a student into Higher Education in the first instance.

It will therefore be the argument of this study that access decisions should preferably not be confounded with decisions about what level of support is necessary, in terms of a specific ability, that of handling academic discourse. The latter kinds of decisions are not pre-entry, high stakes ones, but more appropriately – given the current expanding access to Higher Education in South Africa that was referred to above – low to medium stakes, post-entry placement (support) decisions.

The argument will therefore refer throughout to the shortcomings of using only school language results as a proxy for academic literacy levels.

A further question, to be specifically addressed in this study, is how early one can identify low academic literacy levels. The tests referred to above, especially tests like TALL and TAG, are written at the beginning of students' first year at university. It would be profitable to compare the results of TALL and TAG with either Home Language marks in English or Afrikaans – if students of course have written TALL and TAG – but not all universities use such tests, as has been noted above. There is another study which also looks at the relationship between Home Language marks, results of academic literacy tests and first year performance, which is being undertaken by Sebolai (2015). That study is not yet complete and would naturally augment the findings of the current study. This study, however, proceeds from the assumption that academic literacy can be defined as the language one needs for learning at all levels – university level, secondary school level and even earlier (Steyn 2014). For example, Steyn's dissertation at the Rijksuniversiteit of Groningen deals with a Test of Early Academic Literacy (TEAL). Additionally, Grün's 2015 study intends to justify the design of a test of emergent literacy for pre-schoolers. It might thus be possible to consider the level of such ability much earlier than the first year of university, and therefore this study takes Grade 10 students as such a possible point of identifying low levels of academic literacy.

The early identification of academic literacy levels is clearly advisable. Cliff, Yeld and Hanslo (2003:3) note that “(academic) success is constituted of the interplay between the language (medium-of-instruction) and the academic (typical tasks required in Higher Education) demands placed upon students.” This is a common problem that many students in South Africa face on a daily basis. A country that attempts to promote multilingualism through having eleven official languages, but which fails to assign equal or at least substantial authority and resources to all of its eleven recognised languages may well experience problems with regard to Higher

Education. It is often wrongly assumed that students who are fluent in their mother tongue will rapidly become fully proficient in English. The sad truth is that many people are unaware that “being able to read, write and speak in one language does not make one ‘literate’ in another” (Parkinson 2000:369). When left unaddressed, this problem could potentially impede a student’s academic progress at tertiary educational institutions, or even in the workplace.

### **1.3 Research aims**

The aims of this study may be presented in the following list:

- The main aim of this study is to determine whether universities can acknowledge students’ school language results as a reliable source for the determination of their ability to handle academic discourse at university level, or whether the use of a specialised measure, in the form of a specific assessment of the ability to handle academic discourse, is more preferable and appropriate.
- Accompanying the main aim is the confirmation of the usefulness of assessing the ability to use academic language, as stated in the curriculum. CAPS, for example, stipulates that students need to be able to engage with texts of an academic nature (Weideman, Du Plessis & Steyn 2014:9; Department of Basic Education 2011), which then necessitates that academic literacy levels need to be assessed.
- A further aim would include articulating the kind of emphasis language teaching should take in order to make up for potential shortfalls either in the school curriculum, or in actual teaching.
- A subsidiary aim (see below, Research procedure) is to demonstrate that in assessing academic literacy adequately at school level one needs a refined test or even set of tests.

- Finally, the empirical data collected for this study can possibly be used to substantiate the notion that academic literacy could be influential on overall academic achievement.

#### **1.4 Research procedure**

The research procedure will be carried out through the administration of two selected academic literacy tests to Grade 10 students of two Bloemfontein based high schools. The first test, named The Test of Advanced Language Ability (TALA), is a test that has previously been administered to Grade 12 groups, also in the city of Bloemfontein. The second test has been taken from a test book by Weideman and Van Dyk (2014), which was specifically created for high school students who need to prepare for academic literacy assessments. The use of more than one test could increase the credibility of this study – a study which could address the academic literacy needs of Grade 10 students. The second test will be piloted for the first time and possibly refined, since if one refines an academic literacy test such as the ones to be used in this study, which might give an indication of whether that should become part of language instruction and its assessment at school, and so give a reliable indication of academic literacy levels. If academic literacy is indeed a crucial part of the differentiated set of abilities prescribed by CAPS (and its predecessors), then it is vital to establish what such tests should look like and whether the development of similar tests would be useful, since the measurement of the ability to handle academic discourse would then form part of the overall assessment of the ability to handle (the home) language.

I have chosen Grade 10 students for two reasons, the first being that Grade 11s and Grade 12s are more limited by time constraints because of their more demanding schedules, which makes the Grade 10 group a more accessible and convenient target group. Furthermore, by identifying the academic literacy needs of students in their Grade 10 year, more time is available to address the needs of the students or

remediate problems. The second reason pertains to the various reports issued by Umalusi on the findings for Home Language examinations that “the quality and standard of the assessment in the exit-level examinations need urgent scrutiny” (Weideman, Du Plessis & Steyn 2014:2), which implies that Grade 12 language results need to be treated with care and cannot unconditionally be regarded as a reliable and accurate source of students’ academic literacy levels. Although the Grade 10 results are themselves not unproblematic, the access to examination marks to which the results of the mentioned academic literacy tests will be compared, are more easily available.

The results of the academic literacy tests will then firstly be compared to the students’ results for English Home Language that was obtained in their June examinations. Additionally, the results of the academic literacy tests will be compared to the students’ overall average, sometimes called their GPA (Grade point average), as well as to their overall average excluding their English Home Language mark. These second and third comparisons will be of importance, as they could indicate whether a link exists between a student’s academic literacy levels and their overall academic success, and how strong that relation is.

### **1.5 Overview**

The second chapter will represent a literature review which will focus on the design of a solution pertaining to the problem stated in the first chapter. The review will survey the history of language assessment and traditions of applied linguistics that relate to different paradigms that have affected language assessment. I shall also discuss in more detail key principles which are crucial to responsible language test design.

The third chapter will build on the literature review of the second. Included will be literature relevant to the selection of appropriate tests for the study, and the

justification of the chosen test construct, as well as a discussion of the evolution of this construct, which also yields the test components and task specifications that flow from it.

The next chapter will include the method through which this study will be conducted. Included also will be a discussion of the nature of English Home Language assessments and examinations, as well as a justification of the additional tests which were chosen for this study. Additionally, a discussion will follow regarding the choice of test takers and types of analyses which will be carried out on the results obtained. A set of claims will also be presented regarding what the analyses of the results may disclose.

Chapter five will contain the analysis of the results after the tests have been administered for the first time. The analyses will focus finally on the comparisons mentioned, and the implications of the findings. In addition, it will draw a number of conclusions with regard to the administration of similar tests at school level.

Chapter six will be aimed at describing the possible refinement of the first draft of one of the tests that have been used, as well as the justification for refining that specific test and not the other.

The final chapter will include general findings, a discussion of the limitations of the study and further recommendations, as well as propose possible further research that could guide the subsequent further development and administration of one of the tests and instructional material in Home Languages.

### **1.6 Value of the research**

The empirical evidence collected for this study will in the first instance provide us with insight into the appropriateness or inappropriateness of using students' school language results as an indication of their readiness to handle the demands of using

academic language in universities. In essence, school language marks and marks obtained through administered academic literacy tests will be pitted against one another in order to determine which provides a better indication of university readiness with regard to academic discourse.

This study could also assist in clarifying whether students should undertake a separate academic literacy test at a later stage such as when they apply to universities, or whether an academic literacy test or assessment might not perhaps be included in language instruction at school as CAPS (2011 Department of Basic Education) clearly stipulates (Du Plessis, Steyn & Weideman 2014). Having academic literacy testing assessed at school level could in some respects be beneficial, as the earlier identification of at risk students would facilitate the earlier provision of support mechanisms for these students. The curriculum already allows for the development of the language of learning. However, studies such as those undertaken by Du Plessis, Steyn and Weideman (2014) indicate that language instruction at this level may currently suffer from a number of deficiencies.

This study will also be of value to university administrators, as they need to rely on valid test results, derived from consistent measurements, to ensure that students are placed within applicable programmes. The study will therefore emphasise the importance of responsible academic literacy testing as well as the responsible interpretation of test results. One example of the irresponsible interpretation of test results can be traced back to some current uses of the NBT results. The NBTs were designed “to better inform learners and universities about the level of academic support that may be required for successful completion of programmes” (National Benchmark Tests Project 2013), which clearly categorises the NBTs as placement tests. In spite of this, some universities and tertiary educational institutions use the results of the NBTs to accept or deny students access to their programmes. This is not defensible, as it contradicts the purpose of the test, which is that of a placement test. Perhaps this contradiction relates to an ambiguity with which the NBT test



designers and their collaborators present the purpose of the test. Cliff and Hanslo (2005:1) note that it “goes almost without saying that Higher Education institutions worldwide, and the coordinators of the study programmes these institutions offer, need to adopt a coherent and defensible approach towards the selection of students to these institutions”, which indicates an immediate contradiction between the idea of the selection of students (before they have access to Higher Education) and the placement on appropriate courses after they have gained entry. The first kind of decision is a high stakes decision that will have an effect on the increased or limited earning power of an individual student throughout their working lives. The latter kind is a medium to low stakes decision about what kind of post-admission support might be appropriate for students to develop their ability to handle academic discourse at university. The temptation to use the NBTs as access tests derives in part from them being administered before entry to university. This study will critically examine this practice, in order to propose a possible alternative.

Lastly, but possibly the most valuable contribution of this study, are the potential changes in emphasis of language teaching that will be identified. One of the main aims of our curriculum embodies the preparation of our students to be functional in managing a life after school. Consequently, it goes without saying that our students will then firstly need to be competent and successful as Higher Education students (Department of Basic Education 2011:4), which is an objective worth emphasising in the curriculum and worth undertaking by our schools. This does not mean offering a separate course benefitting a minority of learners, namely those going to higher education institutions, but merely emphasising in Home Language instruction components of the syllabus (CAPS) that already require and prescribe this.

## **Chapter 2**

### **The selection and assessment instrument**

#### **2.1 Introduction**

It will be the argument of this study that the judicious employment of an academic literacy test might be a possible solution to the potential inadequacy of using school language results as an indicator of first year students' ability to handle academic discourse at university level. Before such a solution can be adopted, however, the relationship between language test design theory and applied linguistics must be articulated in order to account for the principles of language test design which guide the selection, design and evaluation of language tests and the development of appropriate test constructs. An observation by Green (2014:173) confirms the complex history and nature of language test design theory and the even more complex endeavour that is language assessment:

Language assessment has been shaped by a wide range of influences, including practicality, political expediency and established customs, as well as developments in language teaching, applied linguistics and other allied disciplines such as educational psychology. Global trends including the growth in international trade, mass migration and tourism have brought new reasons for learning and using languages, and naturally assessment has also been affected by these broader social changes.

This chapter will focus on how principles for language test design can be articulated with reference to applied linguistics and the history of applied linguistic designs, as well as with reference to certain key and other chief principles. This survey is undertaken in order to articulate how a responsible design choice can be made for an assessment instrument that will be appropriate for use in the educational contexts (upper secondary school and higher education) of this study.

## 2.2 Language testing as a sub-discipline of applied linguistics

Language tests, together with language policies and language curricula, form part of the practice of applied linguistics as a discipline of design (Weideman 2014:2). The relationship between language assessment and applied linguistics is emphasised by Weideman (2014) in his assertion that applied linguistics is “a discipline of design: it solves language problems by suggesting a plan, or blueprint, to handle them.” Language tests are technically qualified instruments (Weideman 2011:101). A language test’s functionality is therefore dependent on its capacity to assess language ability through the technical character of its design. Weideman (2011:101) suggests a reciprocal relationship between the norms for the technical designs of applied linguistic artefacts, and the end-user formats of these artefacts. Applied linguistic designs thus operate on two levels: that of a conditioning artefact and that of an end-user format of that artefact. The end-user format should be aligned with the norms that apply to it. This may be presented in the form of a table (Weideman 2011:101):

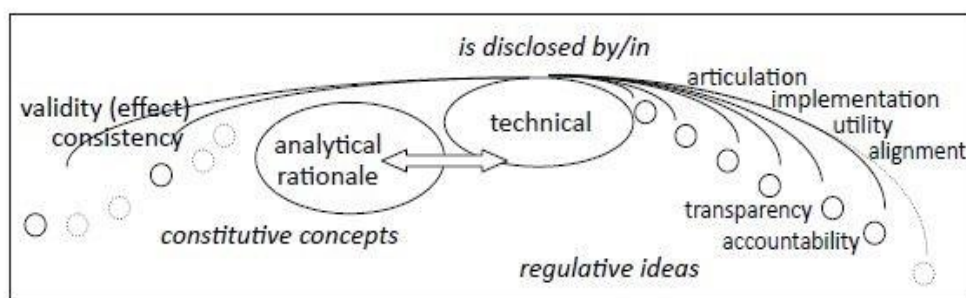
<b>Prior conditioning artefact</b>	<b>End-user format of design</b>
Language curriculum	Language course
Construct and test specifications	Language test
Language policy	Language management plan

**Table 2.1: Levels of applied linguistic artefacts**

A language policy which prescribes the norms and specifications of a language management plan remains an applied linguistic artefact, as it represents a technical design framework for addressing a certain language problem, whilst the management plan itself is also an artefact, as it embodies the final format of the design as prescribed by a language policy. In turn, a language test is regulated by its test construct and specifications, which act as a theoretical justification for the specific design of a test.

Practising language test design within the scope of applied linguistics therefore involves an approach to test design that refers to the conditions discovered for such

designs by applied linguistic theory. These conditions may be conceptualised as constitutive and regulative requirements for all applied linguistic designs including, in this case, conditions or requirements for language assessment (Weideman 2011:102). Du Plessis (2012:36) explains that in “language testing 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 for the intervention.” The reciprocal relationship that exists between the technical mode and the analytical dimensions of an applied linguistics artefact such as a language test can be seen in the representation below (Weideman 2014:7):



**Figure 2.1: Terminal and other functions of an applied linguistic design**

The relationship between the founding function and the qualifying function of an applied linguistic artefact is relevant to the design of the artefact, as well as the principles that originate from the leading technical function of the same artefact. Weideman (2014:7) suggests that in the “connections that the technical aspect of reality has with all the other dimensions, we potentially find the normative moments that might serve as applied linguistic design principles.” Consequently, we encounter constitutive technical concepts and regulative linguistic ideas when we investigate the technical dimension of experience. For example, the technical reliability of a test is dependent on the relationship which the technical mode of experience shares with the kinematic dimension of reality. These connections, enumerated below in the last column, between the technical mode of reality and the others, are represented by Weideman (2014:8) in the following table:

<b>Applied linguistic design</b>	<b>Dimension of experience</b>	<b>Kind of function</b>	<b>Retrociatory analogical moment</b>
is founded upon	numerical	Constitutive	systematicity
	spatial		limits, range
	kinematic		technical reliability
	physical		internal effect
	biotic		differentiation
	sensitive		intuitive appeal
	analytical	Founding	design rationale
is qualified by	technical	Leading function	-
is disclosed by	lingual	Regulative	articulation of design
	social		implementation
	economic		technical utility
	aesthetic		resolving misalignment
	juridical		transparency, fairness
	ethical		accountability, care
	faith		reputability, trust

**Table 2.2: Constitutive and regulative moments in applied linguistic designs**

From each of the corresponding constitutive technical concepts or regulative ideas issues a normative appeal to the designers of applied linguistic artefacts. These normative moments thus condition the design of an applied linguistic artefact such as a language test. The meaning of these design conditions for both language courses and language tests may be articulated as follows (Weideman 2014:8):

- 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.

Formulated thus, the analogical moments and other dimensions of reality that are reflected in the technical can each be taken up as an injunction to language test designers to create tests that conform to certain fundamental principles. By attending to both the regulative and constitutive conditions for language test design as articulated above, one can ensure that a test conforms to criteria of responsible test design, one of the most important of which is that the test construct should be theoretically defensible, a point to which I shall return in a detailed discussion in the next chapter. Weideman (2014:8) claims that these principles or design requirements are common to both language tests and language courses, though they may be specified slightly differently to accommodate, respectively, the typical nature of the assessment instrument (a language test) or of the language instruction (a language course). This is not the only argument for conceptualising both language assessment and language teaching as applied linguistic designs. I return below (see section 2.4) to a further, historical argument after first surveying below the phases of language testing and teaching that are relevant for the choice of assessment instrument in this study.

The principles of test design that we practise today do not present themselves to us in a vacuum. They have been discovered and articulated right through the history of language testing. It is to the disclosure of these principles in the history of

language testing that I turn in the next section in order to have a sounding board for the selection of an assessment measure that is appropriate for this study.

### 2.3 Phases of language testing and teaching

This section will consider how the ways in which we teach and test language have changed as our perceptions have changed regarding what language is, and how language should be defined. Green (2014:173) observes that “different theoretical accounts of language and different theories of measurement have come in and out of favour in different parts of the world”. These shifts in how languages are conceptualised have given rise to certain key phases in the field of language testing. Green (2014) summarises an account of Spolsky’s views of the evolution of language testing and teaching (1995) in the form of a table:

Language testing	Language teaching	Favoured assessment techniques
Pre-scientific/traditional	Grammar translation	Translation, grammar exercises, essays.
Psychometric-structuralist	Audio-lingualism	Multiple choice tests of grammar, vocabulary, phonetic discrimination, reading and listening comprehension. Focus on reliability.
Psycholinguistic-sociolinguistic	Natural approach	Cloze, dictation.
Communicative	Communicative/task-based approach	Assessment tasks intended to reflect ‘real life’ language use. Integrated skills. Focus on validity of content.

**Table 2.3: Phases in language assessment**

The pre-scientific or traditional phase provided three objectives with regard to language learning. First was the enjoyment of the literature of the target language, second the appreciation of its culture, and third the ability to communicate with its

users with ease (Green 2014:175). Students were regularly assessed orally, expected to correct sentence errors, combine sentences, and to participate in translation and dictation exercises, which only sometimes pertained to the objectives mentioned (Green 2014:176). However, with time, multiple concerns were raised, which included that such assessment did not actually assess proficiency in language. Students were not being assessed on their ability to communicate, for example, but rather on their ability to express (Green 2014:177). Eventually, as we shall again observe below, the difference between assessments that measure either individual expression or a shared expression or communication is such that they embody a paradigm shift in language testing (Weideman 2009:63), and one which is highly relevant for this study. Not only were assessments in this traditional mould lacking in commitment to test communicative ability, but they were also yielding unreliable test results (Green 2014:177).

The second phase, the psychometric-structuralist phase, came about as an attempt to attend specifically to the matter of unreliable test results. Lado (1961) claimed that the testing techniques associated with the audio-lingual method were more scientific, since the results were psychometrically obtained. Multiple choice questions were favoured for the objective manner in which the questions were marked as opposed to, for example, essay marking, which could only be done subjectively (Green 2014:178). Additionally, Lado recommended discrete-point testing, which called for the separate and single item based assessment of what he deems the four language skills of listening, speaking, writing and reading (Green 2014:179, Patterson & Weideman 2013b:143). Lado justified the separate assessment of the four language skills on the basis that it would reveal “a more general picture of (a student’s language) proficiency” (Green 2014:180), as well as disclose a student’s true ability to apply language knowledge in real life. In contrast to the long essays or translation pieces that characterised assessment in the traditional methods of language instruction, tests included single short items that



were unrelated, since these would permit the designer potentially to test a bigger variety of components of language ability (Green 2014:181).

Problems were again evident in that many teachers were concerned about the absence of speaking and writing tasks. Moreover, test designers thought it too difficult to create tasks that necessitated the assessment of only one component at a time. In disagreement with Lado, Carroll (1961) therefore proposed integrative testing, arguing that the emphasis of language assessment should be students' ability to combine their language skills in such a way that they are able to understand the target language in its entirety. Termed the psycholinguistic-sociolinguistic phase, it favoured assessment techniques such as cloze procedure. A cloze test requires a student to repair a text in which words have been left out either by filling in or selecting the correct word from a list or giving any appropriate alternative, based on contextual clues (Green 2014:188). Although this type of assessment was well received, other types of assessments employed by test designers within this phase had their drawbacks. One of these was oral examinations, such as implemented in the traditional phase, that were still being utilised even though their results were unreliable. Despite these concerns, the phase still played a role in highlighting the quest for a single, general language ability. Thus Oller (1979:212) discarded discrete-point testing in support of integrative testing and coined the unitary competence hypothesis (Green 2014:197). He noticed that tests which were supposed to measure different language components frequently exhibited congruent results. He regarded this phenomenon as proof that language did not consist of different components which operated distinctly, but rather that the different components of language all work together to form a general language proficiency. Although very influential, the hypothesis did not stand the test of time (Green 2014:197).

In language assessments today, one cannot ignore the importance of each of these phases, as they paved the way for the communicative orientation which is currently

subscribed to by most test designers and language teachers. The communicative phase presents a new approach, which sees the functions of language as the entry point for test and curriculum design rather than the grammar and sounds of language. It regards language as in essence communicative, which in turn implies that language cannot be separated from the social context in which it appears (Green 2014:198; Blanton 1994:225; Bachman & Palmer 1996:62). Viewing language as embedded in a social context, this stance departs also from the importance attached in traditional language studies to individual expression. Rather, that expression is deepened to embrace shared expression or communication (Weideman 2009:63). Additionally, this would mean that language proficiency does not include only the correct use of language, but also knowing what to say, when to say it and to whom it should be said. The communicative approach is therefore innovative in that it represents an open and functional view of language ability in contrast to the restrictive view which dominated the earlier phases. This shift is evident in the acknowledgement that the objectives of language teaching have varied from focussing on teaching language that is aesthetically appealing to what they are today, where its objectives are in essence communicative. The shift is also a sign of global mobility: more than ever, people aspire to travel the world and to conduct business endeavours abroad through effective communication (Green 2014:173,175).

In the open view of language that is part of the communicative revolution, language is therefore not seen as purely expressive in nature, but also as communicative, as suggested by the communicative approach in language teaching. A table (2.4) by Van Dyk and Weideman (2004a:5) summarises the differences between the two perspectives:

<b>Restrictive</b>	<b>Open</b>
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 and</li> <li>○ negotiate human interaction</li> <li>○ in specific contexts</li> </ul>
Main function = expression	Main function = communication
Language learning = mastery of structure	Language learning = becoming competent in communication
Focus: language	Focus: process of using language

**Table 2.4: Two perspectives on language**

Consequently, the communicative approach or communicative language teaching (CLT) addresses the way in which we design language teaching in classrooms from the starting point of an open, disclosed view of language. Firstly, by making use of authentic texts in language instruction, CLT requires that real life language situations are recreated and students become familiar with the social context of such situations. Secondly, by using tasks which integrate the different language skills and media, students in CLT classrooms better experience the interdependence of “language skills” as they are used in combination during the process of communication (Green 2014:200; Weideman 2013a:13). For example, when you receive a written message from someone, you first read the content and process that before you can write a reply. Several different ‘skills’ are used to facilitate one functional action, which is that of responding to a written message. The emphasis is therefore on the purpose or function for which language is used.

This shift in perspective has also been influential in language testing. Initially, language tests were skills-based, general tests associated with methods such as discrete point testing. Nowadays, however, a skills-neutral approach (Weideman 2013a:14) may be favoured, with specific language tests which assess contextually specific language abilities. Bachman and Palmer (1996:75) explain that we should “not consider language skills to be a part of language ability at all, but to be the contextualised realisation of the ability to use language in the performance of specific language use tasks.” After surveying below principles traditionally

identified as crucial to the process of language test design, I will return to the relevance of these historical underpinnings of language teaching and testing for the selection and use of the assessment instrument that will be used in this study. First, however, a final observation needs to be made about the similarity in the phases of development of language teaching and testing designs.

#### **2.4 Language test development phases echo those of other applied linguistic artefacts**

We noted above (section 2.2) that there are certain design principles for language test and language course design that have derived from the history of these designs. In that sense both tests and courses are applied linguistic artefacts. For further evidence that language test design is a sub-discipline of applied linguistics, one should only have to look at the phases identified historically for the development of language testing, the approaches to teaching writing, and the evolution of applied linguistics as a whole, since they exhibit certain key similarities. Weideman (2006:150,152), for example, identifies several comparisons between the approaches to writing as presented by Lillis (2003) and Ivanic (2004) and the phases of applied linguistics, emphasising especially the shift from the initial focus on skills-based approaches to that of viewing language as dependent on social context. My focus, however, will be on the similarities between the traditions of applied linguistics and the development of language testing. Weideman (2013b:239) summarises the different phases of applied linguistics in the form of a table (2.5):

<b>Model/tradition</b>	<b>Characterised by</b>
Linguistic/behaviourist	“scientific” approach
Linguistic “extended paradigm model”	language is a social phenomenon
Multidisciplinary model	attention not only to language, but also to learning theory and pedagogy
Second language acquisition research	experimental research into how languages are learned
Constructivism	knowledge of a new language is interactively constructed
Post-modernism	political relations in teaching; multiplicity of perspectives
A dynamic /complex systems approach	language emergence organic and non-linear, through dynamic adaptation

**Table 2.5: Traditions of applied linguistics**

The first tradition of applied linguistics displays similarities with the psychometric-structuralist approach of language teaching and testing referred to above, in that both favoured behaviourist methods, where emphasis is placed on the four different language skills of reading, writing, listening and speaking (Weideman 2006:158). The most notable parallel, however, is between the linguistic “extended paradigm model” of applied linguistics and the communicative approach. Regarding language use as dependent on the social context in which it occurs, it presented a revolutionary, open view of language which stood in stark contrast to the restrictive view that dominated earlier phases of applied linguistics and especially the phases of language teaching and testing designs (Weideman 2006:159) that have already been referred to. This shift in the way we define language has prompted innovative approaches to the design and development of those solutions to language problems that we may consider to be applied linguistic artefacts. For example, second language acquisition research and constructivism are both approaches which derive from the extended paradigm model which sees language as a “social phenomenon” and have thus encouraged questions such as how children acquire new languages interactively, and how these languages are “interactively constructed”. These shifts are important for language test design, since such “approaches determine the

content, style, the what and the how of the solutions that are proposed”, within applied linguistics (Weideman 2006:147).

The postmodernist tradition of applied linguistics, the sixth of the styles of design identified in Table 2.5 above, is also of importance, since it opposes earlier modernist approaches. The post-modernist phase emphasises the accountability of a test designer for the designs that are developed (Weideman 2013b:243), or what Bachman and Palmer refer to as the consequences or impact of the test. This phase also indicates to what extent abusive, unequal political relations can influence the design of “accountable solutions for language problems” (Weideman 2013b:244; Rambiritch 2012:176). An example of this is given by Weideman (2006:148) when he explains that when such unequal political relations are institutionalised, this can cause immeasurable harm. We often find that when language learners are identified as having limited language proficiency, they are treated in accordance with their assumed limitations. Instead of providing them with a multitude of resources, extensive academic support and positive expectations, they are expected to fail, which brings about that they do not receive the additional assistance that they truly need. Currently, however, we are more aware than before of these injustices, and we have postmodernist approaches to testing and teaching to thank for this.

McNamara (2005:775) discusses the “social turn” in the design of applied linguistic artefacts that is due to post-modernism when he explains that we now view language tests and their results from a more critical perspective, since unfair language test results can have undesirable implications for test takers. Similarly, Shohamy (1997:340) discusses not only the importance of reliable and valid language tests, but also the bias which can be attached to the results of a language test. McNamara and Shohamy (2008:89) observe that in “most societies tests have been constructed as symbols of success, achievement and mobility, and reinforced by dominant social and educational institutions as major criteria of worth, quality and value”. It is therefore of the utmost importance that language tests are designed

which truly measure language ability in terms of the current day conceptualisations of language proficiency, that institutions where these tests are administered abandon the notion of viewing language tests as administrative burdens, and that the results obtained from language tests are approached with open-mindedness (McNamara 2005:776). Termed critical language testing, McNamara and his circle speak to the significance of the social and political context in which language testing and applied linguistics operate (McNamara & Shohamy 2008:93). Alongside other subfields of applied linguistics, the critical turn in language assessment signifies the shift from modernist to postmodernist approaches to design, confirming that language assessment is indeed a critical part of applied linguistic designs (Weideman 2013:243).

### **2.5 Validity and related principles traditionally identified as conditions for test design**

As our views on language teaching and testing have shifted, the principles essential for language test design have also undergone modification, and not surprisingly. Looking back on the history of language testing, the latter half of the 20<sup>th</sup> century has seen Messick's notion of test validity being acknowledged as the overriding principle for language test design (Messick 1980:1012; Weideman 2011:100). Du Plessis (2012:27) emphasises that Messick's main concern regarding test validity involves defining it as the appropriate and adequate interpretation of test results (Messick 1980:1014), as well as to raise awareness of the social implications that test results can have for test takers, education systems and the practice of language testing. Weideman (2012:4), however, points out that "to make validity dependent on interpretation runs the risk of downplaying the quality of the instrument", because if an instrument such as a language test is inadequate it would not matter how cautious or responsibly one approaches the interpretation of test results, the instrument would remain inadequate. For Weideman, the primary condition that a test must be effective or valid (2014:8) derives originally from the normative appeal

that issues from the link between the technical, leading function of a test and the physical aspect of energy-effect.

In an attempt to redefine the overriding principle of test design, as well as to simplify Messick's notion of validity, so as to identify more clearly some of the social concerns of language testing that are present in the latter's notion of consequential validity, Bachman and Palmer emphasised the notion of (technical) utility, an idea which places an emphasis on the usefulness of language tests (Bachman & Palmer 1996:9; Weideman 2011:102). Validity is in this view regarded as a component of the overall utility of language tests, as represented below:

$$\text{Usefulness} = \text{Reliability} + \text{Construct validity} + \text{Authenticity} + \text{Interactiveness} + \text{Impact} + \text{Practicality}$$

**Figure 2.2: Bachman & Palmer's model of test usefulness**

Whilst this model is an attempt at emphasising the necessity of validity as part of a test's usefulness, there are a number of arguments against viewing validity as the solitary principle which can lead responsible language test design. It should rather be regarded as one of the several principles or conditions that can lead responsible language test design (Weideman 2012:2). The model proposed by Bachman and Palmer in fact implies that other principles all play a vital role in the process that is language test design. Reliability, for example, refers to a test's ability to deliver more or less the same results for the same students when written on different occasions, which Du Plessis (2012:31) terms a "function of score consistency". Although inconsistencies cannot be eliminated altogether, it is possible to aspire to regulating and containing sources of inconsistency. Validity, on the other hand, is a term that has received a considerable amount of attention. Messick (1980:1019) saw validity as a holistic concept of the appropriate interpretation of test results (Weideman 2013a:101). However, validity may rather be viewed as belonging to a comprehensive and "systematic set of principles" which in this case refers to the norm that concerns the technical force or effect of a test (Weideman 2012:8). A



further principle identified in Bachman and Palmer's model is that tasks that are required within a test should resemble real life tasks that could be required of the test taker. In essence, this is what the principle of authenticity refers to. Interactiveness, on the other hand, requires a student's extended engagement and use of a specific language ability to complete a test successfully (Du Plessis 2012:34), whilst impact refers to the social consequences that test results could have. A test and the interpretation of its results should therefore be handled with the utmost of care and responsibility. Lastly, practicality suggests the availability of the necessary resources to design and administer a test (Du Plessis 2012:34), a principle that Weideman (2014:8) derives from the link between the technical design function of the test and the economic dimension of reality.

Although some of the principles presented by Bachman and Palmer have received less prominence in actual test design, their work is still of relevance. What is noteworthy, however, is that in language test design Messick's influence has been such that it has stimulated a quest for one overriding principle. There are persuasive arguments, however, that subsuming all principles for responsible test design under one principle merely clouds the clear conceptualisation of the others. Moreover, as Weideman (2012:2) remarks, if there is one overriding principle, it should preferably be related to the qualifying technical design function of a test, since that aspect of the artefact is its guiding and leading function.

In the next section, I shall discuss how we may start by identifying three key principles for the responsible drafting of language tests, based on the notion, confirmed by the above analyses, that language test design belongs within the discipline of applied linguistics (Weideman 2011:100). In the first instance a theoretical justification of the purpose of the test must be articulated. Test designers must be certain of what they want to measure and why it is necessary to measure it. The second principle proposes the responsible interpretation of test results subsequent to the administration of a test, whilst the last principle calls for a

consistent and stable measuring instrument. These three principles are a selection of the constitutive and regulative conditions for test design that were referred to above, and will again be considered below in section 2.6. Weideman (2011) is of the opinion that a discussion of key principles can be justified in that they have figured most prominently in the historical development of language test designs. In what follows I shall therefore discuss a number of key principles of test design that have conventionally been identified, but as articulated through distinctive principles for responsible test design that derive from a more comprehensive framework of applied linguistic design principles.

## **2.6 Key principles for the design of a language test**

According to the key principles identified by Weideman (2011), the process of test design must begin with the articulation of a test's construct (for a potentially contrary view, see discussion below, and Chapelle 2011). What is conventionally termed "construct validity" or even "theory-based validity" (Weir 2005) is, in this view, the theoretical justification of a test design. As we have noted above, this requirement derives from the link between the technical and analytical modes of reality. The construct of a test must include a clear theoretical definition of the ability that is intended to be measured by the test (Weideman, Patterson & Pot 2014:2; Bachman & Palmer 1996:66; Shohamy 1994:341) which can then be referred to as the theoretical rationale for a test (Weideman 2014). Defining an ability is, however, a task that must be attempted in a manner that refers to theoretically current views of language. Test designers must take care in determining exactly what the ability entails and how the ability should be measured, and with reference to currently acceptable perspectives on language. Such a clear definition of an intended ability is of crucial importance, as it supports the fulfilment of other criteria: achieving a reliable, valid, and technically effective test design (Weideman, Patterson & Pot 2014:2), for example. Academic literacy tests, for instance, are designed with a very specific purpose in mind, which is the

measurement of students' academic literacy levels. Designers of such tests need to be aware that academic literacy derives from a student's ability to engage with academic discourse. Therefore, designers need to know which practices constitute "critical features of academic discourse" (Patterson & Weideman 2013b:126) and consequently, what academic discourse demands of users (Patterson & Weideman 2013b:136), before they are able to construct a test that is truly representative of the nature and demands of academic discourse, and also has "authenticity" as defined by Bachman and Palmer (1996:23).

The notion of beginning with a construct when working with language test design, however, is not accepted without critique. Chapelle (2011:19), for instance, refers to the argument of Kane (2006). Kane downplays the idea of starting with a construct, although he does not completely disregard the idea of having a construct. The case he makes revolves around the significance of the interpretive argument which could guide test markers to interpret test scores through a supposed simpler approach. Chapelle (2011:26) presents a table which depicts that the interpretive argument distinguishes between the domain (test designers should be aware of their expectations regarding a target domain), the universe of generalisation (test designers should refrain from using an over-generalised sample of language use from their target domain) and the sample of observations (the size of a sample and observations made regarding a specific sample should be approached with caution). Furthermore, test designers should be aware of the assumptions which often guide test scoring, include a theory-defined construct and be aware of the possible consequences of test scoring (Chapelle 2011:26). According to Kane (2006), these principles would all contribute towards the validity argument. I prefer the idea of starting with a construct, however, since it acts as a guide to test designers, constantly reminding them of the intended ability to be measured by the test.

One should therefore be mindful of the fact that a test construct does not embody only the definition of ability and intention of the measurement at hand, but also

plays a role in acting as a blueprint for the structure of a test. This blueprint would include test specifications derived from the test construct, which determine what task types and assessment formats should be included in a test (Weideman, Patterson & Pot 2014:2) as well as their relative weight. Another indication that a test is valid includes that the task types should be aligned with its test construct. Test designers should, however, take into account that certain challenges can influence the design of a test, such as “administrative, logistical, financial and other resource limitations” (Weideman, Patterson & Pot 2014:2), those which Bachman and Palmer (1996:35) call practicality. Of importance is also comprehending that the processes of design and development are an ongoing cycle. One should always seek to improve the test at hand (Weideman, Patterson & Pot 2014:20). Despite practical constraints that may impede the process, designing a test according to an appropriate test construct is still the best approach to ensure that a test is responsibly developed, as it provides test designers with a clear explanation of what they need to measure. If well articulated, it also sets up a basis for ongoing reflection about what will be measured (Patterson & Weideman 2013a:107), as I will discuss in more detail in the next chapter.

Except for having a theoretically defensible test construct, a test also needs to be a consistent and stable measurement instrument. Termed the technical consistency or reliability of a test, a reliability index of a test should indicate to test designers how consistently a test measures. This requirement clearly reflects the normative appeal that the connection between the technical and kinematic modalities exercises. The consistency of a test is found, *inter alia*, in the performance of the items that make up the task types which constitute a test, and is conventionally measured by a reliability index such as Cronbach’s alpha or Greatest Lower Bound (Weideman 2011:105). The tests utilised for this study have been analysed using Cronbach’s alpha and usually exceed the 0.7 score which is regarded as the benchmark for reliability. The performance of task items should also be judged as regards the average number answers that are correct and the discrimination value of each item

(Weideman 2011:105). After these analyses have been carried out, test designers can identify which task types or items need to be eliminated or improved. A test can then be piloted and refined to a further extent (Weideman 2011:106), and the refinement so achieved usually contributes to an increase in reliability of the test.

The last key principle refers to the appropriate interpretation of test results. This principle derives from the connection between the technical and lingual dimensions of reality. The responsible interpretation of test results are dependent on the clear articulation of the construct. If one does not know clearly what is being measured, one cannot give a lucid interpretation of the results of the assessment (Weideman 2011). On the other hand, the irresponsible interpretation of test results can cause immeasurable harm for test takers and should be avoided at all costs. Academic literacy tests are designed to assist struggling students, and not to stigmatise them (Weideman 2011:107). Therefore, instead of having simple pass or fail results, a risk band system was introduced for the tests to be used in this study. According to this system, test takers are classified according to the level of risk they show as regards their ability to handle academic language. A risk level of 1 would indicate that a test taker runs a very high risk of not being able to engage with academic discourse whilst a risk level of 5 would indicate that a test taker shows little to no risk in engaging competently with academic discourse. This approach therefore minimises the risk of stigmatization (Shohamy 1994:340) that can be caused by a less than kind interpretation of the results of a test, for example, when people automatically interpret a score less than 50% as a 'fail' and those above 50% as a 'pass'. Risk bands give interpretations across a spread of results so that the notion of publishing test results in these risk bands also contribute to a potentially more useful and encompassing expression of the meaning of these results (Weideman 2011:107).

## 2.7 Other principles

Beyond the scope of the three key principles mentioned above, other principles are also of importance to the design of a language test. In section 2.2 I presented a list of principles which emanates from the relationship between the leading technical functions of this applied linguistic artefact and other modalities of experience. From each of the reflections of other dimensions in the technical mode of experience there issues a normative appeal which generates a condition or principle important to the design of a language test (Weideman 2014:8), and in what follows I will give a preliminary assessment of how the tests to be used in this study conform to these conditions.

By providing evidence in the form of statistical analyses, I hope to demonstrate that the tests used for this study do in fact measure what they set out to measure, which is academic literacy levels. The process of justification begins by bringing together several empirical data sets that will provide evidence for the validity of the measurement. As we shall note in the next chapter, such data might include empirical analyses done through programs such as Iteman and TiaPlus, of the reliability of these tests, the discrimination and facility values of the items in the tests, and the subtest intercorrelations, and others. Bringing together in a systematic way a multiplicity of data to argue for the validation of a test seeks to fulfil the first criterion referred to in the list presented on pages 16 to 17, which is to *Systematically integrate multiple sets of evidence in arguing for validity of the test or course design*.

The test takers and the recipients of the test results will be aware of the limitations of the test, which are circumscribed by the measurement of academic literacy levels, and not the assessment of another specifically defined language ability, such as the ability to handle, for example, economic or social discourse. The test takers will also be informed of these limitations during the piloting of the tests, limitations which issue from the link the spatial dimension has with the technical mode of

experience, which asks to *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.*

The results obtained through the pilot tests should display consistency. This consistency can also be seen as the technical reliability of a test and emanates from the link between the kinematic and the technical dimensions of reality. A test measures consistently when it can be given to the same group of students on different occasions and the results stay more or less the same. As I am using two tests for my study, both designed more or less according to the same test construct and specification set, I will be able to argue the presence (or absence) of technical reliability for this specific test construct which refers to *Ensuring that the measurements obtained and the instructional opportunities envisaged are adequately consistent.* Once again, this argument for using the tests will rest upon, and would need to be backed up by adequate empirical analyses.

As the tests and their format are based on a theoretically justified construct and will yield adequate results, I should be able to provide evidence for the validity of the tests (*Ensure effective measurement or instruction by using defensibly adequate instruments or material*). The adequacy of the instrument is related to the technical force of the tests to deliver results, as well as to their reputation as tests that have been used to good effect before. This principle derives, as I have observed above, from the reflection of the physical modality in the technical.

The tests constitute various and different task types and items as to ensure a more inclusive measurement instrument. The requirement to *Have an appropriately and adequately differentiated course or test* derives from the link between the technical function of the test and the organic mode of experience. A differentiated test, in contrast to a monotone design, utilises a variety of task and item types in order to achieve its full potential to measure.

In addition, the tests should be intuitively appealing, as they are well-structured, clearly numbered and presented in a legible font type and size. Care has also been taken to eliminate mistakes, whether it be grammar, spelling or numbering errors. In short, the tests have what is known as “face validity”, which can be perceived as the link between the sensitive dimension of reality and the technical mode of experience. This then refers to the condition of *Making the course or the test intuitively appealing and acceptable*.

A theoretical justification for the test is provided by what I believe is a significant component of our current school language curriculum, and that is the instruction of academic discourse at school level. This does not mean that schools are teaching academic discourse at school level, but as has been noted above, with reference both to the Home Language curriculum (Department of Basic Education 2011) and its possible interpretations (Du Plessis, Steyn & Weideman 2014; Du Plessis 2014b), the development of the ability to use language for academic purposes fits entirely into the Home Language curriculum. Moreover, the tests were designed according to a specific definition of academic discourse and what academic discourse demands of its users: the interactive ability to handle language in higher education (Weideman & Patterson 2013a:109). We have here then the leading design rationale for the tests which refers to the norm of *Mounting a theoretical defence of what is taught and tested in the most current terms*. The definition is current, in the sense that it proceeds from recent views on teaching and testing communicatively, which was discussed above in section 2.3. I return in the next chapter to a more detailed consideration of this theoretical justification for the design of the test that links the analytical dimension to its leading technical function.

The results obtained from the tests will be valuable to the core of this study as they will indicate whether an academic literacy test, designed according to a very specific test construct, could be a more useful indicator of students’ ability to handle



academic discourse than current school language marks. This condition emanates from the link between the lingual and the technical modes of experience (*Make sure that the test yields interpretable and meaningful results, and that the course is intelligible and clear in all respects*).

Information regarding the tests and their construct will be available to anyone enquiring about them. This information can be found on the ICELDA (2014) website which can be accessed by anyone at any given time. Here we find a link between the social and technical modes of experience (*Make not only the course or the test, but information about them, accessible to as many as are affected by them*).

The results of the tests will be useful since they can be used as an academic literacy level indicator (*Present the course and obtain the test results efficiently and ensure that both are useful*). Moreover, the condition of utility, that derives from the link between the economic and technical modes, is also satisfied by the frugal use that the tests make of resources: their multiple choice formats, for example, ensure that they can efficiently measure in minimum time.

The next condition is one that is close to the heart of the study (*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*). As has been noted above, there is no separate instruction of academic literacy at school. One could argue, however, that the tests are indeed aligned with the demands of what is lingually required of pupils across all their subjects. Whether there is currently any alignment in the existing teaching of languages at school and the overall demands of having an ability to use language for academic purposes across all subjects is the critical question that I hope this study will be able to answer, or answer at least in part. Should there be a misalignment between language teaching at school and academic literacy, this is likely to be identified as a shortcoming. Certainly, as Weideman, Du Plessis and Steyn (2014) have shown, the new curriculum (CAPS) for teaching

language requires a high degree of competence in language, a competence that includes academic discourse. This does not mean, of course, that it is actually sufficiently taught at schools.

Details regarding the tests' uses and aims will be available to the public and can once again be found on the ICELDA (2014) website (*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*). This principle connects the technical design to the political or juridical sphere of experience.

The tests have been designed with the utmost of care and consideration as to ensure that they adhere to a certain standard and quality which in turn facilitates fairness in the implementation of the tests (*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*). The requirement in this case stems from the link between the leading technical function of the test in its interaction with the ethical sphere.

As with the previous condition, the tests have been designed with the aim of ensuring not only that the tests and their results can be justified, but also that they have stood the test of time, and have built a reputation of being highly trustworthy indicators of the language ability being measured (*Spare no effort to make the course and the test appropriately trustworthy and reputable*). The condition here relates to linkages between the technical and the aspect of belief.

The tests should be scrutinised once more after they have been piloted and the results made public as to review whether they indeed conform to these conditions. If there are then any conditions to which the tests can more fully adhere, these can be addressed subsequently in order to improve them.

## **2.8 Conclusion**

In the design of an academic literacy test, one condition has been viewed as particularly prominent, perhaps more so than the other key principles and other significant principles discussed above. This is that the construct according to which the tests were designed has to be carefully articulated, since from the construct is generated a list of components from which, in turn, arises a list of specifications and task types. For the sake of my argument, especially that of pursuing language testing research within the scope of applied linguistics, I will discuss in further detail the principles for test design, components of the test, as well as its specifications and task types, in the next chapter.

## Chapter 3

# The test construct and its operationalisation: Design principles and phases

### 3.1 Introduction

In the previous chapter it was observed that one of the key principles of language test design requires that responsible test development includes constructing tests on the basis of a clear definition of what it is they intend to measure. This is referred to as the construct or blueprint (Van Dyk & Weideman 2004a:1) of a test, and an academic literacy test would not be an exception. Correspondingly, one of the guiding conditions for test design as set out in the previous chapter also states that test designers must be able to *Mount a theoretical defence of what is taught and tested in the most current terms* (Weideman 2014:8). This is a prominent design principle, which not only emphasises the importance of utilising current theoretical approaches when teaching or assessing language at school or university, but also emphasises that the testing of language ability must reflect theoretical definitions of language that are in step with recent thinking.

Also mentioned in the previous chapter is the reciprocal relationship that exists between the norms of the technical design of an applied linguistic artefact and the factual end-user format of the same artefact. Viewed thus, a language test is therefore an end-user applied linguistic artefact that is normed or conditioned by its construct and specifications, which act as a theoretical justification for the design of the test (Weideman 2011:101). It is for this reason that we will review the current construct and its lineage below, after which its significance for the assessment instruments to be used in this study, based on certain principles that guide the selection of such assessment instruments, will be discussed in further detail. Thereafter, the components relating to academic literacy deriving from the

construct will be examined, as well as the specifications and task types that flow from it.

### **3.2 The evolution of the construct**

Van Dyk and Weideman (2004a:1) remark that the low academic literacy levels of South African students are regarded as a key reason contributing to the overall poor academic success many of them experience in higher education environments. It is no surprise then that academic literacy tests are employed by South African universities in their attempt to identify at risk students who would possibly need additional academic support. In 1999 the Unit for Language Skills Development (ULSD) was introduced at the University of Pretoria (UP) as an attempt to address the institutional concerns about low academic literacy levels amongst first year students who enrolled for English-medium instruction as non-native speakers of English. Since 2000 until recently, students therefore first had to be declared language proficient before they could obtain a degree at the university, and for some time the UP made use of the English Literacy Skills Assessment for Tertiary Education (ELSA PLUS) to assess the language ability of their new students (Van Dyk & Weideman 2004a:2). In this sense, ELSA PLUS was utilised as a placement test instead of an access requirement, since students were tested after their arrival, and then placed into academic literacy programmes. They were thus not denied access to the university. This kind of assessment is what I referred to in the first chapter as being a lower stakes test, though the stigma that might be attached to its results would, in the eyes of some, make it a medium or even high stakes test.

ELSA PLUS was a test developed by the Hough and Horne consultancy and consists of seven sections which include phonics, dictation, basic numeracy, reading comprehension, language and grammar of spatial relations, a cloze test and a section on vocabulary in context. The test was regarded as administratively efficient, since it is only an hour long. It was also claimed to have empirical validity, objective scoring and lacking in cultural bias. Yet one can clearly recognise that the

test regards language from a restrictive view: the belief that language consists of sound, form and meaning (Van Dyk & Weideman 2004a:3). In a previous section (2.3) I have emphasised the limitation that such a restrictive view of language places on the crucial role of communication that language fulfils, also in academic settings. A limited perspective on what language is, when translated into a test construct, cannot measure language ability in line with current views of language.

At about the same time that the UP was taking these measures, the University of Cape Town's (UCT) Alternative Admissions Research Project (AARP) employed the Placement Test in English for Educational Purposes (PTEEP), based loosely on the construct by Bachman and Palmer (1996:67) which proposes that language ability is dependent on language knowledge and strategic competence (Van Dyk & Weideman 2004a:8), as indicated in the table below:

<b>Language ability</b>		
<b>Language Knowledge</b>		<b>Strategic competence</b>
<b>Organisational knowledge</b>	<b>Pragmatic knowledge</b>	<b>Meta-cognitive strategies, including</b> <ul style="list-style-type: none"> <li>○ Topical knowledge</li> <li>○ Affective schemata</li> </ul>
<b>Grammatical</b> <ul style="list-style-type: none"> <li>○ Vocabulary</li> <li>○ Syntax</li> <li>○ Morphology</li> </ul>	<b>Functional knowledge</b> <ul style="list-style-type: none"> <li>○ The use of language to achieve goals</li> </ul>	
<b>Textual</b> <ul style="list-style-type: none"> <li>○ Cohesion</li> <li>○ Rhetorical or other</li> </ul>	<b>Sociolinguistic knowledge</b> Or: <ul style="list-style-type: none"> <li>○ Dialects</li> <li>○ Registers</li> <li>○ Idiomatic expressions</li> </ul>	

**Table 3.1: Bachman and Palmer's construct**

Many problems have been raised concerning this construct, of which the argument time and again returned to whether it is feasible to distinguish so strongly between many of the elements listed in Table 3.1 above. For example, one of the questions

raised by Van Dyk and Weideman (2004a:9) asks whether choice of register (located under sociolinguistic knowledge) could be separated from “the selection by a language user of a particular organisational form” (organisational knowledge). Nonetheless, AARP reinterpreted this construct for the development of PTEEP, placing the main focus of the test on the reading and writing abilities of first year students. The test was, however, administratively not efficient enough, since it took too long to administer to large groups and even longer to mark. Also, contrary to what the name states, the test was used as an access instrument and not as a placement test (Du Plessis 2012:47).

The inefficiencies and inadequacies of Both ELSA PLUS and PTEEP prompted the quest for an alternative, current test construct which would yield a test that is administratively efficient, and which could be used as a placement test rather than an access test. Taking into account the construct by Bachman and Palmer (1996) as outlined above in Table 3.1, the test construct of PTEEP, and a definition of academic discourse presented by Blanton (1994), the first versions of the Test of Academic Literacy Levels (TALL) and the Toets van Akademiese Geletterdheidsvlakke (TAG) began to be conceptualised (Van Dyk & Weideman 2004a:9). Blanton’s definition articulates an idea of academic discourse that proceeds from a more open and disclosed view of what language ability is in that particular setting (Van Dyk & Weideman 2004a:7). According to Blanton’s definition of academic literacy (1994:226), students should be able to:

1. interpret texts in light of their own experience and their own experience in light of texts;
2. agree or disagree with texts in light of 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 of the above;
7. talk and write about doing any or all of the above;
8. do number 6 and 7 in such a way to meet the expectations of their audience.

One concern with Blanton’s construct is that it does not adequately articulate the cognitive ability or all the subskills which accompany our engagement with

academic discourse (Van Dyk & Weideman 2004a:9). What is more, the definition seems more difficult to operationalise than alternatives (Patterson & Weideman 2013b:138). For this reason, the construct of TALL and TAG includes both the measurement of language ability and the measurement of cognitive ability when engaging with academic discourse. As background to that discussion of the final articulation of the construct, however, I shall firstly discuss a number of further principles for test selection and design, with a view to understanding the current construct more clearly.

### **3.3 Principles for test design and selection**

In order to design and select tests through which the research aims of this study (section 1.3) can be achieved, and which still conform to the chosen test construct that is to be articulated, it is necessary to investigate and identify first the principles that such tests should satisfy.

We know that language is context specific and cannot be separated from its social setting (Patterson & Weideman 2013a:109), which implies that context specific skills are needed by students to read and write successfully at university level. When we then attempt to design, evaluate or select an appropriate academic literacy test, we firstly need to ask what makes academic discourse different from other types of discourse, and secondly what types of tasks would be able to determine effectively how well students engage with various texts that are encountered by them in academic discourse. Weideman (2014:8) asserts that language is not just factual in nature, but rather bound by certain normative principles of the social relationships in which the language in question is embedded. These differentiated relationships show great variation and the conditions under which language operates in them may indicate that language used in such spheres may therefore be typically stamped as a logical, aesthetic, academic, social, political, ethical or economic type of language. These conditions for the spheres in which language is



used determine the “nature of factual texts, namely the concrete language used in a specific context or situation” (Weideman 2011:6), and the types of language that are observed in such contexts are referred to as a typically differentiated variety of material lingual spheres. We can therefore argue that academic discourse is specific to a certain context, as it is bound by certain normative principles that are characteristic of the nature of academic institutions, such as universities or other tertiary educational establishments. Consequently, Patterson and Weideman (2013a:118) propose the following definition of academic discourse:

Academic discourse, which is historically grounded, includes all lingual activities associated with academia, the output of research being perhaps the most important. The typicality of academic discourse is derived from the unique distinction-making activity which is associated with the analytical or logical mode of experience.

From the above mentioned definition we can infer that academic discourse signifies the act of distinction-making, which affirms that a test that attempts to evaluate academic literacy has to proceed from the analytical mode that qualifies academic endeavour. Academic literacy tests, as applied linguistic artefacts, carry the imprints of lingual conditions, which are typified by the sphere of discourse that they regulate. In a word, an academic literacy test has to be specific and contextual. Weideman (2014:7) emphasises this by pointing out that we need to be aware of the “conceptualisation of the limits of the artefact and what it can accomplish”, as one universal language test cannot measure all the possible language abilities that an individual can perform. An academic language test cannot measure, for example, how well a person would be able to deliver a sermon, since academic language proficiency and one’s capability of delivering a sermon successfully depend on typically different lingual abilities. Moreover, when one attempts the design of an applied linguistic artefact, such as an academic literacy test, one would methodically start by identifying analogical constitutive and regulative moments relating to the technical qualification of the test, each of which yields a design principle for such a test.

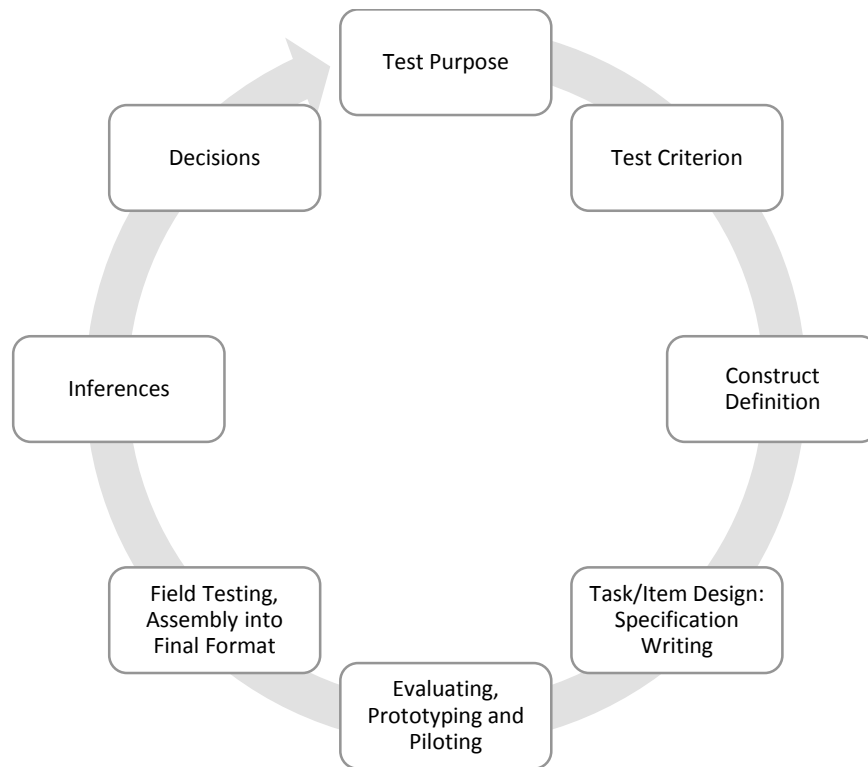
A model proposed by Weideman (2014:7), acting as a framework of design principles for all applied linguistic instruments, asserts that the qualifying function of an applied linguistic artefact such as a language test can be found in its technical dimension, which constitutes the leading function of the artefact. In turn, the foundational function embodies the analytical dimension of the design. A reciprocal connection exists between these two terminal functions, even though the leading technical design function takes priority over the theoretical rationale, since the technical design is the qualifying function of the instrument (Weideman 2014:8). The leading technical function of the test yields certain design principles which guide the design of the test. These design principles originate from the connections that the technical has with each of the other dimensions of reality. Every design principle yields a normative condition which guides the responsible design of a test, and can be regarded as either a constitutive or as a regulative principle (Weideman 2014:7). The principle of a test possessing “face validity” or intuitive appeal, for example, is a constitutive design principle, in that it requires that the test should be “intuitively appealing and acceptable” (Weideman 2014:8). The technical also has connections with other dimensions of experience which yield regulative principles for the design of the test. For example, the social context of test takers must be taken into account when a test is implemented (Rambiritch 2012:189; Weideman 2014:8), as does its fairness, accessibility and the integrity of its use.

The focus of this chapter is on the argument that an academic literacy test principally has to meet the requirements that are set out in its design rationale, which ultimately provides a theoretical justification for the design of the test and what the artefact or test is supposed to measure. The design rationale refers to what is called the test construct, which also encapsulates the purpose of the test, a point that I shall subsequently return to. Van Dyk and Weideman (2004b:17) have listed the possible components that are indicative of a proficient level of literacy in academic discourse, and these will be discussed in greater depth and detail in section 3.4 below.

The process of drafting the tests that I have chosen for this study, which will be discussed in detail in Chapter 4, exhibits the various stages or phrases of the test design cycle articulated by Fulcher (2010). Fulcher's perspective on the test design cycle emphasises the need for starting by identifying a test purpose, which in turn will guide the design of the test at hand. This includes identifying the target domain, defining the test construct, designing the tasks/items that will be included in the test, the piloting of the test and lastly the refinement of the test (Fulcher 2010:95-127). The initial phase of the cycle represent stages in the ongoing and further operationalisation of the construct on which the test is to be based.

### **3.4 The operationalisation of the current construct: Design phases**

The design procedure employed in the tests used for this study provide a good illustration of Fulcher's design stages. Fulcher's stages of the test design cycle (2010:291) capture how responsible test design and test development are phased. He emphasises the need for starting by identifying a test purpose (construct), which in turn would guide the design of the test at hand. Weideman (2015:78) articulates this process slightly differently. Firstly, one identifies a language problem after which the technical imagination of the test designer is coupled with the theoretical knowledge, among other things, of language use in the domain to be assessed, in order to come to an appropriate solution. This renders an analytically founded rationale for the design of the test which constitutes its theoretical justification, leading to the final articulation of the test's construct. Below, however, is a schematic representation of the cycle as articulated by Fulcher (2010:92):



**Figure 3.1: The test design cycle**

In the current case, a definition of academic literacy was formulated which guided the subsequent articulation of the test construct and its components. In order to operationalise the various components of the test construct, the task types and subtests which could potentially measure these listed components, as well as the item specifications of elements of these task types, were then identified. Following the identification of the test construct is therefore the identification of the components that would realise the construct. Explained in another manner, the construct components are those elements of language ability that are acknowledged to be necessary for the successful performance of the ability measured in the test, thus what ‘skills’ a student would need to engage successfully with academic language. Once the construct components have been identified, task types need to be recognised that would measure those components, and lastly, after the task types have been established, they need to be refined through item specification. Item specification relates to how many questions of a certain task type will be included, and what format the item would take, such as multiple choice questions or open-ended questions, for example (Van Dyk & Weideman 2004b:17). The

operationalization of the construct is thus accomplished in several sub-stages of further specification.

### **3.5 Test purpose and construct definition**

The initial stages of test design, referred to in the previous section, are worth considering in more detail, especially as they relate to the stages in the design of the tests used for this study. The design process starts with the ability that the tests are supposed to measure, as well as identifying the components unique to that ability. In this case the test purpose is to assess academic literacy, and to do that through the activities that one would typically use when engaging with language in the domain of academic discourse. Patterson and Weideman (2013a:118) provide the following definition of academic discourse based on the types of activities one would exercise when engaging with academic discourse:

Academic discourse, which is historically grounded, includes all lingual activities associated with academia, the output of research being perhaps the most important. The typicality of academic discourse is derived from the unique distinction-making activity which is associated with the analytical or logical mode of experience.

This definition proposes that the distinction-making activity accompanying academic discourse constitutes what makes academic discourse unique, which at the same time also discloses that academic discourse is a distinctive language with its own conditions, different to other lingual spheres, as opposed to earlier definitions which often took a closed view of language, regarding it as consisting of sound, form and meaning (Weideman, Patterson & Pot 2014:5). The notion of material lingual spheres, mentioned earlier in this study (section 1.1), demonstrates that as people move from one social context to another, they adapt to the situation in which they find themselves and to the conditions that govern language use in that setting (Patterson & Weideman 2013a:109). It is for this reason that one would have to acknowledge that academic discourse is a distinctive social context with its own distinctive prerequisites.

The construct deriving from this specific definition of academic discourse is therefore innovative, as it has undergone the changes that the shift in focus of language instruction and assessment have brought on. This construct acknowledges that language is contextual and emphasises that language cannot be separated from the situation in which it occurs or from the people using the language in a specific setting (Patterson & Weideman 2013a:109). Therefore, a test based on such a definition should take into account the distinction-making activity which Patterson and Weideman (2013a:118) identify as unique to academic discourse, acknowledging that it plays a crucial role in the design of a test. Alongside the distinction-making activity, related components of academic discourse can potentially also be specified. Weideman, Patterson and Pot (2014:7) provide a comprehensive list of activities, or components, which suggests that students who are academically literate should be able to:

- 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 distinctions between essential and non-essential information, fact and opinion, propositions and arguments; distinguish between the cause and effect, classify, categorise 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 purposes 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 the sentence.

More components have since been added to this list in order to make it even more comprehensive. However, the tests used for my study have been designed with the above list of components in mind. When this initial list of components was discussed with various academics and in publications across several disciplines, it

was found that “the elements above not only constitute a number of essential components of what academic literacy entails, but resonate strongly with what academics across the disciplinary spectrum think constitutes the competent use of academic discourse” (Weideman, Patterson & Pot 2014:7).

An advantage of articulating so carefully a test construct which is based on current views of the critical functions of language in a specific context is the possibility of positive washback. When language teaching and testing align, we encounter washback (Brindley 2002:467), which also relates to one of the aims of this study: the appeal to have academic discourse taught (or at least be attended to) and tested at school level so as to prepare students for the demands of tertiary study. Van Dyk and Weideman (2004a:11) explain that washback is evident when looking at the construct of the tests used for this study. This is because the construct is developed according to current views on language teaching and testing, and therefore it echoes what should eventually be taught in academic literacy development classes.

The biggest challenge, however, pertains to how one would translate this construct into task types that will allow mastery of its components to be assessed. In the case of an academic literacy test, the solution is to turn such task types into a range of subtests. Subtests constitute the range of activity types found within a test. Several of the current tests within ICELDA’s range of academic literacy tests make provision for testing the various components of the construct in a number of subtests. This is discussed in greater detail in the next section, with attention to the different focus that each subtest potentially allows.

### **3.6 Specifications and task types**

With reference to the operationalization of a test construct, Van Dyk and Weideman (2004b:17) mention that a true challenge for test designers comprises the alignment between the construct of the test and its subsequent specifications in the process of

operationalization. Achieving such alignment helps to ensure that the test measures what it is supposed to measure. It should be noted that each component of the construct can potentially be measured by more than one task type. For example, academic vocabulary comprehension can be tested through a modified cloze test and in longer reading passages, as well as in a shorter format with less contextual embeddedness. In other words, one task type can yield information regarding various abilities (Weideman, Patterson & Pot 2014:8). The table of task types shown below (Table 3.2) is therefore useful to test designers, since it indicates that a combination of task types can measure more than one ability at a time, and may subsequently help designers to develop tests with a positive internal correlations among the subtests, a point I shall return to below. Table 3.2 below contains a list of test specifications and task types as presented by Van Dyk and Weideman (2004b:18-19) in relation to the components of the construct (left column) that was articulated in the previous section:

<b>Specification/component</b>	<b>Possible Task types</b>
Vocabulary comprehension	Vocabulary knowledge Dictionary definitions Cloze C-procedure
Understanding metaphor and idiom	Longer reading passages
Textuality (cohesion and grammar)	Scrambled text Cloze C-procedure (perhaps) Register and text type Longer reading passages Academic writing tasks
Understanding text type (genre)	Register and text type Interpreting and understanding visual & graphic information Scrambled text Cloze procedure Longer reading passages Academic writing tasks (possibly also) C-procedure



Understanding visual & graphic information	Interpreting and understanding visual & graphic information (potentially) Longer reading passages
Distinguishing between essential/non-essential information	Longer reading passages Interpreting and understanding visual & graphic information Academic writing tasks
Numerical computation	Interpreting and understanding visual and graphic information Longer reading passages
Extrapolation and application	Longer reading passages Academic writing tasks (Interpreting and understanding visual & graphic information)
Communicative function	Longer reading passages (possibly also) Cloze, scrambled text
Making meaning beyond the sentence	Longer reading passages Register and text type Scrambled text Interpreting and understanding visual & graphic information

**Table 3.2: Specifications and task types**

The tests designed by ICELDA generally constitute a number of subtests. The current Test of Academic Literacy for Postgraduate Students (TALPS), for example, starts with a Scrambled text which, as can be seen from Table 3.2, assesses cohesion and grammar, understanding text type, a student's awareness of meaning beyond the sentence, and possibly also the communicative function of language. The second subtest includes questions pertaining to a given graph, which is labelled as Interpreting and understanding visual and graphic information in Table 3.2 above. This subtest tests meaning beyond the sentence, extrapolation and application, numerical computation, understanding visual and graphic information, understanding text type and a student's ability to distinguish between essential and non-essential information. The next subtest assesses Vocabulary knowledge through multiple choice questions, which relates to vocabulary comprehension. The fourth subtest requires of students to match different Text types to one another,

which assesses understanding of text type, making meaning beyond the sentence and possibly also textuality. The following subtest is the longest and constitutes a substantial reading passage with questions. This subtest potentially assesses understanding metaphor and idiom, cohesion and grammar, understanding text type, numerical computation, extrapolation and application, the communicative function of language, making meaning beyond the sentence, distinguishing between essential and non-essential information, and possibly also understanding visual information. The sixth subtest deals with Grammar and text relations in the form of cloze tests, which relate to vocabulary comprehension, textuality, understanding text type and possibly also the communicative functions of language. The last subtest is an Academic writing activity based on the texts presented throughout the test (the test has a theme), and assesses textuality, understanding text type, distinguishing between essential and non-essential information, and extrapolation and application.

Since there is a distinct difference of ability level between what TALPS should measure and a school-level test, the tests used for this study have five subtests (which will be explained in more detail in section 4.3.1). The first subtest relates to sequencing and is therefore aimed at assessing a student's ability to identify the succession within a text, in other words, a text's "systematic progression" (Patterson & Weideman 2013b:140). Following the first subtest is a section on academic vocabulary which can be related to more than one of the listed components, but its primary focus is on the relation between academic terminology and how it enables the critically important distinction-making activity mentioned earlier (section 3.2). The third subtest is aimed at the interpretation of graphic and visual information, which also assesses a student's ability to make inferences and identify evidence for an argument. A longer reading passage is presented as the fourth subtest, which asks questions to test text comprehension. The last subtest relates to grammar and text relations and assesses more than one component,

including cohesion, syntactic connections and vocabulary (Patterson & Weideman 2013b:140-141).

Rambiritch (2012:186) explains that test designers should pay attention, once a test has been put on trial, to the internal correlations of the subtests, in this case, how well the subtests within a test correlate with each other and with the test in its totality. When correlated with each other, test designers hope to see that correlations occur in the sense that a component of the construct is potentially measured in various ways, but on the other hand, one would also like to see that the correlations are not too strong, so as to ensure that each subtest truly tends to assess a different side of the language ability which the test is set to measure. This is true especially when we look at the list of various components of academic discourse above. Ideally a test should assess all of the components, but in various ways. For this reason, we have looked to test specifications and their associated task types, to provide a useful account of which task types measure which of the test components, or specifications.

### **3.7 Conclusion**

This chapter has described how, across the various stages of test design, a test purpose and domain may be identified, and how, aligned with that, a construct may be articulated. It has also discussed how, in relation to various principles of test design, the construct may be operationalised by defining its components, and then devising task types and ultimately subtests, with specified item types, that in their turn relate to and are aligned with these components.

At this stage it is also important to mention that the construct, a principal point for the design of a test, is not fixed and should regularly be scrutinised to improve test reliability and consistency. However, at the time of the development of the tests used for this study, the construct, components and task types discussed above were

widely accepted (ICELDA 2014), being employed not only within the four partnering institutions of ICELDA, but also at a pair of South African universities of technology, in Namibia, Singapore and Vietnam. Any new developments, however, are unlikely to affect the test results in any material way, since they are already covered at least to some extent in current versions of the test (Patterson & Weideman 2013b:146). The current construct can also be theoretically justified in terms of the range of experimentation trials and their constructive results, details of which are to be found in the various reports under the research tab on the ICELDA website (ICELDA 2014). The evolution of the construct as well as its components and subtests discussed here is an essential prelude to the discussion in the next chapter of the research method I followed to acquire the data that will be used for this study.

## **Chapter 4**

### **Research method**

#### **4.1 Introduction**

As indicated in the previous chapters, the tests used for this study were developed according to a specific construct and a specific definition of academic discourse (Patterson & Weideman 2013a:118). Their development included the selection of a similar set of components of academic literacy, as well as a roughly similar set of test specifications and subtests as those of a number of academic literacy tests that have been used in both undergraduate and postgraduate contexts over the last decade. In this chapter I will discuss the shortfalls of the current Home Language examinations and the necessity to utilise two additional measures of academic literacy. The discussion will include the detailed set of specifications that has informed their selection, and the importance of the process of test refinement. The rationale for further analyses of the empirical properties of the tests, employing a range of statistics packages, will be discussed. Next, the argument for making the results of grade 10 learners' performance in Home Language the basis for the comparison with the results of the two academic literacy tests will be set out, as will making the learners' average mark an overall indication of academic performance. Finally, the selection of a regression analysis for carrying out the comparison will also be motivated, and claims regarding the possible results of the analyses will be presented.

#### **4.2. Home Language assessment processes**

Currently, the end of the year examinations for students in the final phase of their school career consists of 3 test papers and an oral component, which is measured throughout the year and constitutes 50 marks of the 300 mark total. The first paper is called Language in context and consists of three sections, namely a

comprehension section, a summary, and the assessment of language structures and conventions such as vocabulary use and sentence structures. The second paper focuses on literature, which includes assessing students on poetry, a drama and a novel, whilst the last paper is regarded as a writing paper. For the last paper students have to write an essay and two transactional texts such as a formal or informal letter, reviews or minutes of a meeting, amongst various options (Department of Basic Education 2011:81-82). This is the format for all final phase examinations. The only difference between the papers are the lengths of the texts used or required. Grade 12 texts are generally thus longer than grade 10 and grade 11 texts. Below, an articulation of the different test papers can be seen in Table 4.1 (Department of Basic Education 2011:81-82), as well as the marks allocated to each component.

Consequently, when I refer to the utilisation of students' Home Language mark, I refer to the combination of the marks received for the above mentioned three test papers. The three separate test papers are therefore regarded as one combined mark for language ability and through this combination of what has been assessed, it is sometimes assumed that students are theoretically considered to have acquired the "language skills required for academic learning across the (school) curriculum" (Department of Basic Education 2011:9). For this reason I will then also regard the students' combined language mark as an indication of their potential to handle academic discourse. However, as mentioned on previous occasions, results obtained from Home Language examinations might be inadequate measures of academic literacy levels. Reasons include low Flesch-Kincaid grade levels of the text comprehension sections (Du Plessis 2014a:7), bulleted summaries have been required instead of written summaries in paragraph form since 2010 (Du Plessis 2014a:9), and the inappropriate use of cartoons and other visual texts as the sole indicators of the current generation's technological background in the language in context section (Du Plessis 2014a:10). In section 4.5 below the comparisons of the different measures of academic literacy levels, including Home Language results, with the students' overall average mark will be discussed in more detail. Now,

however, I will turn my attention to the specific nature of the remaining two measures of academic literacy levels, as well as the rationale for utilising them.

<b>Paper</b>	<b>Test items</b>	<b>Marks</b>
1. Language in context (70)	Comprehension (visual or graphic text): Identify and explain font types and sizes, captions, headings	30
	Summary (may not be the same text as the comprehension text)	10
	Language structures and conventions: Vocabulary Sentence structures Critical language awareness	30
2. Literature (80)	Poetry: Contextual/essay questions	30
	Novel: Contextual or essay question	25
	Drama: Contextual or essay question	25
3. Writing (100)	Essay: Narrative/descriptive/argumentative/ reflective/discursive	50
	Transactional texts: Letters (friendly, formal, informal, press)/curriculum vitae and cover letter/obituary/agenda and minutes of a meeting/report (formal or informal)/review/newspaper article/magazine article/written speech (formal or informal)/dialogue/written interview	50 (25 X 2)
4. Oral (50)	Prepared speech Unprepared speech Listening for comprehension	50 (20 + 15 +15)

**Table 4.1: A summary of the test papers for final phase examinations**

### **4.3 The specific nature of the additional tests**

For this study, I will make use of two further tests. The first is called the Test of Advanced Language Ability (TALA) and was designed by a panel of expert test designers and teachers brought together by the Inter-Institutional Centre for Language Development and Assessment (ICELDA), in a project commissioned by Umalusi (Du Plessis, Steyn & Weideman 2014). This test has been piloted, and refinements suggested by a second panel of experts have been made. The process of the development and refinement of TALA will be described in another study currently being completed by Steyn (2015).

Originally, the test to be used in this study (TALA) had 187 items, which were piloted on 1244 students of the Bloemfontein area. For that specific pilot, the test obtained an impressive reliability score of 0.985. TALA was then refined and reduced to have only 60 items. This refined version of TALA obtained a reliability score of 0.900 (Weideman, Du Plessis & Steyn 2014:14). The second test was taken from a test book by Weideman and Van Dyk (2014), which was then reduced to a 60 item test on the basis of the test specifications of TALA (see Table 4.2 and the discussion in section 4.3.2 below). The second test could also be a candidate for refinement. Both these tests were also developed according to the construct and components mentioned in previous chapters (Chapter 2 and Chapter 3). However, while they demonstrate a similar interpretation of the mentioned construct in comparison with other tests, they make use of a more limited set of subtests, mainly because they are aimed at a different target group, but also for logistical ease. Whereas TALPS (Test of Academic Literacy for Postgraduate Students), for example, is intended to measure academic literacy levels of students who would like to further their postgraduate studies, the tests used for this study are specifically aimed at measuring the academic literacy ability of school students. They are therefore based, among other things, on some of the prerequisites as set out in the Curriculum and Assessment Policy Statement (CAPS) for Home Language students (Department of Basic Education 2011). The tests thus have fewer subtests than



others that derive from the same construct. For example, the Text type task, Verbal reasoning task and Academic Writing task were not included. In the case of the former two (Text type and Verbal reasoning), they were known from previous experiments to have lower reliability levels from previous experiments. Concerning Academic writing, the logistical difficulty of reliably marking such a task in this experiment presented a practical hurdle for its inclusion. Thus, since the tests have fewer subtests, they therefore require a shorter writing time.

The tests, but especially the test that will be piloted first (TALA), were developed according to what are generic ideas of language ability referred to in CAPS (Weideman, Du Plessis & Steyn 2014:2; Department of Basic Education 2011:9). In principle, even though the tests are aimed at Home Language students who need to demonstrate an ability to function within a range of material lingual spheres, they measure general components or abilities which can be attributed to most so called 'high-level' generic functions across the various discourse types (Weideman, Du Plessis & Steyn 2014:12) that the curriculum refers to. The ability to distinguish between essential and non-essential information, for example, while characteristic of one sphere (academic discourse) can be attributed to and occur in a range of other material lingual spheres as well. Steyn (2015:1) points out that by focussing on generic or general abilities which can be seen as functions in various material lingual spheres, we can in some way develop an assessment instrument that cuts across discourse types. At the same time, the tests should be a better measure of the 'high-level' language ability that CAPS frequently refers to; and certainly better than the sometimes less challenging questions that characterise some of the Home Language papers (Du Plessis 2014a:5).

There are other differences between the two tests as well. Of the two tests used for this study, only TALA (The Test of Advanced Language Ability) has undergone post-piloting refinement. This means that the test has been subjected to a test-level and item analysis after piloting results were obtained, in order to determine the

discrimination values of each item as well as some other parameters of item performance. The IteMan program versions 3.6 and 4.2 have been employed. Both IteMan programs have been used, since the 3.6 version generates different calculations of, for example, discrimination values, than the 4.2 version, while 4.2 in turn allows one to consider more conservative measures of item discrimination, and also calculates Differential Item Functioning (DIF) in order to check whether items unfairly discriminate against certain groups in the test population. Another analysis, generated by the TiaPlus program, allows not only the calculation of the point-biserial (rpbis) of test items, which refers to an item's ability to distinguish between students who have a high language ability and students with a low language ability (Steyn 2015:10), but also consider an orthodox measure of validity, namely the intercorrelations among the various subtests, as well as the correlation of each subtest with the overall score. Additional conditions to which productive items should adhere include the item's alignment with the construct of a test, its facility value and, at test level, a factor analysis to determine the homogeneity of a test (Weideman 2011:105).

These analyses are done to improve, through empirical analysis, the quality of a given test (Van der Slik & Weideman 2005:23). A productive item, for example, would not be too easy to answer correctly (anything between 20% and 80% of test takers would be an acceptable facility value), and would be able to distinguish between the top 25% of the test group and the bottom 25% in at least 30% of cases. It would also align with the construct of the test and different items and subtests would have a functional correlation with each other (Weideman 2011:105). It is for this reason that TALA will be administered first, as theoretically it should therefore provide more accurate data than the second test. The second test, as yet unrefined, should not be undervalued, however, since it was modified according to the test specifications of TALA, which will be discussed in more detail below (section 4.3.1). Test specifications prescribe the mark allocation of the various test sections or subtests, the essential components of the construct that should be measured, and

the types of primary questions which would be the most advantageous to utilise (Steyn 2015:23). In this sense, the two tests are closely related, even though the one has undergone refinement and the other not. I will now discuss in more detail the specific nature and purpose of each test.

#### **4.3.1 TALA**

The Test of Advanced Language Ability was developed with the aim to assess in a more reliable manner the high-level language ability of high school students in their final phase of schooling. As mentioned previously in section 1.2, the current assessment of Home Languages for the exit-level examinations cannot be seen as an equivalent to TALA because of the misalignment between the school curriculum and the exit-level examinations (Steyn 2015:13), as is clear from the data and analyses presented in several studies commissioned by Umalusi in the recent past (Steyn 2015; Weideman, Du Plessis & Steyn 2014; Du Plessis, Steyn & Weideman 2014). One can therefore argue that there is a need for utilising measurement instruments, which still adhere to assessment requirements as set out by CAPS, but assess language ability more reliably than the current set of grade 12 examination papers. Since TALA was developed with the intention of assessing high school students within the South African education system, teaching and testing guidelines presented in the CAPS outline were used as directives for its design (Weideman, Du Plessis & Steyn 2014:1).

In particular, CAPS refers to the notion of both a differentiated language ability and a generic language ability (Weideman, Du Plessis & Steyn 2014:12). A differentiated language ability enables a student to function within specific contexts, discourse types or material lingual spheres. CAPS (Department of Basic Education 2011) distinguishes between six different discourse types, namely social, economic, academic, aesthetic, ethical and political (Weideman, Du Plessis & Steyn 2014:10), and communication within some of these contexts can be referred

to as participation in high level language functions. One must acknowledge, however, that certain language skills are utilised within all of the mentioned discourse types, however differentiated they might be. This common ground between the various discourse types is what was previously referred to as generic language ability and includes components such as “comparing and contrasting, classifying and inferring, identifying purpose, creating coherence, defining and explaining” (Steyn 2015:14).

To therefore assess skills that demonstrate similarities to components and functions of academic literacy, but are also characteristic of generic language ability, explicit test specification guidelines must be articulated. Test specifications are of significance, since mark allocations and question guidelines are articulated which guide the design of similar tests. It is for this reason that TALA’s test specifications, which will be described in greater detail in the following section, could form the basis for the modification of the second test. The administration of the second test could, in turn, further substantiate the credibility of TALA as a potentially reliable and valid measure of advanced language ability.

#### **4.3.2 The second test**

The second test was taken from an academic literacy workbook which was edited by Weideman and Van Dyk (2014), and includes the expertise of various other academic literacy scholars. The introduction to this book clearly states that it is aimed at preparing learners in the senior phase of secondary school to handle academic discourse. A separate study done by Erasmus (2014) showed that this specific test that has been used, possessed texts that were appropriate for use with Grade 10 learners. The workbook touches on matters such as the significance of academic literacy, as well as the significance of preparing high school students for possibly writing academic literacy tests when they apply to tertiary education institutions. The workbook contains six tests and their memoranda, which are designed for high school students. The test that I chose initially had a score of a 100

marks, but it was then modified to have a subtest framework similar to that of TALA. This was done by utilising the test specifications of TALA. The table below records these test specifications (Steyn 2015:37):

<b>Subtest and general task type</b>	<b>Component measured/potentially measured</b>	<b>Specifications for items (60 marks): guidelines for questions</b>
<p>A <b>Scrambled text</b> in which the candidate is given an altered sequence of sentences and must determine the correct order in which these sentences must be placed.</p>	<p>Textuality: cohesion and grammar, 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 See sequence and order Understanding text type (genre) Communicative function Making meaning beyond the sentence</p>	<p>(5)</p> <ul style="list-style-type: none"> <li>✓ Sequencing</li> </ul> <p>[Candidates use their knowledge between different parts of the text and the logical development of an academic text to determine the correct order.]</p>
<p><b>Vocabulary knowledge</b> is tested in the form of multiple choice questions</p>	<p>Vocabulary comprehension: understand and use a range of academic vocabulary as well as content or discipline-specific vocabulary in context (however, limited to a single sentence).</p>	<p>(10)</p> <ul style="list-style-type: none"> <li>✓ Vocabulary in context (use)</li> <li>✓ Handling metaphor and idiom (optional)</li> </ul>
<p>The <b>Interpreting graphs and visual information</b> subtest consists of questions on graphs and simple numerical computations.</p>	<p>Understanding text type (genre) Understanding graphic and visual information Distinguish between essential and non-essential information, fact and opinion, propositions and arguments, cause and effect, and classify, categorise and</p>	<p>(8)</p> <ul style="list-style-type: none"> <li>✓ Trends: Perceived trends in sequence, proportion and size. Predictions and estimations based on trends. Averages across categories, etc.</li> <li>✓ Proportions:</li> </ul>

	<p>handle data that make comparisons</p> <p>Numerical computation</p> <p>Extrapolation and application</p> <p>Making meaning beyond the sentence</p>	<p>Identify proportions expressed in terms of fractions or percentages.</p> <p>Compare proportions expressed in terms of fractions or percentages, e.g. biggest difference or smallest difference.</p> <ul style="list-style-type: none"> <li>✓ Comparisons between individual readings within a category in terms of fraction, percentage or the reading in the relevant unit (e.g. in grams or millions or tonnes)</li> <li>✓ Comparisons between the combined readings of two or more categories in terms of fractions, percentage or the reading in the relevant unit</li> <li>✓ Differences between categories</li> <li>✓ Comparisons of categories</li> <li>✓ Inferencing/extrapolation based on the given graphic information.</li> </ul>
<p>In the <b>Text comprehension</b> section, candidates must answer questions about the given text.</p>	<p>Vocabulary comprehension</p> <p>Understanding metaphor and idiom and vocabulary in use</p> <p>Distinguish between essential and non-essential information, fact and opinion , propositions and arguments, cause and effect, and classify, categorise and handle data that make comparisons</p> <p>Extrapolation and application</p> <p>Think critically (analyse the use of techniques and arguments) and reason logically and systematically</p>	<p>(25)</p> <p>Essential</p> <ul style="list-style-type: none"> <li>✓ Distinction making: categorisation, comparison, distinguish between essential and non-essential (5)</li> <li>✓ Inferencing/extrapolation: e.g. identify cause and effect (3)</li> <li>✓ Comparing text with text (2)</li> <li>✓ Vocabulary in context (5)</li> <li>✓ Handling metaphor, idiom and word play (1)</li> </ul> <p>Another (4) from any of these.</p> <p>Possible</p>

	<p>Interact with texts: discuss, question, agree/disagree, evaluate, research and investigate problems, analyse, link texts, draw logical conclusions from texts, and then produce new texts</p> <p>Synthesise and integrate information from a multiplicity of sources with one's own knowledge in order to build new assertions</p> <p>Communicative function Making meaning beyond the sentence</p> <p>Textuality (cohesion and grammar)</p> <p>Understanding text type (genre)</p>	<p>(5) of the following:</p> <ul style="list-style-type: none"> <li>✓ Communicative function: e.g. defining/concluding</li> <li>✓ Cohesion/cohesive ties</li> <li>✓ Sequencing/text organisation and structure</li> <li>✓ Calculation</li> </ul>
<p>In the <b>Grammar and text relations</b> section the questions require the candidate to determine where words may have been deleted and which words belong in certain places in a given text that has been systematically mutilated</p>	<p>Vocabulary comprehension</p> <p>Textuality (cohesion and grammar)</p> <p>Understanding text type (genre)</p> <p>Communicative function</p>	<p>(12)</p> <p>Determined by the specific item. The text is systematically mutilated – one cannot predict beforehand which components will be measured, but a good range is possible and indicated.</p>

**Table 4.2: TALA test specifications**

From Table 4.2 one can identify the five subtests as Scrambled text, Vocabulary knowledge, Understanding graphs and visual information, Text comprehension and Grammar and text relations. The subtests each measure more than one of the components pertaining to academic literacy. What is more, each one of the identified components of academic literacy may potentially be measured by more than one subtest of the same test. Textuality, for example, can be measured by means of a subtest such as Scrambled text, Text comprehension or Grammar and text relations, or all of them.

With the assistance of an experienced high school teacher, the original 100 mark copy of the second test was modified to a 60 mark test. The Scrambled text was kept exactly the same, since the original also constituted five marks. The remaining subtests were all modified in light of the specifications listed above. For Vocabulary knowledge, ten questions of the total 25 questions were kept. The ten questions were chosen with the assistance of the teacher mentioned, and questions which were more likely to be misinterpreted by students were discarded. An example of a question that has been removed from the original test is the following:

**It is commonly believed that the popularity of gadgets and freaky inventions \_\_\_\_\_ to the growing influence of sophisticated technology on our lives.**

- A. contravenes
- B. contributes
- C. contrives
- D. contradicts

This question can be ambiguous since some people may believe that any gadget or invention is beneficial to technological progress whilst other may believe that such inventions are harmful to the sophisticated advances of technological progress. For this reason, the question was removed. The Verbal reasoning subtest was eliminated altogether. Interpreting graphs and visual information was modified by means of keeping three questions which are aimed at the identification of trends in averages, sequence or estimations. An example of such a question is:

**The only faculty that followed the same pattern as the Faculty of Medicine is**

- A. Public Health.
- B. Dentistry.
- C. Physical Science.
- D. Life Science.

The remaining five questions were aimed at proportions identified in terms of fractions or percentages, as seen in this third extract from the test:



**In which year was the Faculty of Life Science’s % of reports double the % of reports of the Faculty of Dentistry?**

- A. 2007
- B. 2008
- C. 2009
- D. 2010

Consequently, only two questions of the original ten were eliminated. The Register and text type subtest was also eliminated entirely. Text comprehension had to be modified to constitute 25 marks, instead of 35 marks. Questions which could be misinterpreted by students were again eliminated first. As seen in the table above (Table 4.2), of the 25 questions, five questions should relate to distinction-making as in the example below:

**SixthSense has a number of functions that can make day-to-day life easier. Which of the following is the odd one out?**

- A. Checking emails anytime, anywhere.
- B. Capturing memories without a camera.
- C. Helping you understand why your flight is late.
- D. Turning the entire world into a computer.

Another three questions should be aimed at extrapolation and inferencing, as seen in this further example:

**One can conclude from paragraph \_\_\_\_\_ that SixthSense is a project that the Fluid Interfaces Group has been working on for quite some time.**

- A. 5
- B. 4
- C. 3
- D. 2

A further two questions should focus on text to text comparison as illustrated by both the previous and the next example:

**The phrase ‘gesture driven’ in paragraph two is best supported and explained by which other paragraph?**

- A. 1
- B. 2
- C. 3
- D. 4

Another five questions should emphasise vocabulary in context. An example of such a question can be seen in the following extract:

**The word ‘augment’ in paragraph 2 means to \_\_\_\_\_ of the physical world.**

- A. adapt to another form
- B. change the perception
- C. change the reality
- D. increase the quality

One question should be aimed at understanding metaphor and idiom, as illustrated in this example:

**The phrase ‘turning the entire world into a computer’ in paragraph 1 means that**

- A. the earth as a whole works as a computer.
- B. everything can form part of this computer.
- C. the earth is controlled by a computer.
- D. the world is nothing but a computer.

Another four questions from the categories mentioned in Table 4.2 can be chosen as well, whilst the remaining five marks should relate to either communicative function, the identification of cohesion or sequencing, or any type of calculation. Lastly, for Grammar and text relations some of the original questions were kept as examples, whilst the remaining questions were kept as they were.

In addition to the modification of test items and subtests, the texts used for the second test were also analysed to ensure that they were appropriate for grade 10 students. The Flesch reading ease of a text for grade 10 students should preferably be above 50% and should fall within a grade 10 level. This would indicate that the text is neither too difficult nor too easy to read (Steyn 2010:5). The first text within the test has a Flesch reading ease of 56.3% and a Flesch-Kincaid level of 10.5, whilst the second text has a Flesch reading ease of 67% and a Flesch-Kincaid level of 8.6 (Steyn 2010:5). The test is therefore likely to be an appropriately modified measure of academic literacy levels and can be viewed as an annexure (A) of this study. For reasons of confidentiality, however, to view TALA one should contact the author of the study. In the next section I will discuss in further detail the target

group, the correlations which will be of importance, and other relevant matters of significance.

#### **4.4 The target group**

The tests that will be used as measurement instruments, and discussed above, are not the only central concern. The groups on which they will be administered (in the case of TALA) and piloted (in the case of the second test) as well as the analysis of the results, are also of significance. The choice of groups will therefore firstly be motivated, after which I will discuss the relevance of the data. The tests for this study will be administered to grade 10 students of two Bloemfontein based schools.

I have chosen grade 10 students for three reasons, the first being that their schedule allows more time for additional activities in comparison to the usually busier schedules of grade 11 or grade 12 students, thus making Grade 10 students a more convenient choice. The second reason pertains to the possible reliability of the students' school results, specifically when they have been combined with continuous assessment marks (as is the case for grade 12 marks for Home Language). In various reports issued by Umalusi, the reliability of assessments concerning Home Language have been questioned in respect of both the standard of the assessments, and their quality and consistency (Weideman, Du Plessis & Steyn 2014:2). The current continuous assessment system which is utilised in South African schools, for example, has recently been criticised in the press for being an unreliable and dishonest depiction of students' language ability (Prins 2014:11). The last reason refers to the position of grade 10 students, since they are in the final phase of high school during which students are urged to prepare for their career choices and tertiary studies. It is from this phase that learners at school progress to studies in higher education, the context in which their language ability becomes critical in light of the considerations of this study. Additionally, if we are able to identify literacy problems during this phase of students' high school years, more time is then available to attempt to remediate these problems.

The first school, which was approached for this study, had 160 Grade 10 students. The second school only had 78 Grade 10 students. The ages ranged anywhere between 15 years to 17 years of age, and whilst the first school is an all-girls school, the second school had more or less the same number of students of either gender. The two test groups are also socio-economically divergent. The first group is made up of learners at a former Model C school which is noted for its excellent academic record. The second is a township school which has learners who come from poorer backgrounds whilst the school itself is probably not as well equipped as the first.

#### **4.5 Procedure**

After the tests have been piloted, with permission from both principals of the respective schools, three sets of comparisons will form the focus of the study. The results obtained by the students for the two academic literacy tests will firstly be compared to their Home Language result. Thereafter the results will also be compared to their average mark (sometimes referred to as their GPA or grade point average, as in America) and lastly, the results will again be compared to their average mark excluding their Home Language mark. The comparisons noted above could indicate not only whether a specific academic literacy test could more accurately measure academic literacy levels than school language results, but could possibly also serve as evidence of the assumed relationship between academic literacy levels and overall academic success. The comparisons will be explored by means of a regression analysis which will be discussed and motivated in further detail in section 4.5.

As I have already noted above in section 4.3, additional analyses that will be carried out include Iteman 3.6 and Iteman 4.3 analyses, as well as a TiaPlus analysis. The Iteman programs generate statistical analyses regarding item performance, and whereas Iteman 4.3 generates individual graphs per item, Iteman 3.6 presents only a table summarising the combined performance of the items. TiaPlus, on the other hand, measures intercorrelations between subtests, determines Differential item

functioning (DIF) and indicates the various reliability indices for the respective tests (these are all discussed in more detail in Chapter 5).

#### **4.6 The regression analysis and choice of variables**

The type of analysis that will be utilised for the comparison between the different measures is called a regression analysis, which is a widely utilised statistical measure used to identify patterns among sets of data. Regression analysis is generally used to explore the useful relationships which may exist amongst variables (Chatterjee & Price 1991:1). For a regression analysis two types of variables are needed, a response variable and one or more predictor variables (Chatterjee & Price 1991:1). Since this study is aimed at determining which measure is the better predictor of academic literacy levels or academic success, the students' average mark will be regarded as the response variable, as it is the variable to which the other variables will be compared, whilst the students' Home Language marks and academic literacy test marks will be seen as the predictor variables of the response variable. Through what is often deemed as an informal type of data analysis, a possible relationship can be calculated between the response variable and the predictor variables, or in other words, between the students' average mark and the three possible measurements of academic literacy levels. Since there is more than one predictor variable, the equation that will be utilised is regarded as a multiple regression equation (Chatterjee & Price 1991:1).

Since a regression analysis is used to identify useful relationships amongst sets of variables, hypotheses about the possible relationships can be formulated in advance, assuming that the hypotheses will turn out to be of a valid nature. For this study, however, I shall refer to these hypotheses as claims. In the following section of this chapter, which is also the concluding section, these claims will be presented.

#### **4.7 The claims**

After the analysis has been carried out, there are five claims I expect to be able to substantiate with the results obtained. I will now describe and motivate these claims, and discuss their significance to the study.

**Claim 1: The refined test will demonstrate a better correlation with students' average mark, and therefore their possible academic success, than the unrefined test.**

Since the first test (TALA) has already been piloted on school students and undergone refinement in that regard, it should be better able to indicate academic literacy levels than the second, unrefined test. The second test should, however, still demonstrate a high measure of correlation even before items have undergone refinement after analyses have been carried out. The claim therefore assumes that tests of academic literacy are related to students' overall performance across school subjects, since the language ability to handle the demands of these other subjects must have some relation to performance in them.

**Claim 2: Both tests will demonstrate a better correlation with students' average mark than their Home Language mark would demonstrate.**

Since the academic literacy tests are aimed specifically at measuring academic literacy levels, and since academic literacy is assumed to relate to language competence across all academic subjects taken at school, their results should be more representative of possible academic success than the results of their Home Language mark, which assesses a single subject. This is indicated, amongst other things, by the fact that Home Language results include the measurement of students' performance in various material lingual spheres, and neither exclusively their ability only in academic discourse, nor in advanced level generic language functions that to an extent approximate academic literacy levels. What is more, if Home Language examination results are as unreliable as Umalusi believes, and has

been reported in the press (Joubert 2014:2), both of these tests of academic literacy will be more reliable indicators of language ability than the Home Language mark.

**Claim 3: Both tests will demonstrate an even better correlation with students' average mark, excluding the Home Language mark.**

With the exclusion of the Home Language mark from the students' average mark, an even better correlation can be expected between the two tests of academic literacy and overall performance, as measured by their average mark.

**Claim 4: The results of the tests will enable me to determine the relative power of the unrefined test and will also enable me to make recommendations for its refinement and that of similar tests.**

The possible variance in results between the refined test and the unrefined test could indicate the extent to which the unrefined test should be modified as to obtain a higher level of reliability. It might also indicate whether the development of similar tests is appropriate and feasible.

**Claim 5: More useful conclusions regarding academic literacy levels can be drawn from academic literacy tests than can be drawn from other types of measurements.**

Given the typicality of academic discourse (Patterson and Weideman 2013a:118) it is highly inappropriate to utilise results which pertain to students' ability to engage with literary texts, public speaking, and so forth, as evidence of a student's ability to engage with and handle academic discourse. In this sense, one is comparing different types of abilities which do not necessarily share any similarities. By rather making use of academic literacy tests to determine academic literacy levels, one can be more confident of a reliable outcome.

## **4.8 Conclusion**

In the next chapter the mentioned analyses will be presented. The claims mentioned above (section 4.6) will then be discussed in more detail, as will other observations that could also be drawn from the Iteman 3.6, Iteman 4.3 and TiaPlus analyses.



## **Chapter 5**

### **Analyses and interpretation**

#### **5.1 Introduction**

This chapter discusses the various test and item-level analyses that have been conducted on both tests that were used for this study: the Test of Advanced Language Ability (TALA) and the second test, a TALA-like derivative test, which were used for this study. These analyses include both an IteMan 3.6 and IteMan 4.3 analysis, as well as a TiaPlus analysis. Additionally, a differential item functioning (DIF) analysis will also be undertaken in order to determine whether some of the test items possibly discriminated against one of the test groups. The two DIF analyses may also provide some basis of comparison between the two tests. Thereafter, the emphasis will shift to a consideration of the results obtained from the regression analysis regarding the comparisons which were anticipated in previous chapters. These comparisons include the link between the students' Home Language mark and their score on both academic literacy tests, the relation between the student's score on both academic literacy tests and their average school mark, and lastly the connection between the students' academic literacy test scores and their school average, excluding their Home Language mark. Finally, the information provided will be utilised to substantiate the accuracy, or lack thereof, of the claims articulated in the previous chapter.

#### **5.2 IteMan 3.6 analysis**

The IteMan 3.6 program generates a statistical analysis of test and item performance (Assessment Systems Corporation 2006). Especially where larger groups are concerned, this type of analysis assists with possible generalisations that can be made regarding test groups (Du Plessis 2012:74). Score distributions and other properties of test performance are represented in the form of graphs and tables, and

scale statistics are provided at the end of the generated report. The complete reports for both TALA and the second academic literacy test can be viewed respectively as annexure B and C, and the most relevant figures will also be discussed in some detail below.

### 5.2.1 TALA

From the report generated by Iteman 3.6, some key statistics of TALA’s re-pilot for this study are important to take note of. The following scale statistics have been generated:

Scale:	1
	-----
N of Items	60
N of Examinees	242
Mean	25.112
Variance	65.678
Std. Dev.	8.104
Skew	0.278
Kurtosis	-0.413
Minimum	6.000
Maximum	50.000
Median	24.000
<b>Alpha</b>	<b>0.818</b>
SEM	3.458
<b>Mean Pcnt Corr</b>	<b>42</b>
Mean Item-Tot.	0.238
<b>Mean Biserial</b>	<b>0.313</b>

**Table 5.1: Scale statistics, TALA**

The significant figures for the purpose of this study have been rendered in bold in the table above (Table 5.1). The first significant statistic to note is that TALA obtained a Cronbach’s Alpha score of 0.818 for this specific administration. Cronbach’s Alpha is a reliability index which measures the technical consistency of a test, and a score of at least 0.7 is required (Weideman 2011:105) for tests whose results affect or influence decisions that will have a significant impact on the subsequent lives of those who wrote the test, for example, tests that are high to medium stakes tests. TALA, for example, can be regarded either as a medium stakes

test or as a high stakes test. TALA can be seen as a medium stakes test when it is used to identify students who would benefit from participating in courses which could develop their competence to use academic language adequately. On the other hand, TALA can also be regarded as a high stakes test in situations where students may be exposed to stigmatisation because they performed poorly on the test. An Alpha of 0.818 can be therefore be regarded as satisfactory, since it indicates that the results generated by the test can be regarded as adequately reliable in terms of the 0.7 benchmark.

The next relevant statistic is the mean biserial score. The mean biserial score indicates the general ability of items in the test to discriminate between students with a high language ability and students with a lower language ability (Steyn 2015:11). As Steyn (2015:11) and Du Plessis (2012:18) indicate, the benchmark chosen for tests by ICELDA is that an overall score above 0.15 is indicative of a suitable mean biserial score. As can be seen from the given statistics (Table 5.1), TALA scored 0.313 for its mean biserial, which is comfortably above the required minimum.

The last notable statistic is that of mean percentage correct. This statistic shows the facility value of the test, and ideally should be in the vicinity of 50% (Weideman 2011:105). During this pilot, TALA scored an average facility value of 42%, which means that it was slightly more difficult than the desirable average of 50%. The further implication is that a test that is too difficult may affect the reliability of its measurement, and may also affect its ability to help predict associated performances, such as overall academic performance that depends to a certain extent on language ability. It is an early indication, therefore, that the test may be marginally inappropriate for the current analysis.

### 5.2.2 The second test

An Iteman 3.6 analysis was also done on the second, TALA-like test. The scale statistics of this test can be seen below:

Scale:	1
-----	
N of Items	60
N of Examinees	240
Mean	33.233
Variance	110.962
Std. Dev.	10.534
Skew	-0.027
Kurtosis	-0.909
Minimum	9.000
Maximum	54.000
Median	33.000
<b>Alpha</b>	<b>0.896</b>
SEM	3.396
<b>Mean Pcnt Corr</b>	<b>55</b>
Mean Item-Tot.	0.333
<b>Mean Biserial</b>	<b>0.432</b>

**Table 5.2: Scale statistics, second test**

Again one can look at the same statistics of significance as referred to in Table 5.1. A satisfactory Cronbach's Alpha of 0.896 was achieved, which is comfortably above the required 0.7. Additionally, a mean biserial score of 0.432 can be noted, as well as an overall facility value of 55%. The latter figure shows that the test was perhaps slightly easier than the expectable 50%. All of these, however, fall more or less within the desired parameters. This is an early indication that this particular test may have greater potential than TALA to predict associated academic performances. In the following section attention will be given to the results obtained from an analysis done with another, more recent version of Iteman.

### 5.3 Iteman 4.3 analysis

Iteman 4.3 (Guyer & Thompson 2011) is a more recent version of the program than Iteman 3.6, which implies that it should not only yield more accurate analyses, but also additional information. Summary versions of the Iteman 4.3 reports are attached as annexures D and E. It is noticeable that these reports contain more

extensive information of test items and subtests, as well as the addition of graphs that assist in visualising this information.

### 5.3.1 TALA

In Table 5.3 below, each subtest's Alpha is provided, as well as the overall Alpha score which was obtained by TALA. A very slight improvement can be seen regarding the Alpha score. Iteman 3.6 indicates an Alpha of 0.818, whilst Iteman 4.3 indicates an Alpha score of 0.819. Since Iteman 4.3 is a more refined version of the program than Iteman 3.6, the Alpha of 0.819 should be regarded as a more correct representation of the test's overall technical consistency. One would also notice that the Alpha scores of each subtest in TALA do not fall within the desired parameters. However, a subtest with a more satisfactory Alpha, such as Text comprehension (marked in bold below), increases the overall Alpha score of the test, even though other subtests, such as Interpreting graphs and visual information, might not have performed as well. An overall satisfactory Alpha score was nonetheless still attained, since the total test consistency is more than the average of that of the subtests.

Score	Alpha	SEM	Split-Half (Random)	Split-Half (First-Last)	Split-Half (Odd-Even)	S-B Random	S-B First-Last	S-B Odd-Even
Scored items	<b>0.819</b>	3.457	0.729	0.601	0.684	0.843	0.751	0.813
Scrambled text	0.598	0.913	0.151	0.221	0.554	0.263	0.362	0.713
Vocabulary knowledge	0.547	1.381	0.364	0.382	0.367	0.534	0.553	0.537
<b>Interpreting graphs &amp; visual information</b>	<b>0.459</b>	1.241	0.339	0.320	0.264	0.506	0.485	0.417
<b>Text comprehension</b>	<b>0.722</b>	2.204	0.586	0.554	0.564	0.739	0.713	0.721
Grammar & text relations	0.628	1.501	0.381	0.301	0.633	0.552	0.462	0.775

**Table 5.3: Iteman 4.3 reliability report of TALA's re-pilot**

In the following table (5.4) below, the mean Rpbis of each subtest, as well as the overall Rpbis score of the test can be seen. Rpbis, the point biserial index, is an accompanying indication of how well a test item discriminates between test takers who selected correct answers and test takers who selected incorrect answers. Rpbis can range anywhere between -1.0 and 1.0 (Du Plessis 2012:81). An item with a negative Rpbis score indicates that students with a higher language ability selected

an incorrect answer where students with a lower language ability chose the correct answer, and is of course quite undesirable. It is undesirable since one would like a test to distinguish positively among test takers with a high language ability and those with a lower language ability (Guyer & Thompson 2011:30). On the other hand, a positive Rpbis score would indicate that test takers with a higher language ability mostly chose the correct answers whilst test takers with a lower language ability chose the incorrect answers. A positive Rpbis score is, of course, a desirable feature of any item in a test. Iteman 4.3 indicates an overall mean Rpbis score of 0.239, which is slightly below Iteman 3.6's indication of an overall biserial score of 0.313. Both values, however, score within the stipulated parameter of above 0.15.

Score	Items	Mean	SD	Min Score	Max Score	Mean P	Mean Rpbis
All items	60	25.112	8.121	6	50	0.419	<b>0.239</b>
Scored Items	60	25.112	8.121	6	50	0.419	0.239
Scrambled text	5	2.541	1.441	0	5	0.508	0.131
Vocabulary knowledge	10	5.306	2.051	0	10	0.531	<b>0.272</b>
Interpreting graphs & visual information	8	2.764	1.687	0	8	0.346	0.190
Text comprehension	25	9.950	4.179	2	22	0.398	<b>0.266</b>
Grammar & text relations	12	4.550	2.461	0	12	0.379	0.231

**Table 5.4: Iteman 4.3 summary statistics of TALA's re-pilot**

Again, the longer subtest, Text comprehension (with 25 items) makes a large contribution to the mean Rpbis score, with the shortest subtest (Scrambled text, with 5 items) making the smallest contribution. For its length, Vocabulary knowledge (10 items), with a mean Rpbis of 0.272, makes a disproportionately large contribution to the overall discrimination value of the test. The reason for this may be explored in subsequent studies.

### 5.3.2 The second test

Shown below in Table 5.5 are the reliability statistics for the second test, as generated by Iteman 4.3. A slightly higher Alpha score of 0.897 is once again indicated, compared to the overall Alpha score of 0.896 which was indicated by Iteman 3.6.

Score	Alpha	SEM	Split-Half (Random)	Split-Half (First-Last)	Split-Half (Odd-Even)	S-B Random	S-B First-Last	S-B Odd-Even
Scored items	<b>0.897</b>	3.394	0.846	0.683	0.843	0.916	0.812	0.915
Scrambled text	0.865	0.734	0.660	0.773	0.769	0.795	0.872	0.869
Vocabulary knowledge	0.444	1.304	0.413	0.299	0.354	0.585	0.461	0.523
Interpreting graphs & visual information	0.801	1.087	0.616	0.659	0.671	0.763	0.794	0.803
Text comprehension	0.751	2.216	0.576	0.617	0.602	0.731	0.763	0.751
Grammar & text relations	0.707	1.538	0.463	0.357	0.779	0.633	0.526	0.876

**Table 5.5: Iteman 4.3 reliability report of the second test**

The summary statistics generated by Iteman 4.3 can be seen in the table below (Table 5.6). A mean Rpbis score of 0.334 is indicated by Iteman 4.3. Whilst lower than the mean biserial score of 0.432 indicated by Iteman 3.6, the more refined (but patently more conservative) overall Rpbis score generated by Iteman 4.3 still falls within the preferred parameters of a test's discrimination ability.

Score	Items	Mean	SD	Min Score	Max Score	Mean P	Mean Rpbis
All items	60	33.233	10.556	9	54	0.554	<b>0.334</b>
Scored Items	60	33.233	10.556	9	54	0.554	0.334
Scrambled text	5	2.513	1.998	0	5	0.502	0.483
Vocabulary knowledge	10	6.233	1.749	0	9	0.623	0.250
Interpreting graphs & visual information	8	5.004	2.438	0	8	0.626	0.493
Text comprehension	25	13.654	4.445	3	23	0.546	0.293
Grammar & text relations	12	5.829	2.839	0	12	0.486	0.319

**Table 5.6: Iteman 4.3 summary statistics of the second test**

Although a number of other statistics generated by the Iteman programs can also be considered when one looks at the given reports attached as annexures, for the purpose of this study, only a number of relevant statistics have been emphasised. I will now turn my attention the reports generated by another program, named TiaPlus.

#### **5.4 TiaPlus analysis**

TiaPlus (CITO 2005) is a statistics program which also measures, amongst other things, the construct validity of a test by estimating the intercorrelations between subtests (Du Plessis 2012:130; Van der Walt & Steyn 2007:147). As mentioned earlier in this study (section 2.7), subtest correlations are of importance, since these indicate the degree to which subtests measure the same or different components of the construct. Even though it may turn out that some subtests, if not all, measure more than one component at a time, correlations should not be too strong, since each subtest should preferably measure a different part of the language ability being assessed (Rambiritch 2012:186). At the same time, one would wish the correlations of the various subtests with the overall test score to be high, since all potentially contribute towards the overall measurement result. The desired parameters for subtest correlations fall between 0.3 and 0.5, whilst correlations between a subtest and the overall test should score 0.7 or even higher (Van der Walt & Steyn 2007:148). For TiaPlus, the P-value score indicates the average facility value of the test and is indicative of the proportion of test takers who selected the correct answer of a test item (Guyer & Thompson 2011:30). This score should, as has been observed above, be in the vicinity of 50%.

Another useful statistic generated by TiaPlus is Differential item functioning (DIF). DIF indicates whether an item is potentially biased against one of the test groups which participated in writing a specific test and can be noted when the same item performs differently for different test groups. Such an item will subsequently be flagged by TiaPlus (CITO 2007). As has been remarked above, this is a useful statistic to consider in this case, since the two test groups are socio-economically divergent, the one being a township school, and the other a former model C school. The DIF analysis should identify items that are biased for or against one of these groups.



Additionally, the statistics generated by TiaPlus also refer to the Alpha reliability index of a test, as well as to the Greatest Lower Bound (GLB) index of the consistency of a test. GLB is an additional internal reliability index, and is better suited for more heterogeneous tests. Heterogeneous tests are composed of test items or subtests which clearly measure more than one ability (CITO 2005:18) or an ability that is so complex that it is difficult to define and articulate in a single trait. TALA, as well as the second test, are both heterogeneous tests, since subtests measure a potentially wide range of components of a very rich construct (Weideman & Van der Slik 2008:168). This can be seen below in the following subtest intercorrelation analyses.

#### 5.4.1 TALA

The subtest intercorrelations of TALA are represented in the table below:

Subtest	Test	1	2	3	4	5
Scrambled text	1	0.30				
Vocabulary know	2	<b>0.74</b>	0.10			
Interpreting gr	3	0.54	0.05	<b>0.30</b>		
Text comprehens	4	<b>0.88</b>	0.15	0.56	<b>0.38</b>	
Grammar & text	5	<b>0.65</b>	0.04	<b>0.39</b>	0.17	<b>0.40</b>
Number of testees :	242	242	242	242	242	242
Number of items :	60	5	10	8	25	12
Average test score:	25.11	2.54	5.31	2.76	9.95	4.55
Standard deviation:	8.10	1.44	2.05	1.68	4.17	2.46
SEM :	3.46	0.91	1.38	1.24	2.20	1.50
Average P-value :	41.85	50.83	53.06	34.56	39.80	37.91
Coefficient Alpha :	<b>0.82</b>	0.60	0.54	0.46	0.72	0.63
GLB :	<b>0.94</b>	0.75	0.64	0.57	0.83	0.79
Asymptotic GLB :	0.91	0.63	0.62	0.55	0.81	0.80

**Table 5.7: Subtest intercorrelations of TALA**

From Table 5.7 it is noticeable that only four of the ten subtest intercorrelations fall within the desired parameters identified by Van der Walt and Steyn (2007). However, only one intercorrelation is marginally too strong (0.56). Of the remaining five correlations, three can also be labelled as marginal cases (0.10, 0.15, 0.17), whilst the remaining two correlations would possibly be regarded as too low (0.05, 0.04). On the other hand, three of the five subtests correlate desirably with

the whole test and indicate a more satisfactory subtest correlation score. The average P-value of the test is slightly too low with a score of 41.85%, as we have noticed from other analyses.

Below is an excerpt of the DIF statistics (Table 5.8), also referred to as the Mantel-Haenszel test (Weideman & Van der Slik 2008:166), generated by TiaPlus. I have chosen only the first ten items, since including all 60 items will take up too much space. The complete list can be seen as an Annexure F. Additionally, the first ten items are of importance since that list includes the only flagged item of the entire test, which is item number 2. The flagged item is indicative of possibly being biased towards one of the test groups and can be identified as an item performing dissimilarly between test groups (Weideman & Van der Slik 2008:171). This means that one may need to scrutinise this item to see whether it should be modified. However, one flagged item for an entire test consisting of 60 items can be regarded as negligible. The flagged item is indicated with an asterisk, and can be found under the z-value column, where an item is identified when it exceeds a z-value score lower than -2.58 or above 2.58 (CITO 2007).

Label	Item	DIF stat	z (stand)
	1	0.0584	-2.0649
	<b>2</b>	<b>0.0470</b>	<b>-2.5978 *</b>
	3	0.2030	-1.7471
	4	0.7538	-0.3675
	5	1.5407	0.5352
	6	1.7119	0.6859
	7	1.8767	0.8491
	8	2.1985	0.6377
	9	1.6676	0.6815
	10	1.5304	0.5057

**Table 5.8: DIF statistics for TALA**

Misclassifications are another measure of test consistency or reliability to take into account. Misclassifications identified by TiaPlus include students who should have passed the test, but did not, and students who passed the test when they in fact have performed more poorly (Weideman & Van der Slik 2008:169). This can occur for many reasons, but relates to the degree of measurement error present in the administration of the test. Since no test is perfect, some measurement error can be expected, and hence misclassification will occur, as shown below in Table 5.9:

Misclassifications:		Alpha based		GLB based	
-Rxx' case	Percentage	: 15.2	Percentage	: 10.6	
	Number	: <b>37</b>	Number	: 26	
-Rxt case	Percentage	: 11	Percentage	: 7.6	
	Number	: 27	Number	: <b>18</b>	

**Table 5.9: Misclassifications for TALA**

TiaPlus identifies possible misclassifications based on Cronbach's coefficient Alpha and the GLB coefficient (referred to in 5.4 above). The Rxx method generates calculations using the reliability ( $\rho$ ) of a test to correlate test scores with possible parallel test scores whilst the Rxt method utilises the square root of the  $\rho$  of a test to correlate the observed test scores with the true test scores (CITO 2005:19,30). Taking into account both the scores, one can identify an average misclassifications score. For TALA, for example, the highest possible number of test takers which could have been misclassified is 37 whilst the least possible number of test takers which could have been misclassified is 18 (Table 5.9). This is an average of 27 misclassifications. This indicates that just fewer than 14 test takers might have been misclassified by this test, assuming that more or less 50% of test takers would have benefitted, whilst 50% were possibly placed at a disadvantage. In essence, this is not an entirely desirable outcome, and would, in the case of a high stakes test, have raised the question of whether those potentially misclassified to their disadvantage should not be offered a second-chance test similar to the first.

On a more positive note, a desired Alpha score of 0.82 is indicated by TiaPlus, as well as a GLB score of 0.94 (Table 5.7), which are both acceptable values. Alpha, however, is usually a much more conservative measure of consistency than GLB, and therefore usually lower (Weideman & Van der Slik 2005:26).

### 5.4.2 The second test

The subtest intercorrelations of the second test can be seen in the following table:

Subtest	Test	1	2	3	4	5
Scrambled text	1	0.64				
Vocabulary know	2	<b>0.70</b>	<b>0.42</b>			
Interpreting gr	3	<b>0.81</b>	<b>0.50</b>	<b>0.49</b>		
Text comprehens	4	<b>0.89</b>	<b>0.43</b>	0.57	0.64	
Grammar & text	5	<b>0.74</b>	<b>0.33</b>	<b>0.37</b>	<b>0.52</b>	<b>0.54</b>
Number of testees :	240	240	240	240	240	240
Number of items :	60	5	10	8	25	12
Average test score:	33.23	2.51	6.23	5.00	13.65	5.83
Standard deviation:	10.53	1.99	1.75	2.43	4.44	2.83
SEM :	3.40	0.74	1.30	1.09	2.22	1.54
Average P-value :	<b>55.39</b>	50.25	62.33	62.55	54.62	48.58
Coefficient Alpha :	<b>0.90</b>	0.86	0.44	0.80	0.75	0.70
GLB :	<b>0.97</b>	0.90	0.64	0.85	0.86	0.89
Asymptotic GLB :	0.96	0.90	0.53	0.84	0.84	0.84

**Table 5.10: Subtest intercorrelations of the second test**

Subtest intercorrelations for the second test are slightly more satisfactory than the subtest intercorrelations for TALA. This is another indication that this test may, among the administrations of the various measurements employed in this study, have been more robust and useful. Of the ten subtest intercorrelations, eight fall within the preferred parameters, whilst only two subtest intercorrelations can be regarded as possibly too strong (0.57 and 0.64). At the same time, four of the five correlations between the subtests and the test as a whole fall within the specified parameters, whilst only one correlation is slightly too low. Overall, this is a much more satisfactory outcome than is the case for TALA. In this instance, the average P-value of the test is again just above expectation, with a score of 55.39%. This indicates that the percentage of test takers who chose correct answers slightly exceed the percentage of test takers who selected incorrect answers.

Concerning misclassifications, the second test has again performed above expectation. As can be seen in the table below (Table 5.11), no misclassifications can be identified for all calculations using both Cronbach's Alpha and GLB. This indicates that possibly not one test taker was disadvantaged by the second test; a rare occurrence.

Misclassifications:		Alpha based		GLB based	
-Rxx' case	Percentage	: 0.1	Percentage	: 0.1	
	Number	: 0	Number	: 0	
-Rxt case	Percentage	: 0.1	Percentage	: 0	
	Number	: 0	Number	: 0	

**Table 5.11: Misclassifications for the second test**

Additionally, the DIF statistics of the second test can be viewed as annexure G. For the second test, no item has been flagged as being biased towards any of the test groups, which once again is wholly satisfactory, even remarkable. One can also take note of the indicated Alpha score of 0.90, and the indicated Greatest Lower Bound score of 0.97, which both are well above the desired parameters. These are further indications that among the specific administrations of assessment measures in this study, the TALA-like test has performed better than the original TALA.

### **5.5 Regression and related analyses**

Three further sets of statistical analyses were carried out for this study in consultation with Robert Schall of the Statistical Consultation Unit (SCU) at the University of the Free State, and the full set of reports is attached as annexures H-L. A correlational analysis was done (Annexure H) as well as a regression analysis (Annexure I) and an ANCOVA analysis (Annexure J, K, L). These three kinds of analyses are either complementary or statistically similar, each yielding results that may variously support the other analyses.

A regression analysis was done on the results acquired through the administration of the two tests on two Bloemfontein based schools. The aim of the analyses is to

establish whether notable comparisons exist between the academic performance of the students and the results the students obtained for three measurement devices. These include the two academic literacy tests, TALA and the second test which was reduced and amended according to the specifications for TALA, and the English Home Language school examination paper of June.

The correlations between the three sets of data produced the following results:

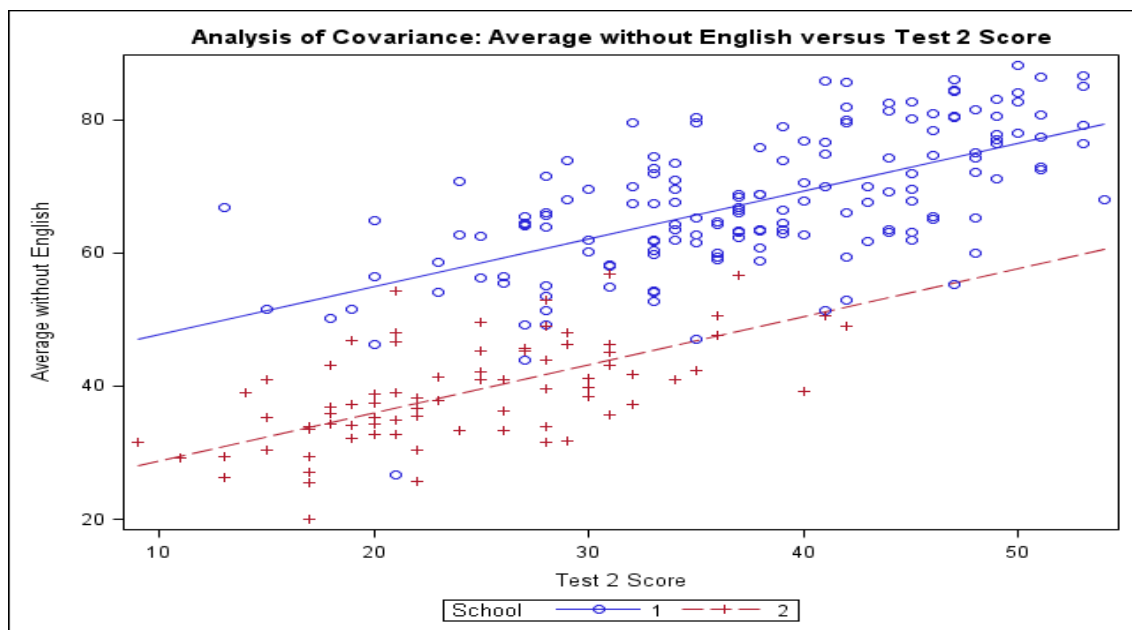
	<b>Average without English</b>	<b>Test 1 (p)</b>	<b>Test 2 (p)</b>	<b>English (p)</b>
<b>Average without English</b>	1.00000	0.45512 (<.0001)	0.78491 (<.0001)	0.81810 (<.0001)
<b>Test 1 (p)</b>	0.45512 (<.0001)	1.00000	0.35253 (<.0001)	0.31814 (<.0001)
<b>Test 2 (p)</b>	0.78491 (<.0001)	0.35253 (<.0001)	1.0000	0.78408 (<.0001)
<b>English (p)</b>	0.81810 (<.0001)	0.31814 (<.0001)	0.78408 (<.0001)	1.00000

**Table 5.12: Correlation analysis results**

Looking at the table above, one soon notices that English seems to be a better predictor of the students' academic average than both Test 1 (TALA) and Test 2 (TALA-derivative). Where English has obtained a score of 0.81810, which is a high correlation, Test 2 has attained a close second position with a score of 0.78491 and Test 1 trails behind with a medium strength score of 0.45512. These scores are highly significant, since p, which represents the probability that correlations are accidental, is very low. The p value shown in Table 5.12 is below .0001, which indicates that the results obtained are not accidental, and therefore the scores obtained are statistically significant. Additionally, while it is somewhat disappointing that TALA did not perform as well as was anticipated, it is on the

other hand exciting to see that the second test comes close to predicting average performance nearly as well as the English Home Language mark. This is impressive, since it is a comparison between a familiar kind of assessment and an unfamiliar test which was written once by the students, and one that is here competing with a mark based on 10 years of accumulative and continuous assessment done under familiar circumstances. The comparison, in other words, is more appropriately characterised as being between the familiar and the unfamiliar, than between reliability and unreliability.

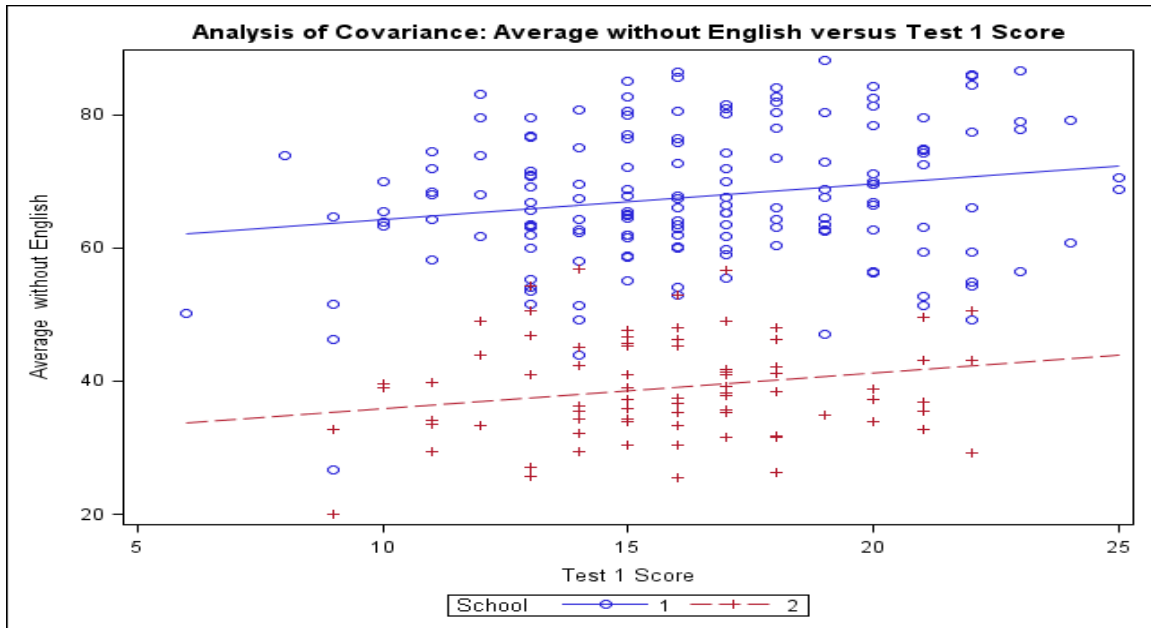
Additionally, in his review of the regression analysis and ANCOVA analysis, which produced results that confirm the correlation analysis, Schall (SCU 2014) reports the following: “These results suggest that information from Test 2 improves the prediction of Average school mark excluding English, relative to a prediction based only on the English mark.” This can be seen when one looks at one of the figure plots generated by the ANCOVA analysis:



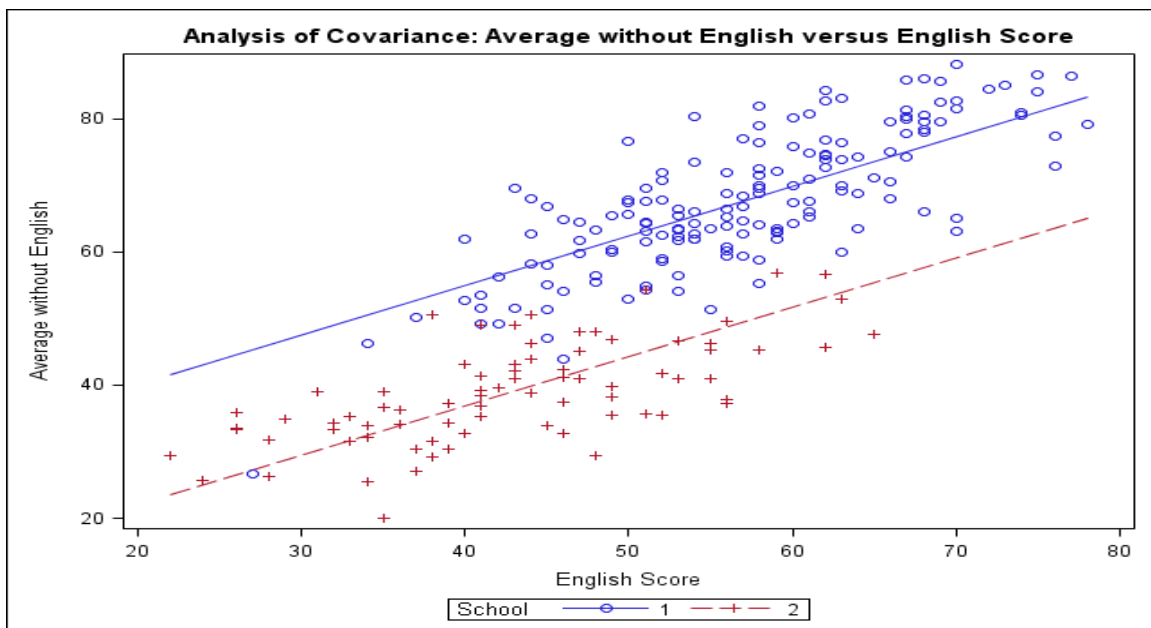
**Figure 5.1: Covariance analysis between average excluding English and test 2**

Looking at the figure plot above, one notices that the data are congregated around the data lines, which substantiates that a relationship is evident between the student averages and the results of test 2. The figure plot for TALA displays a more

scattered data plot, as seen below in Figure 5.2, which together with the gradient, discloses that a more limited relationship can be identified between the results of TALA and the student averages. In contrast, when one looks at the data plot for English and the average excluding the English Home mark, one would immediately notice that the data are very closely congregated around the data lines, as seen below in Figure 5.3. This, again, indicates that a relationship exists between the two sets of data.



**Figure 5.2: Covariance analysis between average excluding English mark and test 1**



**Figure 5.3: Covariance analysis between average excluding English and English**



Lastly, the results of the regression analysis also substantiate the finding that the English Home mark predicts the average of the students more accurately than the academic literacy test marks (Annexure I). Three different regression models were used with varying independent variables and one fixed dependent variable, namely the average school mark of the students, excluding the English Home Language mark.

The independent variables of the first model included the results of TALA and the results of the second test, as well as the results of the student's English Home Language subject mark. The regression analysis done on the first model confirmed that the English Home Language mark is the most accurate in determining the average mark of the students. Additionally, the first model also confirmed that TALA is not an adequate predictor of the student's average, but that the second test improves the prediction force of the English Home Language mark when the two are used in conjunction.

For the second model, the independent variables included only the results of TALA and of the English Home Language mark. The regression analysis done on the second model found and confirmed that the results of TALA are not significant predictors of the student's average mark. Moreover, TALA's results do not improve the predictive force of the English Home Language mark when both are used in conjunction.

The last model's independent variables included the results of the second test and the results of the English Home Language mark. The regression analysis of this model confirmed the outcome of the first model, that the results of the second test improve the predictive force of the English Home Language mark, but that the English Home Language mark continues to be the most adequate predictor of the student's academic average and academic potential.

## 5.6 Discussion

The results of the regression analysis are to an extent unexpected, since they indicate that the English Home Language mark predicts average academic performance more accurately than the administered academic literacy tests that were administered. Even though this was an unexpected outcome, it may, however, be premature to conclude that the academic literacy tests used give less valid measures of the ability to handle the demands of academic language. A variety of explanations may be offered which refer to the complexity and richness of the analyses attempted, and that indicate that these issues may require further investigation. To my mind, six main reasons for this result can be given and argued.

I have stated in previous chapters the reasons why I have chosen Grade 10 students instead of Grade 11 or Grade 12 students (Chapter 3). Not only were Grade 10 students more readily available to participate in this study, making them a more convenient sample group, but their marks are also not under that much scrutiny, at this stage, especially when compared to the Grade 12 marks (Weideman, Du Plessis & Steyn 2015:2). One should also note that the first test, TALA, might have been too difficult for Grade 10 students, as the test was originally designed for Grade 12 students. The performance of TALA during its previous pilot at higher grade level was, in fact, superior to its less sterling performance here. TALA, for instance, obtained a notable Alpha score of 0.958 during its first pilot on Grade 12 students (Weideman, Du Plessis & Steyn 2015:14), while its reduced 60 mark version also scored above 0.9. TALA's Afrikaans counterpart, TOGTAV, obtained equally impressive alpha scores. These can be seen in the table (5.13) that follows (Weideman, Du Plessis & Steyn 2015:14):

Version of test	Reliability (alpha)
TALA first pilot (187 items; n = 1244)	0.958
TOGTAV 1 first pilot (196 items; n = 368)	0.955
TOGTAV 2 first pilot (187 items; n = 357)	0.944
TALA (reduced 60-item version; n = 1244)	0.900
TOGTAV 2 (reduced 60-item version; n = 357)	0.831

**Table 5.13: TALA's performance during its first pilot**

One could also question whether it was too ambitious to compare a 300 mark examination, combined with the accumulation of 10 years of previous assessments, with one or two 60 mark tests. It is known that a longer test measures more reliably than a short test. When more test items are included in an assessment, they lend more opportunities for test takers to achieve their true potential, which is likely in turn to reflect in more reliable results.

Another possibility to take into account is that learners are usually well prepared and taught for the assessments they are given at school, in contrast to the two academic literacy tests which were given to the learners in an unfamiliar format under unfamiliar circumstances. An assessment external to a system, one component of which is thorough teaching and diligent preparation, is likely to find it difficult to compete with a system-internal measure.

Additionally, it may be that the average performances across school subjects are too homogeneous, all tending towards the mean, or academic average, that was the criterion for this study. A correlation of different subjects with the overall mark may show this, but this explanation falls outside the scope of this study, and will not be further pursued here.

A study by Van Rooy and Coetzee-Van Rooy (2014:6) provides another possible explanation. Even though the correlations between the academic literacy tests and the average academic score of the students were weaker than expected, it should not be doubted whether a connection exists between performance in academic

language and overall academic success. Van Rooy and Coetzee-Van Rooy (2014) observe that certain language related abilities are needed by students to be able to pursue tertiary studies and that these abilities may not be adequately addressed by the school curriculum. They are, however, of the opinion that school marks in combination with an additional measure of academic literacy are needed to identify students who should benefit from academic literacy interventions (Van Rooy & Coetzee-Van Rooy 2014:5). Their study concludes, furthermore, that the results of academic literacy interventions are in fact the best predictor of first year performance. In retrospect, one could wonder whether tests, no matter how well constructed and developed, would perhaps not always lack the richness and depth which programmes of longer duration offer in terms of accuracy and reliability. 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, it would appear that further separate studies are needed to establish again the relationship between language ability and overall academic performance. Although extensive research would have to be done to identify what type of relationship exists between the two, and how strong that relationship is, it would be beneficial not only to the further pursuit of investigations I have initiated in this study, but to many others as well. When that relationship has been identified, many uncertainties could possibly be dealt with more effectively. Whether this study is of functional importance would be one of the questions that could subsequently be answered.

Lastly, and perhaps most contentiously, the results bring forth the question of whether there is any difference between general and specific language ability. This in turn then questions whether language test design should at all be aimed at testing specific language abilities within a certain language context. Patterson and Weideman (2013a:107,109) present a convincing argument that academic

discourse is in fact a distinctive material lingual sphere with a specific set of properties, different to other lingual spheres (2013a:107), which aligns with the popular belief that language is context specific and dependent. It goes without saying that one would use a different type of language when writing for aesthetic purposes than one would use when writing an informative or instructive text. Not only does the purpose of the text change, but the entire style, word choice, and structure of the text change together with the specific set of skills needed for every specific language task. Consequently, it should be beneficial to assess ability in all the different material lingual spheres and the specific set of skills associated with them, using different assessments. I therefore support the idea that there is a difference between general and specific language ability, since it is evident that not all language related skills can be attributed to a single material lingual sphere.

In the concluding section of this chapter, the claims posed in the previous chapter will be discussed in detail using the results which were obtained through the regression and related analyses.

### **5.7 Answering the claims**

The claims articulated in the previous chapter may now be treated as follows:

**Claim 1: The refined test will demonstrate a better correlation with students' average mark, and therefore their possible academic success, than the unrefined test.**

This claim was refuted, since the Home Language mark correlated more closely with students' average academic mark than the academic literacy tests did. It is, however, of importance to note that TALA did perform better at its appropriate level during its previous Grade 11 and Grade 12 pilot (Steyn 2015). In the next chapter I shall discuss this matter in further detail.

**Claim 2: Both tests will demonstrate a better correlation with students' average mark than their Home Language mark would demonstrate.**

**Claim 3: Both tests will demonstrate an even better correlation with students' average mark, excluding the Home Language mark.**

Both the second and third claim, similar in essence, were refuted, although it might be premature to conclude that only Home Language marks could be used to channel students into academic literacy programmes. Van Rooy and Coetzee-Van Rooy (2014:4) explain that a discrepancy exists between what students are taught in school concerning writing in English and what lecturers at universities expect of students. Additionally, Van Rooy and Coetzee-Van Rooy (2014:4) also observe that the textbooks, and other materials which are used in school situations, do not necessarily encompass appropriate examples of what academic language entails. Students are therefore highly unlikely to be prepared for the level of writing and reading with which they are faced with at university. This emphasises why English Home Language marks alone may not be an adequate indication of what level of students should be included in academic literacy programmes. The finding that the second test yields additional information about learner's performance in conjunction with the Home Language mark is an indication that an academic literacy measure may yet be useful, or at least more useful than the English Home Language marks on their own.

**Claim 4: The results of the tests will enable me to determine the relative power of the unrefined test and will also enable me to make recommendations for the refinement of similar tests.**

The analyses which have been carried out have indeed identified which test items have performed within the specified parameters and also which items need to be refined. In the next chapter I shall therefore also address this matter in more detail, whilst specifically focusing on the improvement of items which did not perform satisfactorily.

**Claim 5: More useful conclusions regarding academic literacy levels can be drawn from academic literacy tests than can be drawn from other types of measurements.**

After scrutinising the results of the various analyses which have been carried out, this claim has also been refuted, though potentially contradictory findings have been raised by Van Rooy and Coetzee-Van Rooy (2014). Whilst they have noticed that school marks seem to be less able to predict the possible academic performance of students, they have concluded that the results of assessments on support modules or academic literacy interventions are much more accurate and reliable in determining possible academic success or potential at first year level (Van Rooy & Coetzee-Van Rooy 2014:5). Additionally, they also mention that the school curriculum does not place enough emphasis on developing abilities that are needed for tertiary studies, especially those pertaining to language (Van Rooy & Coetzee-Van Rooy 2014:7). In conjunction with this, Du Plessis's (2014) recent study questions the validity of a part of the assessment of the Home Language examination, while another report, by Du Plessis, Steyn and Weideman (2014), confirms that there may be problems not only with parts, but with the entire Home Language examination in Grade 12. This again emphasises the need to have additional measurements of academic literacy and consequently urges us to pursue further research into this. I therefore believe that the development, administration and refinement of academic literacy tests are worth paying further attention to, although it appears that they should preferably be administered as close as possible (Grade 12 or post-Grade 12) to university enrolment. This does not mean that the academic literacy of students develops only at that stage, but merely that a measurement closer to university enrolment (where the bulk of the current research is focused) might yield more easily interpretable results. That seems a more appropriate time to make the desired measurement, and, in view of these researcher's findings, also more likely to be successful predictors of academic performance, at least for the first year at university. For measuring academic literacy at lower levels, such as Grade 10, one would have to design tests at the

appropriate level, and perhaps of substantial length, a point that I return to below and in the next chapter.

### **5.8 Conclusion**

Since the second test has performed above expectation during its administration, it could therefore be advantageous to refine this test and pilot it again on students in the future. Another possibility is to extend the test, which could increase its reliability. In the next chapter, I will therefore discuss the refinement of the second test, after which I will review TALA's performance, or lack thereof, in more detail.



## Chapter 6

### Refinement of the second test

#### 6.1 Introduction

In this chapter, I will explore some of the notable empirical properties which are evident in TALA (Test of Advanced Language Ability) and are derived from statistical analyses done on the results obtained from its previous administrations (which was why TALA was chosen for this specific study). However, since TALA did not perform as well as expected, a possible refinement of the second test used in this study will rather be presented and motivated.

#### 6.2 Why refine the second test?

In the previous chapter we learned that TALA did not perform as well as was expected. This was surprising, since TALA, which has already been refined, had been administered on a previous occasion and had generated excellent results on all the analyses that were carried out on its results. Below a table can be seen which includes the reliability index scores of TALA's previous pilot (Weideman, Du Plessis & Steyn 2014:14):

Version of test	Reliability (alpha)
TALA first pilot (187 items; n = 1244)	0.958
TALA (reduced 60-item version; n = 1244)	0.900

**Table 6.1: TALA's reliability indices of its first pilot**

For its first pilot, for instance, it obtained a 0.958 Cronbach's Alpha score when tested on a group of 1244 students. After being reduced from a 187-item test to a 60-item test, it continued to achieve a 0.900 Cronbach's Alpha score (Weideman, Du Plessis & Steyn 2014:14), which can be regarded as an indication of its high level of reliability and consistency. Additionally, the idea that a longer test assesses

more reliably is substantiated by these scores, since the longer TALA version achieved an even better Cronbach's Alpha than the shorter TALA version.

A possible explanation can, however, be provided for the distinct differences in results between the two piloted occasions. TALA was constructed and intended for Grade 11 and Grade 12 students (Steyn 2015:28), but for this study, it was administered to Grade 10 students. Initially, the test developers thought a difference of one or two school year levels would not affect the outcome of the results as much as it did in the end. This noticeable difference in results between the grades can possibly be attributed in part to the difficulty of the texts which were utilised in the test. The second test's texts were more appropriate for Grade 10 students whereas TALA's texts were chosen more specifically for Grade 11 and Grade 12 students. The Flesch-Kincaid level of the texts included for the second test, for example, scored 10.5 and 8.6 respectively (Steyn 2015:5). The Flesch-Kincaid level for Grade 10 students should preferably be pitched no higher than level 10, which means that the first text is pitched at the appropriate level for Grade 10 students, and that the second text was slightly below the level 10 score, but still within an appropriate range. TALA's Flesch-Kincaid levels, on the other hand, were 12 and 11.7 respectively (Steyn 2015:29). These scores are tabulated below:

	<b>TALA</b>	<b>The second test</b>
<b>Flesch-Kincaid level (Text 1)</b>	12	10.5
<b>Flesch-Kincaid level (Text 2)</b>	11.7	8.6

**Table 6.2: Flesch-Kincaid levels of texts in both TALA and the second test**

Consequently, the texts used for TALA were probably too difficult for Grade 10 students, and it is therefore possible that the test items based on them were also too difficult. One can therefore argue that TALA is too difficult for Grade 10 students, whereas the second test is a more appropriate measure for the same group of Grade 10 students, a conclusion that is borne out by some of its empirical properties reviewed in the previous chapter, particularly the averages (55% versus 42%)

obtained by the candidates on the respective tests, as well as their different reliability indices.

The refinement of a test includes the modification of test items which did not perform as desirably as they should have in light of the Iteman and TiaPlus analyses (discussed in the previous chapter). Replacing entire texts does not ordinarily fall within the scope of initial test refinement. Also, whereas this was the first administration of the second test, TALA was administered on a previous occasion and has already undergone refinement. Additionally, because the second test outperformed TALA for this study, and has been shown to be the more appropriate test at this level (Grade 10), I shall shift my focus to the refinement of the second test only.

One may question whether it is necessary to attempt to refine this test, in light of the claims refuted in the previous chapter. Even though on its own the test did not predict academic performance as well as the English Home Language marks did, however, it is notable, first, that this second test came close to predicting academic performance as well as did the English Home Language marks. As I have observed, one should consider that the test was up against a school system which stretches over ten years of continuous assessment and customarily thorough preparation for assessments. Additionally, in the previous chapter, it was reported that the second test, when combined with the English Home Language marks, predicted the academic average of the students even more accurately than any single other measure. Finally, especially in view of the finding of Van Rooy and Coetzee-Van Rooy (2014) about the inability of school marks for language being able to predict performance at tertiary level, having an assessment of academic literacy (of sufficient length) may, in light of the second test's strengthening of the prediction, yet be useful. Van Rooy and Coetzee-Van Rooy (2014:3) argue that because of the crucial difference between English instruction at school and the expectations regarding performance in academic English at university, one cannot solely rely on

school marks to identify at risk students, a finding borne out by Du Plessis's (2014) study. The better predictor, Van Rooy and Coetzee-Van Rooy found, was an academic literacy intervention of longer duration. How students were identified, in their case, as being eligible for such an intervention, derived from the prior administration of an academic literacy test. One could therefore argue that a longer, and hence potentially more reliable assessment measurement, might give one an even better chance of improving the predictive quality of such a test. This is why I believe that refining this test before its further administration is a valuable potential contribution to our understanding of the risk associated with academic literacy levels. The tests in Weideman and Van Dyk (2014), it should be noted, are not yet refined, and with the exception of the one test from that book that was adapted for use in this study, have neither been piloted nor refined further. Moreover, if an appropriate level test is administered early, it may still be a useful indication perhaps not of current performance, but of future performance (at tertiary level). The refinement of this test will be the main objective of this chapter.

### **6.3 Potential refinements to test items**

#### **6.3.1 Parameters for a productive item**

The IteMan and TiaPlus analyses indicated that the second test performed more desirably than TALA, comparing Alpha scores, Rpbis scores and Mean Percentage Correct scores (previous chapter). However, some items of the second test also did not fall within the prescribed parameters of performance and are therefore eligible for reconsideration. I shall identify some of these items in a table (Table 6.3) below, discuss why their performance was not satisfactory, and suggest how they may be modified.

There are several parameters of item productivity for the test used in this study. First, the Rpbis score of a correct item should be higher than any of the other incorrect options given for that same item. The Rpbis score should additionally be

a positive number and preferably be above 0.15. Second, the facility or P-value of an item should be in the vicinity of 0.5 (Guyer & Thompson 2011), but for this study I have chosen to accept values ranging from 0.2 to 0.8 as suitable. The items of the second test which did not perform within these parameters are listed below in Table 6.3:

	<b>Rpbis</b>	<b>P-value</b>
<b>Item 6</b>	<b>-0.434</b>	0.596
<b>Item 7</b>	0.235	<b>0.904</b>
<b>Item 12</b>	0.320	<b>0.929</b>
<b>Item 13</b>	<b>-0.118</b>	<b>0.129</b>
<b>Item 25</b>	<b>0.111</b>	0.517
<b>Item 27</b>	<b>0.135</b>	0.658
<b>Item 28</b>	<b>-0.045</b>	<b>0.179</b>
<b>Item 32</b>	<b>-0.129</b>	<b>0.146</b>
<b>Item 45</b>	<b>0.114</b>	0.429
<b>Item 52</b>	<b>0.091</b>	0.383

**Table 6.3: Summary of items which did not perform satisfactorily as indicated by Iteman 4.3**

Items can also simply be removed from an assessment, which reduces the number of items in a test if they are not subsequently replaced by others, for example, by items that have performed well in other pilots. However, for the purpose of this study, I shall keep to the possible refinement of the items mentioned above based on information I have taken from the Iteman 4.3 analysis (Annexure E).

### **6.3.2 Refinements of individual items**

Item 6 did not discriminate as well as it should have, since more students who got a lower overall score on the test as a whole, got this one right than did those whose overall performance was in the upper segment of the results. This item can be modified by changing some of the words to possibly lessen the ambiguity of the question. Seen below is the question as it appeared in the test and, thereafter, its modification in bold:

6. To patent an invention is a great \_\_\_\_\_ for any inventing genius.
- A. bereavement
  - B. achievement
  - C. endorsement
  - D. inducement

- 6. To patent an invention is a great \_\_\_\_\_ for any aspiring inventor.**
- A. bereavement**
  - B. inducement**
  - C. confinement**
  - D. achievement**

The P-value of item 7 scored 0.904 (Annexure E), which is too high, and indicates that almost all the test takers chose the correct answer, meaning the item was too easy for these specific test groups. Once again the original item is displayed below, followed by its suggested possible modification:

7. In order for any inventor to make a success of an invention he/she has to be \_\_\_\_\_ to the process.
- A. addicted
  - B. connected
  - C. committed
  - D. indented

- 7. In order for any inventor to make a success of an invention he/she has to be \_\_\_\_\_ to the process.**
- A. addicted**
  - B. embedded**
  - C. committed**
  - D. indented**

Item 12 had a P-value score that was too high. It obtained a P-value score of 0.929 (Annexure E), which indicates that the question is too easy. Below both the original question and its possible modification are given below:

12. The ability to \_\_\_\_\_ a simple idea into a mind-blowing invention is an art that only truly genius inventors are able to perfect.
- A. inform
  - B. transform
  - C. conform
  - D. deform

12. The ability to \_\_\_\_\_ a simple idea into a mind-blowing invention is an art that only truly genius inventors are able to perfect.

- A. reform
- B. transform
- C. conform
- D. perform

Item 13 obtained a P-value of 0.129 (Annexure E) which indicates that the question is too difficult. Additionally, more students who got a lower overall score on the test as a whole, got this one right than did those whose overall performance was in the upper segment of the results. This item can be modified by perhaps moving the correct answer to option A. “Initiates” can also be discarded by replacing it with a less likely choice as alternative:

13. One invention often \_\_\_\_\_ the introduction of another, which results in technology being pushed even further.

- A. initiates
- B. mediates
- C. arrogates
- D. rotates

13. One invention often \_\_\_\_\_ the introduction of another, which results in technology being pushed even further.

- A. mediates
- B. propagates
- C. arrogates
- D. rotates

The Rpbis value of item 25 is slightly too low with a score of 0.111 (Annexure E) and indicates that there are more students with a lower language ability who chose the correct answer than there are students with a higher language ability who chose the correct answer. This item can be modified by changing the incorrect answer which was chosen by many high language ability students, which is option A:

25. The phrase “turning the entire world into a computer” in paragraph 1 means that

- A. the earth as a whole works as a computer.
- B. everything can form part of this computer.
- C. the earth is controlled by a computer.
- D. the world is nothing but a computer.

25. The phrase “turning the entire world into a computer” in paragraph 1 means that
- A. the earth can be programmed like a computer.
  - B. everything can form part of this computer.
  - C. the earth is controlled by a computer.
  - D. the world is nothing but a computer.

The Rpbis value of item 27 is slightly too low with a score of 0.135 (Annexure E). As is the case with the previous item, more emphasis needs to be placed on the correct answer, which is option C. One way to do this might be to move the correct answer to another position:

27. A good description of what SixthSense does, is given in paragraph \_\_\_\_
- A. 1
  - B. 2
  - C. 3
  - D. 4

- 27. A good description of what SixthSense does, is given in paragraph \_\_\_\_**
- A. 4
  - B. 3
  - C. 2
  - D. 1

Item 28’s P-value score was also too low. It obtained a score of 0.179 (Annexure E). Additionally, the Rpbis score is also too low. This item can be modified by changing option A, since it was incorrectly chosen as the correct answer by higher language ability students, as seen below:

28. The word ‘augment’ in paragraph 2 means to \_\_\_\_\_ of the physical world
- A. adapt to another form
  - B. change the perception
  - C. change the reality
  - D. increase the quality

- 28. The word ‘augment’ in paragraph 2 means to \_\_\_\_\_ of the physical world**
- A. adapt the conception
  - B. change the perception
  - C. change the reality
  - D. increase the quality

Item 32 may have been too difficult for the test groups since it obtained a P-value score of only 0.146 (Annexure E). Consequently, more students chose the wrong



answer at option A when the correct answer is actually option D. By exchanging the placements of the possible answers, this item can possibly be modified to perform better:

32. The phrase “to have access to relevant information” in paragraph 4 is related to the phrase

- A. “easy access computing” in paragraph 1.
- B. “go one step further” in paragraph 1.
- C. “gesture driven computing” in paragraph 2.
- D. “a system that can display” in paragraph 3.

**32. The phrase “to have access to relevant information” in paragraph 4 is related to the phrase**

- A. “a system that can display” in paragraph 3.**
- B. “go one step further” in paragraph 1.**
- C. “gesture driven computing” in paragraph 2.**
- D. “easy access computing” in paragraph 1.**

For item 45, more students who got a lower overall score on the test as a whole, got this one right than did those whose overall performance was in the upper segment of the results. This is indicated by a low Rpbis score of 0.114 (Annexure E). This means that the phrasing of the question should perhaps be improved in order to eradicate this problem:

45. From the first sentence of paragraph 8 one can conclude that

- A. other practical functions could well be added to SixthSense.
- B. SixthSense could be sold commercially when it has been completed.
- C. SixthSense would only be worthwhile once it is sold commercially.
- D. if any additions were made to SixthSense, they would be able to sell it.

**45. Concentrating on the first sentence of paragraph 8, one can conclude that**

- A. other practical functions could well be added to SixthSense.**
- B. SixthSense could be sold commercially when it has been completed.**
- C. SixthSense would only be worthwhile once it is sold commercially.**
- D. if any additions were made to SixthSense, they would be able to sell it.**

The last item, number 52, also obtained a too low Rpbis score of 0.091 (Annexure E). This item can be modified by eliminating the incorrect answer which was often chosen as the correct answer, which is option A:

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

- A. it
- B. way
- C. you
- D. completely

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

- A. not**
- B. way**
- C. you**
- D. completely**

If these items can be refined and if in a second pilot they are shown to be more productive, the test is likely to be more useful. However, if a longer test is indicated to increase reliability further, as well as predictive value, as has been suggested above, one might wish to augment this test with another one of similar length, so one would have an assessment of 120 marks instead of 60 marks. The specifications for such a test may well be the same as for the current one, since the subtest intercorrelations reported on in the previous chapter are an indication of construct validity, and the difficulty level of the texts used is appropriate.

#### **6.4 Conclusion**

It would be interesting to pilot the test once more with the items that have been refined, and to run the analyses on the results again. This, however, goes beyond the scope of the current study and would subsequently need to be attempted as a separate investigation. In the next, and concluding chapter, I will give a summary of the study, combined with general findings and possible recommendations for studies of a similar nature.

## **Chapter 7**

### **Conclusions and recommendations for future research**

#### **7.1 Introduction**

The number of students who enrol at South African universities has increased substantially over the past two decades. University enrolment often requires of prospective students to provide proof of their academic potential. This can either include writing an academic literacy test, such as the NBTs which are written at many universities across the country, or it can be based on a student's matric examination results. The second option, however, has been in dispute over the last couple of years, as the credibility and reliability of the matric results are queried more and more.

#### **7.2 Summary**

This study was initiated to determine whether matric results can be used as a reliable measurement of academic literacy, especially when aimed at access to and performance at tertiary education institutions, or whether scores on academic literacy tests would be a more reliable source of the ability to handle academic discourse. Two different academic literacy tests were administered at two Bloemfontein based schools to Grade 10 students. The first test is the Test of Advanced Language Ability (TALA), designed and developed according to a construct based on the idea that high level language ability, as specified in the school language curricula, is related also to using language for academic purposes, and requires a specific set of language skills. The second test is from a test book that is about to be published (Weideman & Van Dyk 2014) and was reduced using the specifications of TALA, since TALA had already been piloted on Grade 12 students on a previous occasion and had obtained desirable results. The results of the two tests were then compared to the student's June examination school results,

and more particularly, their mark for English Home Language and their June average.

I anticipated that the results of the academic literacy tests would correlate more closely with the student's overall average mark than the English Home Language mark, since a noticeable misalignment can be observed between the prescriptions made in the school curriculum and the actual assessments that take place (Weideman, Du Plessis & Steyn 2014:5). Even though the curriculum prescribes that students should practise skills associated with academic discourse, it can be argued that not enough emphasis is placed on the importance of academic language in the school curriculum (Coetzee & Coetzee-Van Rooy 2014:4). However, when the results became available, a more definite correlation was evident between the student's overall average mark and their English Home Language mark than between the results of the academic literacy tests and the student's average mark. TALA, surprisingly, did not perform as well as it did in its first pilot. The second test, on the other hand, which was deemed the less stronger of the two, being derived from an untested and as yet unrefined measurement, performed better. Although the second test did not correlate as closely with the average mark as the English Home Language mark, it improves the predictive force of the English Home Language mark when the two are used in conjunction. In the next section I shall discuss this matter in more detail, along with other recommendations for research of the same nature.

### **7.3 Recommendations**

Some of the complexities revealed by the analyses in this study were unforeseen, as might be expected. I shall therefore now present some recommendations for others who wish to attempt a study of the same nature.

Firstly, when attempting to compare two different measures of academic literacy, for instance, specifically designed academic literacy tests and English Home Language assessments, one first needs to determine whether the two can be compared, especially when one considers the length of the two measurements. The 60-mark academic literacy tests were possibly too short to be compared to a 300-mark English Home Language examination, since a longer test has a greater chance of yielding a reliable result than a shorter one. Additionally, where the academic literacy tests can be regarded as self-standing entities, a school examination forms part of a more complex system with accumulative preparation. This observation supports the argument of Van Rooy and Coetzee-Van Rooy (2014) which proposes that the most accurate and reliable measures of the ability to handle academic discourse are academic literacy interventions which, much the same as the school system, have the advantage of being attempted over a much longer period of time than a test. Once again, length of intervention seems to trump short term involvement as regards reliability.

Secondly, academic literacy tests should be administered only to the intended target group. For this study, it was thought that the differences between Grade 10 students and Grade 12 students would not impact the results of the tests in any significant manner. It was found, however, that even two years of school level difference can be meaningful. The results obtained for TALA in this study is one such an example. When it was piloted for the first time on its intended target group of Grade 12 students, it performed astonishingly well. When attempted on Grade 10 students, however, TALA did not perform desirably. An appropriate test is therefore a necessity, and recommendable.

What is heartening about the test results obtained for this study is that none of them showed a significant degree of Differential Item Functioning (DIF). In fact, in only one of them a single item showed DIF. In this respect, the TiaPlus analyses of both tests indicated only one flagged item for TALA (Annexure F) and not a single

flagged item for the second test (Annexure G). On the other hand, none of the other results of the other measurement devices for this study have been scrutinised for DIF, or for reliability, or for the productivity of their items. They rely for their reliability on the accumulation provided by the long-term nature of the assessments, which can provide a false sense of consistency, and one that moreover lacks an empirical basis.

Additionally, it might be to one's greater advantage to administer such academic literacy tests at a later stage. Although my reasons for choosing Grade 10 students are well motivated, using Grade 11 students or Grade 12 students might prove to be more meaningful, since they are often more motivated to perform well academically. On the other hand, however, administering such tests at an early stage could identify at risk students beforehand, which would leave some time to design appropriate courses or adjust current instruction (in line with the syllabus requirements), or to enrol such students in support programmes. This matter is therefore one worth investigating further.

Fourthly, I would recommend that one should not disregard the idea of combining different measurement devices. Even though current matriculation results might often be deemed unreliable, they cannot be entirely unuseful. As seen in this study, the English Home Language examination marks of the Grade 10 students predicted the academic average of the students more accurately than the academic literacy tests. Additionally, however, when the English Home Language marks were combined with the results of the second test, an even better prediction of the academic average of the students who participated in this study was obtained. One could therefore argue for the effectiveness of combining more than one measurement device in a study like this.

Moreover, it would possibly be advantageous to do separate regression analyses that compare marks per various school subjects and their relation to average school

performance. This could shed more light on the possible homogeneity of the results of school subjects, which would explain, in turn, why one measurement device (English Home Language) in this study predicts average performance better than a more robust assessment, albeit one that is external to the (internal and familiar) school assessment system.

Lastly, it could also be beneficial to prepare the test takers of academic literacy tests more thoroughly. Although the tests were scanned together with the test takers, and the test administrators tried to keep the test takers calm at all times, it should be considered that more time may be needed before the administration of the test, in order to prepare the test takers more comprehensively. By listening to a more comprehensive explanation of the test, test takers may perhaps have performed better.

#### **7.4 Limitations of the study**

As it is with most studies, foreseen and unforeseen limitations reveal themselves during the course of the investigation. I shall discuss in more detail below some of the limitations of this particular study.

The first limitation of a study like this is the accessibility of test takers, as well as their school marks. Not only did it take diligent planning to ensure that I could administer both tests to two different Bloemfontein schools on two separate occasions, but I also tried to ensure that all the test takers wrote both tests in order to obtain a substantial number of results, to enhance the credibility of the research. Nonetheless, recruiting a greater number of test takers from more schools may prove to be problematic. I was fortunate enough that both the schools I chose were willing to participate in this study. It might happen that schools are not willing to participate in these types of studies, since often controversial school marks are used in the analyses. Another limitation to take into account is the geographical spread

of current schools; a diversity of schools in this and in other areas might have added credibility to the study.

The type of analyses that were carried out on the school marks is another limitation of this study. Even though the schools which participated in this study were generous in providing the English Home Language mark of the students, as well as the students' average mark for the June examinations and the students' average mark excluding the English Home Language mark, some other limitations are obvious. For instance, to investigate why the English Home Language mark predicted the average mark the most accurately, it would have been interesting to examine in detail some of the marks that the students obtained for other subjects. This, however, goes beyond the scope of the current study. Sebolai (2015:19) argues for the importance of developing and designing tests (and also entire programmes, courses or subjects) according to a specific construct, in order to ensure that a test truly measures what it is supposed to measure and consequently, to obtain credible, useful results. He observes, furthermore, that a possible misalignment is present between the school curriculum, the teaching, and the assessments which are used in the curriculum (Sebolai 2015:19). In other words, even though the English Home Language mark is the more accurate predictor of academic potential for this study, the results of Sebolai's investigation, to be carried out at exactly the right level (first year), may well turn out to be dissimilar, and therefore indicate the opposite. As soon as these results are available, they may deserve further consideration.

Another, more obvious limitation of this study is the question of what counts as "Home Language". Using a Home Language other than English might have delivered different results, just as using different schools or ages might also have delivered different results. This means that this particular study can be attempted again for different languages, that might improve and enrich the understanding we have of the problem.



Lastly, 60-mark assessments might possibly have been too short, especially when compared to the 300-mark examinations which were used for the analyses. Administratively, 60-mark assessments were the ideal choice since they could be administered in more or less an hour's time. When the results of the analyses became known, it was apparent that 60-mark assessments might have been too short, even if they were theoretically more reliable. Longer assessments might have a greater chance to compete with the school's entire assessment system.

In the following, and final, section, I shall return to a consideration of which further enquiries may be indicated by the conclusions to this investigation.

### **7.5 Further investigations**

A few final points of interest will now be discussed in further detail.

Firstly, one might wish to develop a survey for the test takers to complete after they have written the tests. The feedback from such surveys can be significant in disclosing the opinions and experiences of the test takers. This, in turn, can be used to improve the administrative procedure of the tests or even, in some cases, the test itself. Especially for a study like this, which can be replicated, such surveys could shed light on some problem areas.

Additionally, it might also be advantageous to work more closely with the teachers of the test takers, if, of course, they are willing to do so. We often forget that teachers only facilitate the school curriculum, they do not develop or in some cases even agree with the specifications of the curriculum implemented. It is also imperative to remember that investigators are outsiders to the school curriculum and the test takers. On the other hand, a teacher is an insider, and might have foreseen some of the problems I did not foresee. Moreover, it might also be interesting, at least, to include teachers' opinions in a study like this. Earlier on I

mentioned that a teacher helped me with the reduction of the second test, and I now feel that I should have included her and others in more aspects of this study.

As mentioned earlier in section 7.3, TALA should rather be administered to its intended target group of Grade 12 students, instead of Grade 10 students. At the same time, however, I have also learnt that using Grade 12 students for such a study does not always prove possible, since Grade 12 students have a much stricter academic schedule to keep to. A study by Sebolai (2015) intends to seek solutions to overcome the problem of measuring too early, which will hopefully help to anticipate the exact problem I have encountered.

Lastly, because the second test performed above expectation, it would probably be beneficial to employ the refinement suggested in the previous chapter and administer the test again. The analyses done on the results of the test indicated that the test has potential, but it would be ideal (if logistically possible) to have a longer than 60-mark assessment as well.

## **7.6 Conclusion**

This study has served to enrich our understanding of the relationship between language ability and academic performance. As awareness grows of the role of language in academic performance at levels lower than tertiary education, that understanding will no doubt continue to grow. I hope that this study may have served to stimulate our growing understanding of these issues so that we are able to identify at an earlier stage those most in need of further language development.



## Bibliography

- Assessment Systems Corporation. 2006. *User's manual for IteMan 3.6 conventional item analysis program*. St Paul, Minnesota: Assessment Systems Corporation.
- Bachman, L. & Palmer, A. 1996. *Language testing in practice*. Oxford: Oxford University Press.
- Blanton, L. 1994. Discourse, artefacts, and the Ozarks: Understanding academic literacy. *Journal of second language writing* 3(1):219-235.
- Brindley, G. 2002. Issues in language assessment. In: R.B. Kaplan (ed.). *The Oxford handbook of applied linguistics*. Oxford: Oxford University Press. P. 459-470.
- Carroll, J. 1961. Fundamental considerations in testing for English proficiency of foreign students. *Testing the English proficiency of foreign students*. Washington: Centre for Applied Linguistics. P. 31-40.
- Chapelle, C. 2011. Validity argument for language assessment: The framework is simple. *Language testing* 29(1):19-27.
- Chatterjee, S. & Price, B. 1991. *Regression analysis by example*. Hoboken, New Jersey: John Wiley & Sons, Inc.
- CITO. 2005. *TiaPlus: Classical test and item analysis*. Arnhem: Cito Measurement and Research Department.
- CITO. 2007. *TiaPlus: Test and item analysis*. Arnhem: Cito Measurement and Research Department.
- Cliff, A., Hanslo, M. & Yeld, N. 2003. Assessing the academic literacy skills of entry-level students using the Placement Test in English for Educational Purposes (PTEEP). Paper read at the Bi-annual conference of the European Association for Research in Learning and Instruction (EARLI). Padova, Italy.
- Cliff, A. & Hanslo, M. 2005. The use of 'alternate' assessments as contributors to processes for selecting applicants to Health Sciences faculties. *Southern African linguistics and applied language studies* 24(1):1-19.
- Department of Basic Education. 2001. *National plan for higher education in South Africa*. Pretoria: Department of Basic Education.

Department of Basic Education. 2005. *Student enrolment planning in public higher education*. Pretoria: Department of Basic Education.

Department of Basic Education. 2011. *Curriculum and assessment policy statement: Grades 10-12 English Home Language*. Pretoria: Department of Basic Education.

Du Plessis, C. 2012. *The design, refinement and reception of a test of academic literacy for postgraduate students*. Unpublished MA dissertation. Bloemfontein: University of the Free State.

Du Plessis, C. 2014a. Issues of validity and generalisability in the grade 12 English Home Language examination. *Per linguam* 30(2):1-19.

Du Plessis, C. 2014b. Writing as construct in the Grade 12 Home Language curriculum and examination. *Journal for language teaching* 48(2):121-141. [Online]. Available: <http://dx.doi.org/10.4314/jlt.v48i42.6>. Accessed: 12 April 2015.

Du Plessis, C., Steyn S. & Weideman, A. 2014. Towards a construct for assessing high level language ability in the South African National Certificate. Submitted to *Africa education review*.

Erasmus, M. 2014. Comparing of two academic literacy tests, 'TALA' and 'Gadgets and freaky inventions', to determine which one would be most suitable to identify students at risk at Grade 10 level. Honours long essay. Bloemfontein: University of the Free State.

Fulcher, G. 2010. *Practical language testing*. London: Hodder Education.

Green, A. 2014. *Exploring language assessment and testing*. New York: Routledge.

Grühn, S. 2015. Initial validation of a test of emergent literacy. Unpublished MA dissertation. Groningen: Rijksuniversiteit.

Guyer, R. & Thompson, N. 2011. *User's manual for IteMan 4.2*. St Paul, Minnesota: Assessment Systems Corporation.

Inter-Institutional Centre for Language Development and Assessment (ICELDA). 2014. [Online]. Available: <http://icelda.sun.ac.za/index.php/sample-tests>. Accessed: 22 April 2014.

Jeffery, A. 2014. Goeie en slegte nuus van 20 jaar. *Rapport*. 25 April. P.7.

- Joubert, J. 2014. African languages flagged. *The Sunday Times*. 3 August. P.2.
- Kane, M. 2006. Validation. In: R Brennen (ed). *Educational measurement*, 4<sup>th</sup> ed. Westport: Greenwood. P. 17-64.
- Lado, R. 1961. *Language testing*. New York: McGraw Hill.
- McNamara, T. 2005. Second language testing and assessment. In: E Hinkel (ed). *Handbook of research in second language teaching and learning*. London: Lawrence Erlbaum Associates. P. 775-778.
- McNamara, T. & Shohamy, E. 2008. Language tests and human rights. *International journal of applied linguistics* 18(1):89-95.
- Messick, S. 1980. Test validity and the ethics of assessment. *American psychologist* 35(11):1012-1027.
- National Benchmark Tests Project. 2013. [Online]. Available: <http://www.nbt.ac.za/>. Accessed: 16 March 2014.
- Oller, J. 1979. *Language tests at school*. London: Longman Group Limited.
- Parkinson, J. 2000. Acquiring scientific literacy through content and genre: A theme-based course for science students. *English for specific purposes* 19(1)369-387.
- 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. [Online]. Available: <http://dx.doi.org/10.4314/jlt.v47i1.5>. Accessed: 18 April 2014.
- Patterson, R. & Weideman, A. 2013b. The refinement of a construct for tests of academic literacy. *Journal for language teaching* 47(1):125-152. [Online]. Available: <http://dx.doi.org/10.4314/jlt.v47i1.6>. Accessed: 19 April 2014.
- Prins, L. 2014. Opdrag oor punte glo gegee. *Die Volksblad*. 11 August. P.11.
- Rambiritch, A. 2012. *Transparency, accessibility and accountability as regulative conditions for a postgraduate test of academic literacy*. Unpublished doctoral dissertation. Bloemfontein: University of the Free State.
- Read, J. 2012. Issues in post-entry language assessment in English-medium universities. *Language teaching* 46(4):1-18.
- Sebolai, K. 2015. *The incremental validity of three tests of academic literacy in the context of a South African university of technology*. MS. PhD thesis. Bloemfontein: University of the Free State.

Shohamy, E. 1994. The validity of direct versus semi-direct oral tests. *Language testing* 11(2):99-123.

Shohamy, E. 1997. Testing methods, testing consequences: are they ethical? Are they fair? *Language testing* 14(3):340-349.

SouthAfrica.info. 2014. [Online]. Available:  
<http://www.southafrica.info/about/education/education.htm#.U2JywldFGuk>.  
Accessed: 1 May 2014.

Statistical Consultation Unit. 2014. Bloemfontein: University of the Free State.

Steyn, S. 2010. Devising a test to aid in the process of preparing prospective students for academic language proficiency tests at tertiary institutions. Unpublished report for ICELDA.

Steyn, S. 2014. *The design and refinement of a test of early academic literacy*. MA thesis. Groningen: Rijksuniversiteit.

Steyn, S. 2015. *A theoretical justification for the design and refinement of a test of advanced language ability for learners at FET level*. In preparation: MA dissertation. Bloemfontein: University of the Free State.

University of Pretoria (UP). 2014. *Calendar of the Faculty of Humanities*. Pretoria: University of Pretoria.

Van der Slik, F. & Weideman, A. 2005. The refinement of a test of academic literacy. *Per linguam* 21(1):23-35.

Van der Walt, J. & Steyn, H. 2007. Pragmatic validation of a test of academic literacy at tertiary level. *Ensovoort* 11(2):138-153.

Van Dyk, T. & Weideman, A. 2004a. Switching constructs: On the selection of an appropriate blueprint for academic literacy assessment. *Journal for language teaching* 38(1):1-13.

Van Dyk, T. & Weideman, A. 2004b. Finding the right measure: From blueprint to specification to item type. *Journal for language teaching* 38(1):15-24.

Van Rensburg, C. & Weideman, A. 2002. Language proficiency: Current strategies, future remedies. *Journal for language teaching* 36(1):152-164.

Van Rooy, B. & Coetzee-Van Rooy, S. 2015. The language issue and academic performance at a South African University. Forthcoming in *Southern African linguistics and applied language studies*.

- Weideman, A. 2006. Overlapping and divergent agendas: writing and applied linguistics research. In: C van der Walt (ed.). *Living through languages: An African tribute*. Stellenbosch: African Sun Media. P. 147-164.
- Weideman, A. & Van der Slik, F. 2008. The stability of test design: Measuring difference in performance across several administrations of a test of academic literacy. *Acta academia* 40(1):161-182.
- Weideman, A. 2009. *Beyond expression: a systematic study of the foundations of linguistics*. Grand Rapids: The Reformational Publishing Project.
- Weideman, A. 2011. Academic literacy tests: Design, development, piloting and refinement. *Journal for language teaching* 45(2):100-113. [Online]. Available: <http://dx.doi.org/10.4314/jlt.v45i2.6>. Accessed: 18 April 2014.
- Weideman, A. 2012. Validation and validity beyond Messick. *Per linguam*. 28(2):1-14.
- Weideman, A. 2013a. Academic literacy interventions: What are we not yet doing, or not yet doing right? *Journal for language teaching* 47(2):11-24.
- Weideman, A. 2013b. Applied linguistics beyond postmodernism. *Acta academia* 45(4):236-255.
- Weideman, A. 2014. Innovation and reciprocity in applied linguistics. *Literator*. [Online]. Available: <http://dx.doi.org/10.4102/lit.v35il.1074>. Accessed: 22 April 2014.
- Weideman, A. 2015. *Responsible design: The foundations of applied linguistics*. In preparation.
- Weideman, A., Du Plessis, C. & Steyn, S. 2014. Diversity, variation and fairness: equivalence in national level language assessments. Paper prepared for presentation at the Language, Education and Diversity conference, Auckland, November 2015.
- Weideman, A., Patterson, R. & Pot, A. 2014. Construct refinement in tests of academic literacy. Paper read at AAAL 2014 Colloquium on post-entry language assessment in universities worldwide. Forthcoming in J. Read (ed.). *Post-admission language assessment in universities: International perspectives*. Forthcoming from Springer.
- Weideman, A. & Van Dyk, T. (eds). 2014. *Academic literacy: Test your competence*. Centre for Language Development and Assessment.



Weir, C. 2005. *Language testing and validation: An evidence-based approach*.  
Hampshire: Palgrave-Macmillan.



**Annexure A**  
**The second test**

## Gadgets and freaky inventions

### Scrambled text

The sequence of the sentences in the following has been altered. Say what the correct order is by marking your choice on the loose answer sheet.

#### Why patent your invention?

- A. The reason to patent your invention is to give the owner (you) the exclusive right to commercially exploit the invention for the life of the patent (usually 20 years).
- B. So you've invented the next big something, it's revolutionary, it's cheap to make and everyone will want one so you want to put it on the market.
- C. When you have finally patented your new masterpiece, you will be the only lucky one that will be able to benefit from it.
- D. But wait a minute, before you can start the production lines rolling, you've got to protect it from someone else who thinks your idea will make them a fortune too.
- E. The first thing you need to do before you start dreaming about the millions you can make, you have to protect it by patenting it.

[Adapted from <http://www.abc.net.au/tv/newinventors/txt/s1097642.htm>, *How to patent your invention*. Accessed 17 August 2010]

- |   |          |          |          |          |          |
|---|----------|----------|----------|----------|----------|
| 1. Which sentence did you put <b>first</b> ?  | <b>A</b> | <b>B</b> | <b>C</b> | <b>D</b> | <b>E</b> |
| 2. Which sentence did you put <b>second</b> ? | <b>A</b> | <b>B</b> | <b>C</b> | <b>D</b> | <b>E</b> |
| 3. Which sentence did you put <b>third</b> ?  | <b>A</b> | <b>B</b> | <b>C</b> | <b>D</b> | <b>E</b> |
| 4. Which sentence did you put <b>fourth</b> ? | <b>A</b> | <b>B</b> | <b>C</b> | <b>D</b> | <b>E</b> |
| 5. Which sentence did you put <b>fifth</b> ?  | <b>A</b> | <b>B</b> | <b>C</b> | <b>D</b> | <b>E</b> |

[5]

### Vocabulary knowledge

Choose the best possible answer from the list of options:

- 6. To patent an invention is a great \_\_\_\_\_ for any inventing genius.
  - A. bereavement
  - B. achievement
  - C. endorsement
  - D. inducement
  
- 7. In order for any inventor to make a success of an invention he/she has to be \_\_\_\_\_ to the process.
  - A. addicted
  - B. connected
  - C. committed
  - D. indented

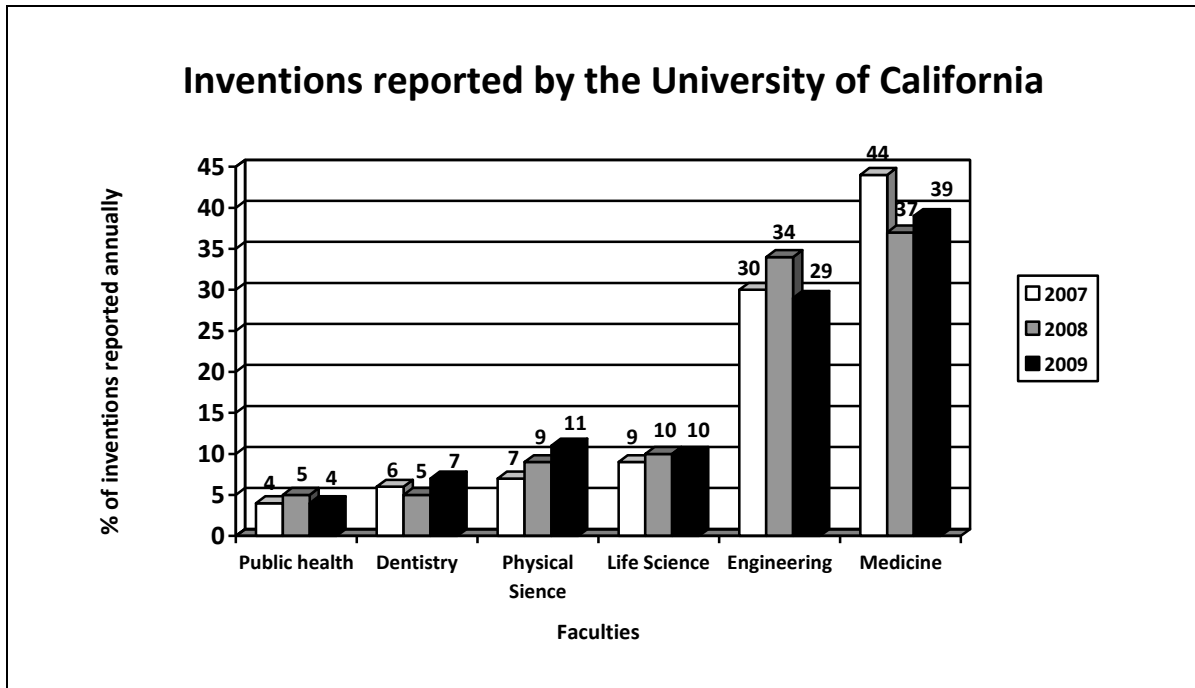
8. Gadgets and inventions often \_\_\_\_\_ an inventor's passion for the art of inventing.
- A. distract
  - B. dissolve
  - C. display
  - D. dissipate
9. Inventors often \_\_\_\_\_ an entire lifetime to complete one single invention.
- A. involve
  - B. invade
  - C. invoke
  - D. invest
10. While some gadgets \_\_\_\_\_ an almost instantaneous buzz of excitement, others take a while to catch on.
- A. generate
  - B. operate
  - C. aggregate
  - D. integrate
11. Many gadgets are invented in order to be able to \_\_\_\_\_ technology with comfort.
- A. intensify
  - B. interfere
  - C. integrate
  - D. interject
12. The ability to \_\_\_\_\_ a simple idea into a mind-blowing invention is an art that only truly genius inventors are able to perfect.
- A. inform
  - B. transform
  - C. conform
  - D. deform
13. One invention often \_\_\_\_\_ the introduction of another, which results in technology being pushed even further.
- A. initiates
  - B. mediates
  - C. arrogates
  - D. rotates
14. It is often the case that inventors change and adapt their gadgets or inventions in order to \_\_\_\_\_ their potential and abilities.
- A. compromise
  - B. internalise
  - C. familiarise
  - D. maximise

15. Some gadgets do not take off as well as others because the inventors do not \_\_\_\_\_ their opportunities as they should.
- utilize
  - memorise
  - analyse
  - visualise

[10]

### Interpreting graphs and visual information

Study the following graph, that summarises the annual invention reports of the University of California from 2008-2010, before you answer the questions below. The percentage of inventions refers to the proportion of the contribution of inventions of each faculty to the UCLA total.



[Adapted from [www.researchucla.edu](http://www.researchucla.edu). Accessed 16 August 2010. Some data altered/fictitious]

16. The Faculties of Public Health, Dentistry, Physical Science and Life Science together reported the same % inventions in 2008 as the faculty of \_\_\_\_\_ reported in 2009 alone.
- Physical Science
  - Life Science
  - Engineering
  - Medicine
17. The Faculty of Medicine reported the same % of inventions in 2007 as which of the following pairs of faculties reported together in 2008?
- Public Health and Life Science
  - Medicine and Engineering
  - Life Science and Engineering
  - Engineering and Public Health

18. The only faculty with a steady increase over all three years is \_\_\_\_\_ .
- A. Public Health.
  - B. Life Science.
  - C. Physical Science.
  - D. Medicine.
19. If this steady increase mentioned above continues, the forecast for that faculty for 2010 would be \_\_\_\_\_
- A. 10%.
  - B. 13%.
  - C. 16%.
  - D. 19%.
20. The only faculty that followed the same pattern as the Faculty of Medicine is \_\_\_\_\_.
- A. Public Health
  - B. Dentistry
  - C. Physical Science
  - D. Life Science
21. In which year was the Faculty of Life Science's % of reports double the % of reports of the Faculty of Dentistry?
- A. 2007
  - B. 2008
  - C. 2009
  - D. 2010
22. In which year did the Faculties Dentistry and Life Science together have 50% of the % of reports of the Faculty of Engineering?
- A. 2007
  - B. 2008
  - C. 2009
  - D. 2010
23. During 2007 the Faculty of Medicine reported \_\_\_\_\_ times more inventions than the Faculty of Public Health.
- A. 7
  - B. 9
  - C. 11
  - D. 13

[8]

## Text comprehension

Read the text below, then answer the questions that follow.

# SixthSense blurs digital and the real

## How SixthSense works

**L**APTOPS AND SMARTPHONES ALLOW easy access computing power, but researchers at the Massachusetts Institute of Technology want to go one step further by turning the entire world into a computer.

**2** At this year's (2008) Computer-Human Interaction (CHI) conference in Boston, the Fluid Interfaces Group at MIT's Media Lab unveiled the latest prototype of SixthSense, a wearable, gesture-driven computing platform that can continually augment the physical world with digital information.

**3** Imagine being able to check your email on any blank wall, simply by drawing an @ sign in the air with your finger, or being able to check the time by using that same finger to draw a circle, which produces the image of an analogue watch right on your wrist. You want to take a digital photograph? Just put your thumbs and forefingers together to make a picture frame. Better yet, imagine a system that can display the reason for your flight delay directly on the boarding paare holding in your hand.

**4** "We're trying to make it possible to have access to relevant information in a more seamless way," says Dr Pattie Maes, who heads the Fluid Interfaces Group at MIT. "We have a vision of a computing system that understands, at least to some extent, where the user is, what the user is doing, and who the user is interacting with," says Dr. Maes.

**5** The SixthSense prototype has changed since it was first introduced to the public last year. Originally, it consisted of a web camera strapped to a bicycle helmet. But the current prototype promises to be a bit more consumer friendly. It consists of a small camera-projector combination (about the size of a cigarette pack) worn around the neck of the user. An accompanying smartphone runs the SixthSense software, and handles the connection to the internet.

**6** "You can turn any surface around you into an interactive surface," says Pranav Mistry, an MIT graduate student working on the SixthSense project. "Let's say I'm in a bookstore, and I'm holding a book. SixthSense will recognize that, and will go up to Amazon. Then, it will display online reviews of that book, and prices, right on the cover of the book I'm holding." Mistry notes that the system is customisable as well, so that if you don't want Amazon reviews of a book, you could choose instead to find out what the *New York Times* thinks of it.

**7** The hardware included in the SixthSense system is not that expensive. The current prototype costs about \$350 to build. But this attempt to merge the digital world with the physical world requires some serious programming and engineering. "All the work is in the software," says Dr Maes. "The system is constantly trying to figure out what's around you, and what you're trying to do. It has to recognize the images you see, track your gestures, and then relate it all to relevant information at the same time."

**8** Pranav Mistry sees some commercial applications for the system in the near future. For example, he wants to develop a sign language application that would "speak out" a translation while someone was signing.

**9** And if SixthSense catches on, what will we all make of the sight of dozens of people checking their e-mails on the walls of airports and train stations? Dr. Pattie Maes laughs: "Well, I think it might actually be more socially acceptable than those Bluetooth earpieces people use these days. At least with our system you can actually see that people are interacting with information, instead of watching someone that looks like they're just talking to themselves on a street."

[Adapted from [www.InventHelp.com](http://www.InventHelp.com). Accessed August 16, 2010]



24. From the title one can conclude that SixthSense
- A. makes the digital world less clear.
  - B. makes reality seem blurry and unclear.
  - C. merges the digital world and reality.
  - D. blanks the digital world out of reality. (1)
25. The phrase “turning the entire world into a computer” in paragraph 1 means that
- A. the earth as a whole works as a computer.
  - B. everything can form part of this computer.
  - C. the earth is controlled by a computer.
  - D. the world is nothing but a computer. (1)
26. The word ‘unveiled’ in paragraph 2 means that
- A. SixthSense was eventually shown in public.
  - B. a veil was removed at the conference.
  - C. there was a bride at the conference.
  - D. a protecting veil was draped over SixthSense. (1)
27. A good description of what SixthSense does, is given in paragraph \_\_\_\_
- A. 1
  - B. 2
  - C. 3
  - D. 4 (1)
28. The word ‘augment’ in paragraph 2 means to \_\_\_\_\_of the physical world
- A. adapt to another form
  - B. change the perception
  - C. change the reality
  - D. increase the quality (1)
29. The phrase “gesture driven” in paragraph 2 is best supported and explained by paragraph \_\_\_\_
- A. 1.
  - B. 2.
  - C. 3.
  - D. 4. (1)
30. SixthSense has a number of functions that can make day-to-day life easier. Which of the following is the odd one out?
- A. Checking emails anytime, anywhere.
  - B. Capturing memories without a camera.
  - C. Helping you understand why your flight is late.
  - D. Turning the entire world into a computer. (1)
31. The word ‘that’ in the first sentence of paragraph 3 refers to
- A. “your email”
  - B. “an @ sign”
  - C. “your finger”
  - D. “check the time” (1)

32. The phrase “to have access to relevant information” in paragraph 4 is related to the phrase
- A. “easy access computing” in paragraph 1.
  - B. “go one step further” in paragraph 1.
  - C. “gesture driven computing” in paragraph 2.
  - D. “a system that can display” in paragraph 3.
- (1)
33. The word ‘vision’ in paragraph 4 means
- A. they use their eyes to see the system.
  - B. it is at this stage only an idea.
  - C. the idea came to them in a dream.
  - D. it is a picture on a television screen.
- (1)
34. One can conclude from paragraph \_\_\_\_\_ that SixthSense is a project that the Fluid Interfaces Group has been working on for quite some time.
- A. 5
  - B. 4
  - C. 3
  - D. 2
- (1)
35. The phrase “last year” in paragraph 5 refers to the year
- A. 2010
  - B. 2009
  - C. 2008
  - D. 2007
- (1)
36. The original prototype was changed because it
- A. didn’t work.
  - B. wasn’t user-friendly.
  - C. was too complicated.
  - D. was insufficient.
- (1)
37. The software of SixthSense works through
- A. the internet.
  - B. a camera projector.
  - C. a smartphone.
  - D. a web camera.
- (1)
38. Pranav Mistry is
- A. the inventor of SixthSense.
  - B. the developer of the software.
  - C. a professor at MIT.
  - D. one of a team working on SixthSense.
- (1)
39. The word ‘it’ in the third sentence of paragraph 6 refers to
- A. MIT.
  - B. SixthSense.
  - C. the software.
  - D. a book.
- (1)

40. The word ‘notes’ in paragraph 6 means that Mistry
- A. is taking down notes.
  - B. is writing himself a reminder.
  - C. gives written acknowledgement.
  - D. mentions extra information. (1)
41. The word ‘customisable’ in paragraph 6 means
- A. recognisable
  - B. interactive
  - C. adaptable
  - D. selective (1)
42. SixthSense consists of various components. Which of the following is the most inexpensive component? The
- A. system.
  - B. prototype.
  - C. software.
  - D. hardware. (1)
43. The word ‘It’ in the last sentence of paragraph 7 refers to the
- A. software.
  - B. system.
  - C. hardware.
  - D. prototype. (1)
44. In the phrase “relate it” in the last sentence of paragraph 7, the word ‘it’ refers to
- A. the way the prototype works in a general sense.
  - B. how the images of the user is recognised.
  - C. the tracked gestures and images of the user.
  - D. the way the software relates things to information. (1)
45. From the first sentence of paragraph 8 one can conclude that
- A. other practical functions could well be added to SixthSense.
  - B. SixthSense could be sold commercially when it has been completed.
  - C. SixthSense would only be worthwhile once it is sold commercially.
  - D. if any additions were made to SixthSense, they would be able to sell it. (1)
46. The word ‘signing’ in paragraph 8 refers to
- A. someone signing autographs.
  - B. a contract being signed.
  - C. the language: Sign language.
  - D. a message board signalling something. (1)
47. The word ‘it’ in the second sentence of paragraph 9 refers to
- A. SixthSense catching on and being successful.
  - B. the sight of people checking their emails on walls.
  - C. the sign language application that is being developed.
  - D. The Bluetooth earpieces people are currently using. (1)

48. The phrase “our system” in paragraph 9 refers to
- A. the software.
  - B. the hardware.
  - C. the smartphones.
  - D. SixthSense.

(1)

[25]

## Grammar and text relations

**In the text below some words have been deleted. First read through the whole text, then answer the questions that follow.**

### The Uno – It’s Unique – but can it pop a Wheelie

The 2008 National Motorcycle Show in Toronto has always been heavily influenced by the American V-twin crowd. It highlights of the area’s top custom builders who have on display a fine array of one-off custom machines.

This year’s, however, had one very unusual one-off custom, the Uno. The orange and grey coloured Uno made first public appearance balanced on its two side-by-side wheels and its footpegs. Looking more like it should have ridden by George Jetson as he pulled up to his space platform, it looked out of place amid the other custom creations in the building.

Operation of the 54.4 kg (120 lb) is simple; in fact it’s so simple there are no controls except for an on-off switch. To go forward you simply push your body weight forward to tilt the machine. To back up, just lean back on the seat to tilt backwards and back it goes. The farther you lean, the faster it accelerates. The gyro

tells the ECU how much to accelerate and that in turn delivers the proper amount of current to the electric motors, one for each wheel.

The independent suspension allows the unit to lean like a motorcycle during a turn. The wheel will then compress the suspension so the wheel moves up inside the body while the outer wheel continues to make contact with the ground. The gyro detects the sideways motion and instructs the ECU accordingly. Since each wheel has its electric motor, the outer wheel speeds up in order to complete the turn.

If the rider is forward and needs to stop, he simply leans back. The electric motors have inherently high torque so stopping is very quick. If you continue to lean backward, the Uno will go backwards. All the while, the Uno feels quite stable. A full battery pack will provide about 3 hours of time and charging time is only 17 minutes if using a fast charger.

[Adapted from Motorcycle Mojo Magazine, May 2008 edition, *The Uno – It’s Unique – but can it pop a Wheelie?* Accessed 16 August 2010]

**In the following texts, you have to indicate the possible place where a word may have been deleted, and which word belongs there. Here are two examples:**

The 2008 National Motorcycle Show in Toronto has always been heavily influenced by the American V-twin crowd. It highlights  of the  area’s top  custom builders who  have on display a fine array of one-off custom machines.

This  year’s , however,  had one  very unusual one-off custom, the Uno.

Where has the word been deleted?

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

Which word has been left out here?

- A. also
- B. some**
- C. surrounding
- D. professional

Where has the word been deleted?

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

Which word has been left out here?

- A. show**
- B. constantly
- C. quirky
- D. random

Here are a following two examples:

The <sup>70&71</sup> [i] orange and [ii] grey coloured Uno made [iii] first public appearance balanced [iv] on its two side-by-side wheels and its footpegs. Looking <sup>72&73</sup> [i] more like it should [ii] have [iii] ridden by George Jetson as he [iv] pulled up to his space platform, it looked out of place amid the other custom creations in the building.

Where has the word been deleted?

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

Which word has been left out here?

- A. its**
- B. bright
- C. only
- D. dull

Where has the word been deleted?

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

Which word has been left out here?

- A. mysteriously
- B. actually
- C. creatively
- D. been**

Operation <sup>74&75</sup> [i] of the 54.4 kg (120 lb) [ii] is simple; [iii] in fact [iv] it's so simple there are no controls except for an on-off switch. To go forward you simply push your body weight forward to tilt the machine. To back up, <sup>76&77</sup> [i] just lean back [ii] on the seat to tilt [iii] backwards and [iv] back it goes. The farther you lean, the faster it accelerates. The gyro tells the ECU how much to accelerate and that in turn delivers the proper amount of current to the electric motors, one for each wheel.

49. Where has the word been deleted?

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

50. Which word has been left out here?

- A. honestly
- B. and
- C. one
- D. machine

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

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

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

- A. it
- B. way
- C. you
- D. completely

The <sup>82&83</sup> [i] wheel will then [ii] compress the [iii] suspension so the [iv] wheel moves up inside the body while the outer wheel continues to make contact with the ground. The gyro <sup>84&85</sup> [i] detects the sideways motion [ii] and instructs the ECU accordingly. Since each wheel [iii] has its [iv] electric motor, the outer wheel speeds up in order to complete the turn.

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

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

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

- A. automatically
- B. only
- C. inner
- D. front

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

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

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

- A. first
- B. automatically
- C. own
- D. generally

If the rider is <sup>86&87</sup> [i] forward and [ii] needs to stop, he simply leans [iii] back. The electric motors [iv] have inherently high torque so stopping is very quick. If you continue to lean backward, the Uno will go backwards. All the while, the Uno <sup>88&89</sup> [i] feels quite stable. A full battery [ii] pack will provide [iii] about 3 hours of [iv] time and charging time is only 17 minutes if using a fast charger.

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

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

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

- A. will
- B. moving
- C. urgently
- D. comfortably

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

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

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

- A. travel
- B. and
- C. nearly
- D. totally

[12]  
**TOTAL: 60**

## **Annexure B**

### **Iteman 3.6 analysis of TALA**

# TALA re-pilot, Eunice & Heidedal (June 2014)

ITEMAN (tm) for 32-bit Windows, Version 3.6  
Copyright (c) 1982 - 1998 by Assessment Systems Corporation

Conventional Item and Test Analysis Program

Item analysis for data from file C:\AlbertDocuments\ICELDA\ice01174.txt  
Date: 28 Aug 2014 Time: 16:00

\*\*\*\*\* ANALYSIS SUMMARY INFORMATION \*\*\*\*\*

Data (Input) File: C:\AlbertDocuments\ICELDA\ice01174.txt  
Analysis Output File: C:\AlbertDocuments\ICELDA\ice01174.out  
Score Output File: NONE  
Exceptions File: NONE  
Statistics Output File: NONE

Scale Definition Codes: DICHOT = Dichotomous MPOINT = Multipoint/Survey

Scale: 1  
-----  
Type of Scale DICHOT  
N of Items 60  
N of Examinees 242

\*\*\*\*\* CONFIGURATION INFORMATION \*\*\*\*\*

Type of Correlations: Point-Biserial

Correction for Spuriousness: YES

Ability Grouping: NO

Subgroup Analysis: NO

Express Endorsements As: PERCENTAGES

Score Group Interval Width: 1



**Subtest 1: Scrambled text**

Seq. No.	Scale -Item	Item Statistics		Alternative Statistics			Key	
		Pcnt Correct	Point Biser.	Alt.	Pcnt Endorsing	Point Biser.		
1	1-1	81	.16	A	1	-.01		
				B	13	-.16		
				C	4	-.19		
				D	81	.16		*
				E	1	-.06		
				Other	1	-.10		
2	1-2	64	-.04	A	7	-.04		
				B	11	-.02		
				C	64	-.04		*
				D	7	-.15		
				E	9	.10		?
				Other	2	-.18		
3	1-3	34	.18	A	23	-.17		
				B	34	.18		*
				C	17	.03		
				D	4	-.17		
				E	19	-.12		
				Other	4	-.20		
4	1-4	41	.16	A	41	.16	*	
				B	11	-.12		
				C	12	-.03		
				D	5	-.07		
				E	29	-.22		
				Other	2	-.02		
5	1-5	35	.19	A	25	-.15		
				B	29	-.11		
				C	3	-.08		
				D	4	-.16		
				E	35	.19		*
				Other	4	-.10		

**Subtest 2: Vocabulary knowledge**

Seq. No.	Scale -Item	Item Statistics		Alternative Statistics			Key	
		Pcnt Correct	Point Biser.	Alt.	Pcnt Endorsing	Point Biser.		
6	1-6	64	.45	A	3	-.14		
				B	64	.45		*
				C	20	-.40		
				D	12	-.26		
				Other	1	-.03		
7	1-7	57	.32	A	21	-.45		
				B	11	-.03		
				C	10	-.12		
				D	57	.32		*
				Other	2	-.02		

Seq. No.	Scale -Item	Item Statistics		Alternative Statistics			Key
		Pcnt Correct	Point Biser.	Alt.	Pcnt Endorsing	Point Biser.	
8	1-8	91	.19	A	4	-.12	*
				B	3	-.16	
				C	91	.19	
				D	2	-.14	
				Other	0	-.12	
9	1-9	55	.36	A	21	-.35	*
				B	5	-.26	
				C	55	.36	
				D	19	-.14	
				Other	0	.09	
10	1-10	40	.40	A	16	-.10	*
				B	30	-.30	
				C	40	.40	
				D	13	-.27	
				Other	1	-.08	
11	1-11	71	.24	A	71	.24	*
				B	20	-.17	
				C	3	-.24	
				D	5	-.15	
				Other	2	-.16	
12	1-12	28	.32	A	31	.06	*
				B	6	-.13	
				C	28	.32	
				D	33	-.41	
				Other	1	-.19	
13	1-13	29	.10	A	11	-.43	*
				B	29	.10	
				C	48	.17	
				D	12	-.19	
				Other	1	-.14	
CHECK THE KEY B was specified, C works better							?
14	1-14	37	.09	A	21	-.07	*
				B	36	-.10	
				C	4	-.20	
				D	37	.09	
				Other	1	-.11	
15	1-15	60	.25	A	60	.25	*
				B	2	-.06	
				C	19	-.27	
				D	18	-.16	
				Other	1	-.11	

**Subtest 3: Interpreting graphs and visual information**

Seq. No.	Scale -Item	Item Statistics		Alternative Statistics			
		Pcnt Correct	Point Biser.	Alt.	Pcnt Endorsing	Point Biser.	Key
16	1-16	30	-.01	A	30	-.01	*
				B	36	.07	?
				C	8	-.10	
				D	24	-.19	
				Other	2	-.07	
		CHECK THE KEY A was specified, B works better					
17	1-17	32	.28	A	16	-.34	
				B	22	.14	
				C	29	-.32	
				D	32	.28	*
				Other	0	-.02	
18	1-18	27	.32	A	25	-.16	
				B	33	-.20	
				C	14	-.18	
				D	27	.32	*
				Other	0		
19	1-19	18	.12	A	18	.12	*
				B	47	-.09	
				C	27	-.17	
				D	7	.01	
				Other	0	-.11	
20	1-20	49	.34	A	20	-.28	
				B	24	-.24	
				C	49	.34	*
				D	6	-.13	
				Other	1	.00	
21	1-21	33	.11	A	36	-.13	
				B	15	-.07	
				C	33	.11	*
				D	14	-.14	
				Other	2	-.02	
22	1-22	57	.24	A	14	-.00	
				B	57	.24	*
				C	18	-.30	
				D	12	-.23	
				Other	0		
23	1-23	31	.13	A	15	-.16	
				B	17	-.14	
				C	36	-.09	
				D	31	.13	*
				Other	2	.04	

**Subtest 4: Text comprehension**

Seq. No.	Scale -Item	Item Statistics		Alternative Statistics					
		Pcnt Correct	Point Biser.	Alt.	Pcnt Endorsing	Point Biser.	Key		
24	1-24	40	.30	A	26	-.08			
				B	40	.30	*		
				C	10	-.15			
				D	23	-.35			
				Other	0	-.03			
25	1-25	32	.02	A	7	-.22			
				B	32	.02	*		
				CHECK THE KEY		C	47	.05	?
				B was specified, C works better		D	14	-.15	
				Other	1	-.05			
26	1-26	48	.26	A	48	.26	*		
				B	18	-.21			
				C	22	-.16			
				D	12	-.20			
				Other	0	-.00			
27	1-27	61	.28	A	19	-.31			
				B	61	.28	*		
				C	12	-.18			
				D	7	-.10			
				Other	1	.05			
28	1-28	38	.27	A	18	-.32			
				B	35	-.11			
				C	8	-.14			
				D	38	.27	*		
				Other	1	-.01			
29	1-29	31	.03	A	31	.03	*		
				B	22	-.29			
				C	31	-.00			
				D	16	.02			
				Other	0	.19			
30	1-30	24	.24	A	35	-.20			
				B	29	-.19			
				C	24	.24	*		
				D	11	-.05			
				Other	1	.03			
31	1-31	29	.41	A	37	-.33			
				B	14	-.23			
				C	19	-.03			
				D	29	.41	*		
				Other	1	-.15			
32	1-32	76	.35	A	3	-.21			
				B	76	.35	*		
				C	7	-.23			
				D	12	-.23			
				Other	1	-.21			
33	1-33	19	.41	A	19	.41	*		
				B	23	-.13			

Seq. No.	Scale -Item	Item Statistics		Alternative Statistics			Key
		Pcnt Correct	Point Biser.	Alt.	Pcnt Endorsing	Point Biser.	
				C	32	-.38	
				D	23	.04	
				Other	3	-.18	
34	1-34	73	.35	A	11	-.23	
				B	8	-.27	
				C	7	-.18	
				D	73	.35	*
				Other	1	-.15	
35	1-35	36	.17	A	30	-.04	
				B	15	-.02	
				C	17	-.32	
				D	36	.17	*
				Other	2	-.16	
36	1-36	29	.17	A	19	-.27	
				B	29	.17	*
				C	16	-.11	
				D	33	.01	
				Other	2	-.18	
37	1-37	16	.16	A	46	-.03	
				B	20	-.15	
				C	16	-.16	
				D	16	.16	*
				Other	2	-.07	
38	1-38	33	.06	A	33	.06	*
				B	26	-.14	
				C	15	-.17	
				D	24	.04	
				Other	2	-.12	
39	1-39	24	.27	A	32	-.31	
				B	29	-.03	
				C	14	-.09	
				D	24	.27	*
				Other	1	-.15	
40	1-40	28	.43	A	45	-.26	
				B	17	-.23	
				C	10	-.13	
				D	28	.43	*
				Other	1	-.17	
41	1-41	54	.51	A	20	-.29	
				B	12	-.30	
				C	12	-.25	
				D	54	.51	*
				Other	2	-.20	
42	1-42	28	.16	A	18	-.23	
				B	45	-.02	
				C	28	.16	*
				D	8	-.14	
				Other	1	-.19	

Seq. No.	Scale -Item	Item Statistics		Alternative Statistics			
		Pcnt Correct	Point Biser.	Alt.	Pcnt Endorsing	Point Biser.	Key
43	1-43	46	.22	A	29	-.08	
				B	46	.22	*
				C	5	-.17	
				D	19	-.26	
				Other	1	-.13	
44	1-44	55	.30	A	11	-.09	
				B	55	.30	*
				C	7	-.30	
				D	25	-.23	
				Other	1	-.14	
45	1-45	39	.20	A	29	-.25	
				B	39	.20	*
				C	17	-.16	
				D	13	.02	
				Other	2	-.13	
46	1-46	54	.45	A	54	.45	*
				B	39	-.51	
				C	5	-.06	
				D	1	-.13	
				Other	2	-.05	
47	1-47	25	.27	A	25	.27	*
				B	52	-.24	
				C	9	-.14	
				D	12	-.08	
				Other	1	-.06	
48	1-48	56	.33	A	14	-.16	
				B	11	-.22	
				C	18	-.27	
				D	56	.33	*
				Other	1	-.06	

-----

**Subtest 5: Grammar and text relations**

Seq. No.	Scale -Item	Item Statistics		Alternative Statistics			
		Pcnt Correct	Point Biser.	Alt.	Pcnt Endorsing	Point Biser.	Key
49	1-49	19	.06	A	44	.11	?
				B	21	-.26	
				C	19	.06	*
				D	12	-.08	
				Other	3	-.18	
CHECK THE KEY C was specified, A works better							
50	1-50	12	.18	A	19	-.09	
				B	37	-.03	
				C	30	-.16	
				D	12	.18	*
				Other	2	-.15	

Seq. No.	Scale -Item	Item Statistics		Alternative Statistics			Key
		Pcnt Correct	Point Biser.	Alt.	Pcnt Endorsing	Point Biser.	
51	1-51	49	.23	A	22	-.21	*
				B	49	.23	
				C	11	-.19	
				D	17	-.09	
				Other	2	-.11	
52	1-52	62	.33	A	15	-.19	*
				B	62	.33	
				C	15	-.30	
				D	5	-.10	
				Other	2	-.17	
53	1-53	61	.38	A	13	-.24	*
				B	14	-.26	
				C	61	.38	
				D	8	-.20	
				Other	4	-.12	
54	1-54	29	.31	A	18	-.30	*
				B	32	-.03	
				C	17	-.20	
				D	29	.31	
				Other	4	-.12	
55	1-55	32	.24	A	17	-.21	*
				B	32	.24	
				C	35	-.09	
				D	10	-.12	
				Other	7	-.18	
56	1-56	41	.12	A	41	.12	*
				B	15	-.11	
				C	21	-.08	
				D	17	-.10	
				Other	6	-.16	
57	1-57	47	.25	A	18	-.15	*
				B	47	.25	
				C	18	-.19	
				D	9	-.13	
				Other	8	-.15	
58	1-58	31	.18	A	31	.18	*
				B	20	.03	
				C	31	-.21	
				D	10	-.15	
				Other	8	-.15	
59	1-59	31	.28	A	31	.28	*
				B	19	-.26	
				C	18	-.15	
				D	24	-.08	
				Other	8	-.11	
60	1-60	40	.21	A	16	-.05	*
				B	18	-.19	
				C	16	-.17	
				D	40	.21	

-----  
ITEMAN (tm) for 32-bit Windows, Version 3.6  
Copyright (c) 1982 - 1998 by Assessment Systems Corporation

Conventional Item and Test Analysis Program

Item analysis for data from file C:\AlbertDocuments\ICELDA\ice01174.txt  
Date: 28 Aug 2014 Time: 16:00

There were 242 examinees in the data file.

Scale Statistics

-----  
Scale: 1  
-----  
N of Items 60  
N of Examinees 242  
Mean 25.112  
Variance 65.678  
Std. Dev. 8.104  
Skew 0.278  
Kurtosis -0.413  
Minimum 6.000  
Maximum 50.000  
Median 24.000  
Alpha 0.818  
SEM 3.458  
Mean Pcnt Corr 42  
Mean Item-Tot. 0.238  
Mean Biserial 0.313

-----



ITEMAN (tm) for 32-bit Windows, Version 3.6  
 Copyright (c) 1982 - 1998 by Assessment Systems Corporation  
 Conventional Item and Test Analysis Program  
 Item analysis for data from file C:\AlbertDocuments\ICELDA\ice01174.txt  
 Date: 28 Aug 2014 Time: 16:00

SCALE # 1

Score Distribution Table

Number Correct	Freq- uency	Cum Freq	PR	PCT	
. . . No examinees below this score . . .					
5	0	0	1	0	+
6	1	1	1	0	
7	0	1	1	0	
8	0	1	1	0	
9	1	2	1	0	
10	2	4	2	1	+##
11	2	6	2	1	##
12	6	12	5	2	###
13	2	14	6	1	##
14	10	24	10	4	####
15	4	28	12	2	+###
16	10	38	16	4	####
17	7	45	19	3	###
18	8	53	22	3	###
19	10	63	26	4	####
20	13	76	31	5	+#####
21	11	87	36	5	#####
22	11	98	40	5	#####
23	16	114	47	7	#####
24	13	127	52	5	#####
25	9	136	56	4	+#####
26	8	144	60	3	###
27	6	150	62	2	##
28	11	161	67	5	#####
29	9	170	70	4	#####
30	4	174	72	2	+###
31	6	180	74	2	##
32	11	191	79	5	#####
33	7	198	82	3	###
34	8	206	85	3	###
35	8	214	88	3	+###
36	6	220	91	2	##
37	9	229	95	4	#####
38	3	232	96	1	#
39	3	235	97	1	#
40	0	235	97	0	+
41	1	236	98	0	
42	1	237	98	0	
43	1	238	98	0	
44	1	239	99	0	
45	1	240	99	0	+
46	1	241	99	0	
47	0	241	99	0	
48	0	241	99	0	
49	0	241	99	0	
50	1	242	99	0	+
51	0	242	99	0	
52	0	242	99	0	
. . . No examinees above this score . . .					
					-----+-----+-----+-----+-----+
					5 10 15 20 25
					Percentage of Examinees

## **Annexure C**

### **Iteman 3.6 analysis of the second test**

# GADGETS & FREAKY INVENTIONS PILOT (JUNE 2014)

ITEMAN (tm) for 32-bit Windows, Version 3.6  
Copyright (c) 1982 - 1998 by Assessment Systems Corporation

Conventional Item and Test Analysis Program

Item analysis for data from file c:\AlbertDocuments\ICELDA\ice01169.txt  
Date: 28 Aug 2014 Time: 11:29

\*\*\*\*\* ANALYSIS SUMMARY INFORMATION \*\*\*\*\*

Data (Input) File: c:\AlbertDocuments\ICELDA\ice01169.txt  
Analysis Output File: c:\AlbertDocuments\ICELDA\ice01169.out  
Score Output File: NONE  
Exceptions File: NONE  
Statistics Output File: NONE  
Scale Definition Codes: DICHOT = Dichotomous MPOINT = Multipoint/Survey

Scale: 1  
-----  
Type of Scale DICHOT  
N of Items 60  
N of Examinees 240

\*\*\*\*\* CONFIGURATION INFORMATION \*\*\*\*\*

Type of Correlations: Point-Biserial

Correction for Spuriousness: YES

Ability Grouping: YES

Subgroup Analysis: NO

Express Endorsements As: PERCENTAGES

Score Group Interval Width: 1

**Subtest 1: Scrambled text**

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics					
		Pcnt Correct	Disc. Index	Point Biser.	Alt.	Pcnt Total	Endorsing		Point Biser.	Key
							Low	High		
1	1-1	55	.76	.56	A	7	17	0	-.25	
					B	55	17	93	.56	*
					C	3	5	0	-.17	
					D	1	2	0	-.06	
					E	33	58	7	-.45	
					Other	1	0	0	-.11	
2	1-2	51	.58	.43	A	22	35	6	-.31	
					B	7	12	0	-.21	
					C	7	11	0	-.17	
					D	51	29	87	.43	*
					E	12	9	6	-.10	
3	1-3	43	.76	.58	A	12	17	0	-.26	
					B	11	14	3	-.16	
					C	14	24	6	-.26	
					D	17	29	3	-.28	
					E	43	9	85	.58	*
					Other	4	0	0	-.11	
4	1-4	44	.69	.51	A	44	15	84	.51	*
					B	17	33	1	-.37	
					C	16	20	9	-.13	
					D	17	24	6	-.22	
					E	5	8	0	-.16	
					Other	1	0	0	-.06	
5	1-5	59	.47	.34	A	13	14	10	-.08	
					B	10	23	3	-.26	
					C	59	36	84	.34	*
					D	13	17	1	-.24	
					E	4	8	0	-.16	
					Other	2	0	0	-.06	

**Subtest 2: Vocabulary knowledge**

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics									
		Pcnt Correct	Disc. Index	Point Biser.	Alt.	Pcnt Total	Endorsing		Point Biser.	Key				
							Low	High						
6	1-6	60	-.51	-.43	A	5	0	12	.19					
					B	60	82	31	-.43	*				
					CHECK THE KEY				C	28	8	50	.32	?
					B was specified, C works better				D	6	5	6	-.01	
									Other	2	0	0	-.13	
7	1-7	90	.20	.23	A	1	3	0	-.13					
					B	5	12	0	-.23					
					C	90	79	99	.23	*				
					D	2	2	1	.01					
					Other	2	0	0	-.19					
8	1-8	77	.52	.44	A	12	24	3	-.26					
					B	4	12	0	-.23					
					C	77	44	96	.44	*				
					D	6	15	1	-.31					
					Other	1	0	0	-.14					

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics						
		Pcnt Correct	Disc. Index	Point Biser.	Alt.	Pcnt Total	Endorsing		Point Biser.	Key	
							Low	High			
9	1-9	77	.37	.34	A	15	30	1	-.39		
					B	4	8	0	-.18		
					C	4	3	4	.02		
					D	77	58	94	.34	*	
					Other	0	0	0	-.08		
10	1-10	63	.53	.43	A	63	35	88	.43	*	
					B	21	38	4	-.36		
					C	8	9	3	-.13		
					D	7	11	4	-.16		
					Other	2	0	0	-.23		
11	1-11	45	.37	.30	A	35	48	26	-.24		
					B	8	15	0	-.26		
					C	45	29	66	.30	*	
					D	12	8	6	-.08		
					Other	1	0	0	.05		
12	1-12	93	.21	.32	A	3	11	0	-.27		
					B	93	79	100	.32	*	
					C	4	11	0	-.24		
					D	0	0	0			
					Other	0	0	0			
13	1-13	13	-.06	-.12	A	65	36	90	.40	?	
					B	13	15	9	-.12	*	
					C	10	20	1	-.22		
					D	12	27	0	-.38		
					Other	0	0	0	-.09		
					CHECK THE KEY B was specified, A works better						
14	1-14	60	.68	.49	A	22	45	3	-.43		
					B	5	6	0	-.14		
					C	13	21	4	-.21		
					D	60	24	93	.49	*	
					Other	1	0	0	-.10		
15	1-15	47	.70	.49	A	47	12	82	.49	*	
					B	1	5	0	-.18		
					C	34	44	12	-.30		
					D	17	39	6	-.35		
					Other	0	0	0			

-----

**Subtest 3: Interpreting graphs and visual information**

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics					
		Pcnt Correct	Disc. Index	Point Biser.	Alt.	Pcnt Total	Endorsing		Point Biser.	Key
							Low	High		
16	1-16	51	.73	.50	A	8	15	1	-.23	
					B	17	30	4	-.27	
					C	51	15	88	.50	*
					D	24	39	6	-.33	
					Other	0	0	0		
17	1-17	63	.62	.51	A	13	26	3	-.29	
					B	19	36	4	-.39	
					C	63	30	93	.51	*
					D	5	6	0	-.15	

Seq. No.	Item Statistics				Alternative Statistics					
	Scale -Item	Pcnt Correct	Disc. Index	Point Biser.	Alt.	Pcnt Total	Endorsing Low	High	Point Biser.	Key
					Other	1	0	0	-.09	
18	1-18	65	.68	.56	A	5	8	3	-.13	
					B	5	5	3	-.07	
					C	65	26	94	.56	*
					D	24	61	0	-.59	
					Other	1	0	0	-.06	
19	1-19	61	.62	.50	A	20	29	4	-.27	
					B	61	32	94	.50	*
					C	8	15	0	-.27	
					D	10	23	1	-.33	
					Other	1	0	0	-.07	
20	1-20	77	.48	.41	A	4	8	0	-.17	
					B	77	52	100	.41	*
					C	13	29	0	-.37	
					D	6	11	0	-.19	
					Other	0	0	0	-.08	
21	1-21	79	.53	.49	A	5	12	1	-.21	
					B	79	45	99	.49	*
					C	12	30	0	-.38	
					D	4	12	0	-.29	
					Other	0	0	0		
22	1-22	44	.63	.45	A	44	23	85	.45	*
					B	27	36	12	-.24	
					C	21	29	3	-.28	
					D	8	12	0	-.22	
					Other	0	0	0		
23	1-23	60	.68	.53	A	12	30	1	-.37	
					B	14	24	1	-.28	
					C	60	26	94	.53	*
					D	13	18	3	-.24	
					Other	1	0	0	-.07	

-----

**Subtest 4: Text comprehension**

Seq. No.	Item Statistics				Alternative Statistics					
	Scale -Item	Pcnt Correct	Disc. Index	Point Biser.	Alt.	Pcnt Total	Endorsing Low	High	Point Biser.	Key
24	1-24	77	.28	.25	A	5	12	1	-.22	
					B	10	11	4	-.11	
					C	77	62	90	.25	*
					D	8	14	4	-.18	
					Other	0	0	0	-.11	
25	1-25	52	.16	.11	A	34	33	31	-.06	
					B	52	45	62	.11	*
					C	10	14	6	-.16	
					D	3	6	1	-.17	
					Other	0	0	0	-.11	
26	1-26	77	.52	.44	A	77	47	99	.44	*
					B	8	21	1	-.29	
					C	4	14	0	-.28	

Seq. No.	Item Statistics				Alternative Statistics					
	Scale -Item	Pcnt Correct	Disc. Index	Point Biser.	Alt.	Pcnt Total	Endorsing Low	High	Point Biser.	Key
27	1-27	66	.13	.13	D	10	17	0	-.26	
					Other	1	0	0	-.05	
					A	3	6	0	-.19	
					B	15	23	12	-.18	
					C	66	55	68	.13	*
					D	15	15	21	-.03	
28	1-28	18	.05	-.04	A	38	35	41	.04	?
					B	28	26	28	-.02	
					C	16	24	10	-.16	
					D	18	15	21	-.04	*
					Other	0	0	0	-.04	
					CHECK THE KEY D was specified, A works better					
29	1-29	65	.44	.32	A	3	6	0	-.21	
					B	14	30	1	-.35	
					C	65	45	90	.32	*
					D	19	18	9	-.12	
					Other	0	0	0		
30	1-30	68	.58	.40	A	8	24	1	-.33	
					B	13	21	1	-.25	
					C	11	17	1	-.18	
					D	68	38	96	.40	*
					Other	0	0	0		
31	1-31	60	.69	.54	A	13	24	4	-.27	
					B	20	42	6	-.41	
					C	60	20	88	.54	*
					D	5	9	1	-.16	
					Other	3	0	0	-.15	
32	1-32	15	-.11	-.13	A	66	52	84	.19	?
					B	10	17	1	-.19	
					C	8	9	6	-.08	
					D	15	20	9	-.13	*
					Other	2	0	0	-.13	
					CHECK THE KEY D was specified, A works better					
33	1-33	65	.64	.50	A	14	36	0	-.42	
					B	65	29	93	.50	*
					C	17	23	7	-.22	
					D	2	6	0	-.20	
					Other	3	0	0	-.15	
34	1-34	50	.49	.36	A	50	29	78	.36	*
					B	21	30	12	-.25	
					C	8	17	1	-.27	
					D	20	18	9	-.12	
					Other	2	0	0	-.16	
35	1-35	61	.42	.28	A	5	12	0	-.24	
					B	22	27	16	-.14	
					C	10	18	4	-.21	
					D	61	35	76	.28	*
					Other	3	0	0	-.09	
36	1-36	67	.58	.46	A	6	18	0	-.30	
					B	67	35	93	.46	*
					C	13	29	3	-.30	
					D	13	15	4	-.23	
					Other	1	0	0	-.12	

Seq. No.	Item Statistics				Alternative Statistics					Key
	Scale -Item	Pcnt Correct	Disc. Index	Point Biser.	Alt.	Pcnt Total	Endorsing Low	High	Point Biser.	
37	1-37	37	.27	.24	A	25	32	18	-.20	
					B	22	24	24	-.03	
					C	37	27	54	.24	*
					D	15	15	4	-.22	
					Other	1	0	0	-.08	
38	1-38	58	.53	.42	A	9	18	3	-.24	
					B	13	15	6	-.13	
					C	18	30	4	-.34	
					D	58	32	85	.42	*
					Other	2	0	0	-.12	
39	1-39	62	.49	.34	A	8	23	0	-.33	
					B	62	33	82	.34	*
					C	18	26	7	-.23	
					D	13	17	10	-.10	
					Other	0	0	0	-.11	
40	1-40	62	.52	.35	A	7	20	0	-.34	
					B	6	15	0	-.25	
					C	22	24	15	-.12	
					D	62	33	85	.35	*
					Other	3	0	0	-.13	
41	1-41	50	.48	.34	A	12	27	0	-.33	
					B	9	18	0	-.32	
					C	50	24	72	.34	*
					D	29	27	28	-.05	
					Other	1	0	0	-.11	
42	1-42	37	.41	.35	A	12	27	4	-.33	
					B	30	41	18	-.26	
					C	20	17	21	-.03	
					D	37	14	54	.35	*
					Other	2	0	0	.00	
43	1-43	51	.22	.17	A	24	20	24	-.02	
					B	51	39	62	.17	*
					C	12	12	6	-.15	
					D	13	27	9	-.24	
					Other	0	0	0	-.08	
44	1-44	35	.51	.35	A	13	23	0	-.29	
					B	18	33	6	-.31	
					C	35	18	69	.35	*
					D	33	24	25	-.02	
					Other	0	0	0	-.08	
45	1-45	43	.20	.11	A	43	32	51	.11	*
					B	32	27	41	.07	
					C	14	23	3	-.25	
					D	10	17	4	-.19	
					Other	0	0	0	-.08	
46	1-46	72	.46	.36	A	5	14	0	-.27	
					B	9	15	1	-.24	
					C	72	45	91	.36	*
					D	13	26	7	-.21	
					Other	0	0	0		
47	1-47	44	.42	.31	A	32	26	24	-.04	
					B	44	30	72	.31	*



Seq. No.	Scale -Item	Item Statistics			Alternative Statistics					
		Pcnt Correct	Disc. Index	Point Biser.	Alt.	Pcnt Total	Endorsing Low	High	Point Biser.	Key
48	1-48	77	.38	.34	C	8	20	1	-.31	
					D	16	24	3	-.30	
					Other	0	0	0		
					A	10	18	7	-.19	
					B	4	8	0	-.18	
					C	7	17	0	-.30	
					D	77	55	93	.34	*
					Other	1	0	0	-.08	

**Subtest 5: Grammar & text relations**

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics					
		Pcnt Correct	Disc. Index	Point Biser.	Alt.	Pcnt Total	Endorsing Low	High	Point Biser.	Key
49	1-49	36	.33	.27	A	33	35	34	-.11	
					B	36	21	54	.27	*
					C	21	30	6	-.25	
					D	8	11	4	-.11	
					Other	2	0	0	-.06	
										A
B	19	32	4	-.29						
C	20	17	21	-.01						
D	44	27	71	.29						*
Other	3	0	0	-.10						
51	1-51	35	.24	.17	A	35	21	46	.17	*
					B	19	29	3	-.31	
					C	36	36	46	.03	
					D	9	12	6	-.13	
					Other	0	0	0	-.08	
					52	1-52	38	.17	.09	A
B	10	15	1	-.14						
C	38	24	41	.09						*
D	18	29	6	-.28						
Other	1	0	0	-.08						
CHECK THE KEY C was specified, A works better										
53	1-53	36	.32	.20	A	30	36	26	-.13	
					B	28	30	18	-.12	
					C	5	8	0	-.17	
					D	36	24	56	.20	*
					Other	1	0	0	-.06	
					54	1-54	38	.46	.36	A
B	10	17	1	-.19						
C	38	21	68	.36						*
D	21	29	13	-.21						
Other	0	0	0	-.08						
55	1-55	53	.61	.45						A
					B	15	30	6	-.31	
					C	11	23	0	-.31	
					D	53	23	84	.45	*
					Other	1	0	0	-.10	
					56	1-56	53	.58	.41	A
B	22	32	10	-.20						

Seq. No.	Item Statistics				Alternative Statistics					
	Scale -Item	Pcnt Correct	Disc. Index	Point Biser.	Alt.	Pcnt Total	Endorsing Low High		Point Biser.	Key
57	1-57	68	.43	.32	C	53	24	82	.41	*
					D	11	20	3	-.24	
					Other	1	0	0	-.10	
					A	68	47	90	.32	*
					B	17	17	6	-.14	
					C	7	21	0	-.35	
58	1-58	66	.34	.24	D	9	14	4	-.18	
					Other	0	0	0	-.08	
					A	10	14	4	-.19	
					B	66	55	88	.24	*
					C	16	17	6	-.14	
					D	7	12	1	-.16	
59	1-59	55	.63	.48	Other	1	0	0	-.11	
					A	13	20	3	-.25	
					B	10	17	4	-.17	
					C	22	38	6	-.36	
					D	55	24	87	.48	*
					Other	0	0	0	-.08	
60	1-60	59	.68	.54	A	59	24	93	.54	*
					B	8	15	1	-.25	
					C	20	32	3	-.32	
					D	13	26	3	-.32	
					Other	1	0	0	-.11	

ITEMAN (tm) for 32-bit Windows, Version 3.6  
Copyright (c) 1982 - 1998 by Assessment Systems Corporation  
Conventional Item and Test Analysis Program  
Item analysis for data from file c:\AlbertDocuments\ICELDA\ice01169.txt

Date: 28 Aug 2014

Time: 11:29

There were 240 examinees in the data file.

Scale Statistics

-----

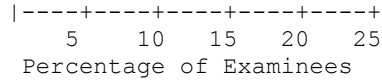
Scale:	1
	-----
N of Items	60
N of Examinees	240
Mean	33.233
Variance	110.962
Std. Dev.	10.534
Skew	-0.027
Kurtosis	-0.909
Minimum	9.000
Maximum	54.000
Median	33.000
Alpha	0.896
SEM	3.396
Mean Pcnt Corr	55
Mean Item-Tot.	0.333
Mean Biserial	0.432
Max Score (Low)	26
N (Low Group)	66
Min Score (High)	41
N (High Group)	68

-----

SCALE # 1

Score Distribution Table

Number Correct	Freq- uency	Cum Freq	PR	PCT	
. . . No examinees below this score . . .					
8	0	0	1	0	
9	1	1	1	0	
10	0	1	1	0	+
11	1	2	1	0	
12	0	2	1	0	
13	3	5	2	1	#
14	1	6	3	0	
15	4	10	4	2	+##
16	0	10	4	0	
17	6	16	7	2	##
18	5	21	9	2	##
19	5	26	11	2	##
20	9	35	15	4	+####
21	7	42	17	3	###
22	6	48	20	2	##
23	4	52	22	2	##
24	3	55	23	1	#
25	6	61	25	2	+##
26	5	66	28	2	##
27	8	74	31	3	###
28	14	88	37	6	#####
29	5	93	39	2	##
30	6	99	41	2	+##
31	8	107	45	3	###
32	5	112	47	2	##
33	11	123	51	5	#####
34	8	131	55	3	###
35	7	138	57	3	+###
36	7	145	60	3	###
37	9	154	64	4	####
38	7	161	67	3	###
39	6	167	70	2	##
40	5	172	72	2	+##
41	6	178	74	2	##
42	9	187	78	4	####
43	3	190	79	1	#
44	6	196	82	2	##
45	7	203	85	3	+###
46	5	208	87	2	##
47	6	214	89	2	##
48	6	220	92	2	##
49	6	226	94	2	##
50	4	230	96	2	+##
51	5	235	98	2	##
52	0	235	98	0	
53	4	239	99	2	##
54	1	240	99	0	
55	0	240	99	0	+
56	0	240	99	0	
. . . No examinees above this score . . .					



## **Annexure D**

### **Iteman 4.3 analysis of TALA**



# ***Classical Item and Test Analysis Report***

## ***TALA re-pilot (Eunice & Heidedal): June 2014***

***Report created on 2014/08/28***

***Iteman: Software for Classical Analysis***

***Copyright © 2013 - Assessment Systems Corporation***



## Introduction

This report provides the results of a classical item and test analysis by the computer program Iteman Version 4.3 (Assessment Systems Corporation, 2013) for TALA re-pilot (Eunice & Heidedal): June 20141. The output is divided into three sections:

1. Specifications
2. Summary statistics
3. Item-by-item results.

The statistical output is also recorded in a comma-separated value (CSV) file of the same name.

## Specifications

The Windows paths for the input files used in this analysis were:

C:\AlbertDocuments\ICELDA\ice01176.txt  
C:\AlbertDocuments\ICELDA\ice01177.txt

The Windows paths for the output files produced by this analysis were:

C:\AlbertDocuments\ICELDA\ice01178.rtf  
C:\AlbertDocuments\ICELDA\ice01178.csv  
C:\AlbertDocuments\ICELDA\ice01178 Scores.csv

Table 1 presents the specifications and basic information concerning the analysis. This provides important documentation of the setup of the program for historical purposes.

**Table 1: Specifications**

Specification	Value	Specification	Value
Number of examinees	242	Total Items	60
Scored Items	60	Pretest Items	0
Multiple Choice Items	60	Polytomous Items	0
Number of domains	5	External scores	No
Minimum P	0.15	Maximum P	0.84
Minimum item mean	0.00	Maximum item mean	15.00
Minimum item correlation	0.15	Maximum item correlation	1.00
ITEMAN 3.0 Header	No	Exclude omits from option statistics	No
Number of ID columns	0	ID begins in column	0
Responses begin in column	1	Omit character	X
Not Admin character	N	Produce quantile tables	Yes
Correct for spuriousness	Yes	Produce quantile plots	Yes
Save data matrix	No	Include omit codes in matrix	N/A
Scaled score setting 2	N/A		
Classify based on	Total Score	Cutpoint	1.000
Low group label	Low	High group label	High

## Summary statistics

Table 2 presents the summary statistics of the test, for all items, scored items only, and for each domain (content area). Definitions of these statistics are found in the Iteman manual.

**Table 2: Summary statistics**

Score	Items	Mean	SD	Min Score	Max Score	Mean P	Mean Rpbis
All items	60	25.112	8.121	6	50	0.419	0.239
Scored Items	60	25.112	8.121	6	50	0.419	0.239
Scrambled text	5	2.541	1.441	0	5	0.508	0.131
Vocabulary knowledge	10	5.306	2.051	0	10	0.531	0.272
Interpreting graphs & visual information	8	2.764	1.687	0	8	0.346	0.190
Text comprehension	25	9.950	4.179	2	22	0.398	0.266
Grammar & text relations	12	4.550	2.461	0	12	0.379	0.231

Table 3 presents a reliability analysis of the tests. Alpha (also known as KR-20) is the most commonly used index of reliability, and is therefore used to calculate the standard error of measurement (SEM) on the raw score scale. Also presented are three configurations of split-half reliability, first as uncorrected correlations, and then as Spearman-Brown (S-B) corrected correlations. This is because an uncorrected split-half correlation is referenced to a "test" that only contains half as many items as the full test, and therefore underestimates reliability.

The cutscore on this exam was 1.000, producing a pass rate of 100.0%. The Livingston index of classification consistency at the cut-score was 0.982.

**Table 3: Reliability**

Score	Alpha	SEM	Split-Half (Random)	Split-Half (First-Last)	Split-Half (Odd-Even)	S-B Random	S-B First-Last	S-B Odd-Even
Scored items	0.819	3.457	0.729	0.601	0.684	0.843	0.751	0.813
Scrambled text	0.598	0.913	0.151	0.221	0.554	0.263	0.362	0.713
Vocabulary knowledge	0.547	1.381	0.364	0.382	0.367	0.534	0.553	0.537
Interpreting graphs & visual information	0.459	1.241	0.339	0.320	0.264	0.506	0.485	0.417
Text comprehension	0.722	2.204	0.586	0.554	0.564	0.739	0.713	0.721
Grammar & text relations	0.628	1.501	0.381	0.301	0.633	0.552	0.462	0.775



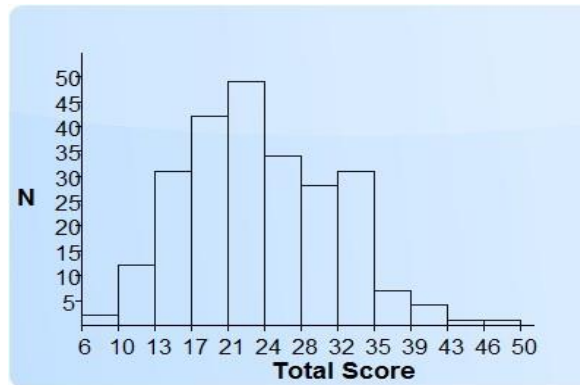
Table 4 presents the item statistics and flags for the item(s) that were flagged during the analysis

**Table 4: Summary Statistics for the Flagged Items**

Item ID	P / Item Mean	R	Flag(s)
2	0.636	-0.039	K, LR
8	0.909	0.194	HP
13	0.293	0.102	K, LR
14	0.372	0.092	LR
16	0.302	-0.009	K, LR
19	0.178	0.122	LR
21	0.335	0.108	LR
23	0.306	0.130	LR
25	0.318	0.022	K, LR
29	0.306	0.033	K, LR
38	0.335	0.063	K, LR
49	0.190	0.062	K, LR
50	0.124	0.177	LP
56	0.413	0.124	LR

Figure 1 displays the distribution of the raw scores for the scored items across all domains. Table 5 displays the frequency distribution for total score shown in Figure 1.

**Figure 1: Total score for the scored items**

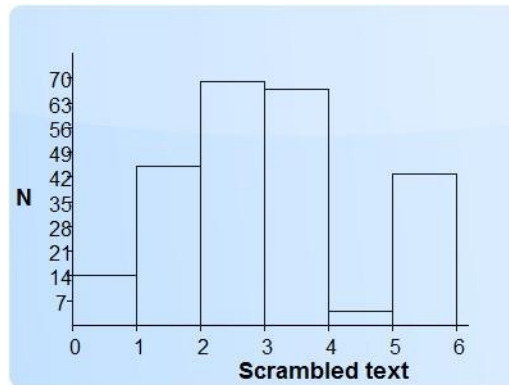


**Table 5: Frequency Distribution for Total Score**

Range	Frequency
5 to 9	2
10 to 13	12
14 to 17	31
18 to 21	42
22 to 25	49
26 to 29	34
30 to 33	28
34 to 37	31
38 to 41	7
42 to 45	4
46 to 49	1
50	1

Figure 2 displays the distribution of the raw scores for Scrambled text. Table 6 displays the frequency distribution of domain scores shown in Figure 2.

**Figure 2: Raw scores for Scrambled text**

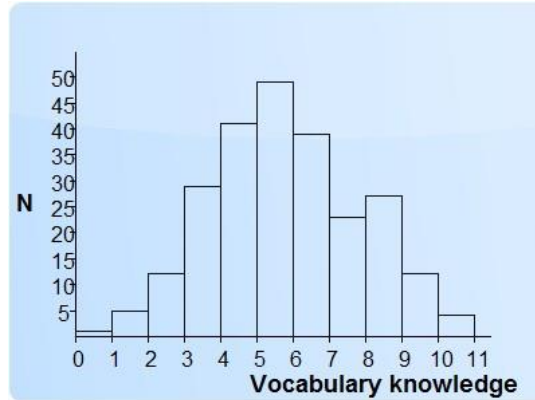


**Table 6: Frequency Distribution for Scrambled text**

Score	Frequency
0	14
1	45
2	69
3	67
4	4
5	43

Figure 3 displays the distribution of the raw scores for Vocabulary knowledge. Table 7 displays the frequency distribution of domain scores shown in Figure 3.

**Figure 3: Raw scores for Vocabulary knowledge**

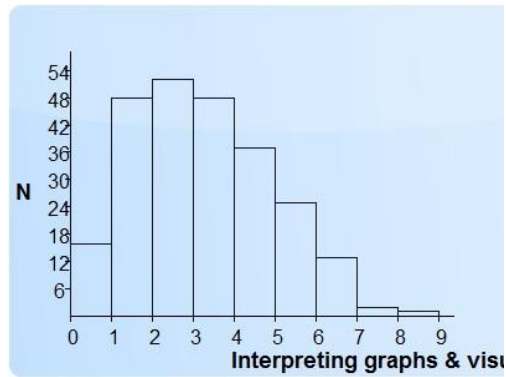


**Table 7: Frequency Distribution for Vocabulary knowledge**

Score	Frequency
0	1
1	5
2	12
3	29
4	41
5	49
6	39
7	23
8	27
9	12
10	4

Figure 4 displays the distribution of the raw scores for Interpreting graphs & visual information. Table 8 displays the frequency distribution of domain scores shown in Figure 4.

**Figure 4: Raw scores for Interpreting graphs & visual information**

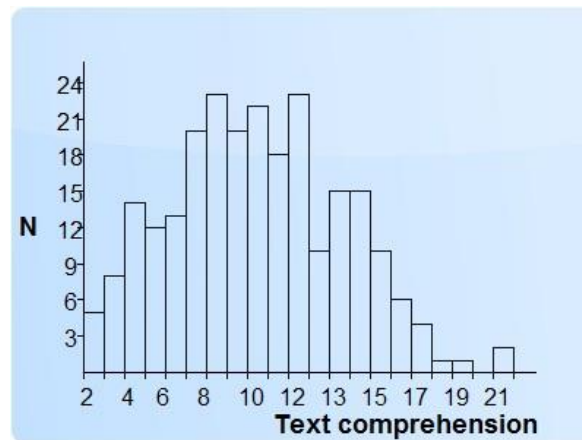


**Table 8: Frequency Distribution for Interpreting graphs & visual information**

Score	Frequency
0	16
1	48
2	52
3	48
4	37
5	25
6	13
7	2
8	1

Figure 5 displays the distribution of the raw scores for Text comprehension. Table 9 displays the frequency distribution of domain scores shown in Figure 5.

**Figure 5: Raw scores for Text comprehension**

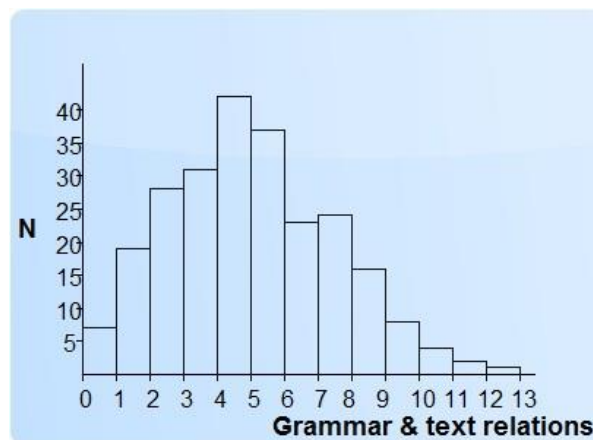


**Table 9: Frequency Distribution for Text comprehension**

Range	Frequency
1 to 2	5
3	8
4	14
5	12
6	13
7	20
8	23
9	20
10	22
11	18
12	23
13	10
14	15
15	15
16	10
17	6
18	4
19	1
20	1
21	0
22	2

Figure 6 displays the distribution of the raw scores for Grammar & text relations. Table 10 displays the frequency distribution of domain scores shown in Figure 6.

**Figure 6: Raw scores for Grammar & text relations**



**Table 10: Frequency Distribution for Grammar & text relations**

Score	Frequency
0	7
1	19
2	28
3	31
4	42
5	37
6	23
7	24
8	16
9	8
10	4
11	2
12	1

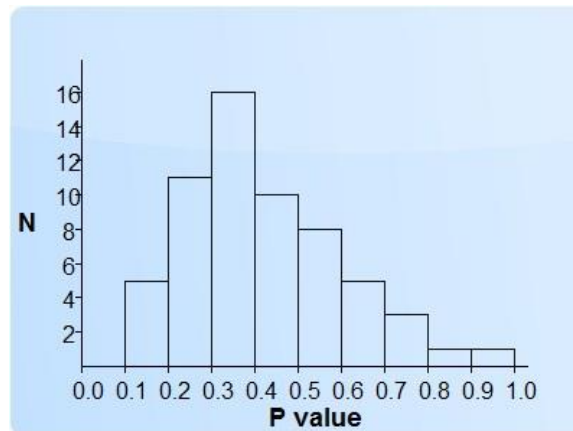
Table 11 displays the correlations of domain scores.

**Table 11: Correlations for Domain Scores**

Domain	Scrambled text	Vocabulary knowledge	Interpreting graphs & visual information	Text comprehension	Grammar & text relations
Scrambled text	1.000	0.095	0.053	0.146	0.040
Vocabulary knowledge	0.095	1.000	0.296	0.559	0.385
Interpreting graphs & visual information	0.053	0.296	1.000	0.380	0.171
Text comprehension	0.146	0.559	0.380	1.000	0.396
Grammar & text relations	0.040	0.385	0.171	0.396	1.000

Figure 7 displays the distribution of the P values for the dichotomously scored items (correct/incorrect). Table 12 displays the frequency distribution of the P values shown in Figure 7.

**Figure 7: P values for the scored items**

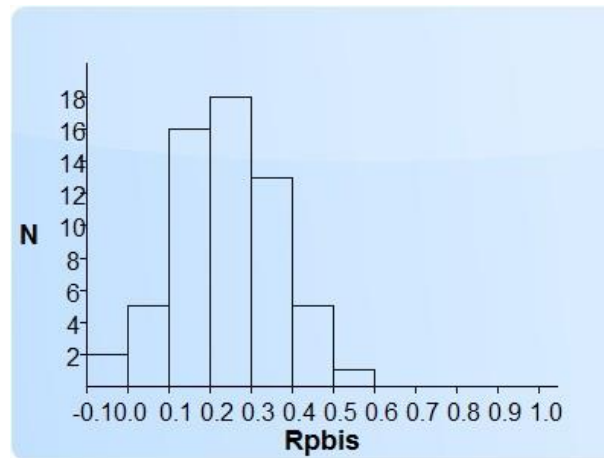


**Table 12: Frequency Distribution for the P values**

Score	Frequency
0.0 to 0.1	0
0.1 to 0.2	5
0.2 to 0.3	11
0.3 to 0.4	16
0.4 to 0.5	10
0.5 to 0.6	8
0.6 to 0.7	5
0.7 to 0.8	3
0.8 to 0.9	1
0.9 to 1.0	1

Figure 8 displays the distribution of the Point-Biserial Correlations for the dichotomously scored items (correct/incorrect). Table 13 displays the frequency distribution of the Point-Biserial correlations shown in Figure 8.

**Figure 8: Rpbis for the scored items**



**Table 13: Frequency Distribution for the Rpbis**

Score	Frequency
-0.1 to 0.0	2
0.0 to 0.1	5
0.1 to 0.2	16
0.2 to 0.3	18
0.3 to 0.4	13
0.4 to 0.5	5
0.5 to 0.6	1
0.6 to 0.7	0
0.7 to 0.8	0
0.8 to 0.9	0
0.9 to 1.0	0

Figure 9 displays the scatterplot of P (difficulty) by Rpbis (discrimination) for the dichotomously scored items (correct/incorrect).

**Figure 9: P by Rpbis**

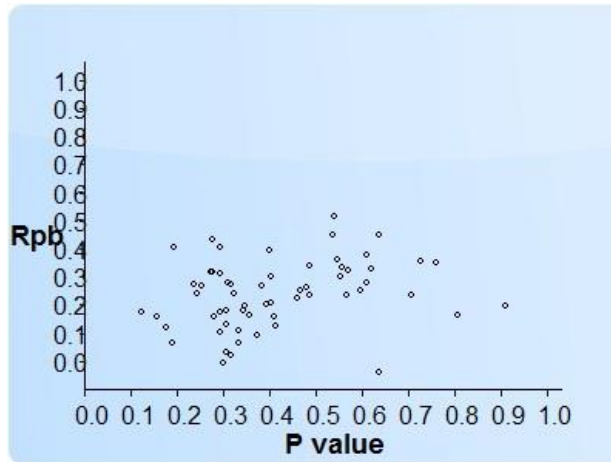
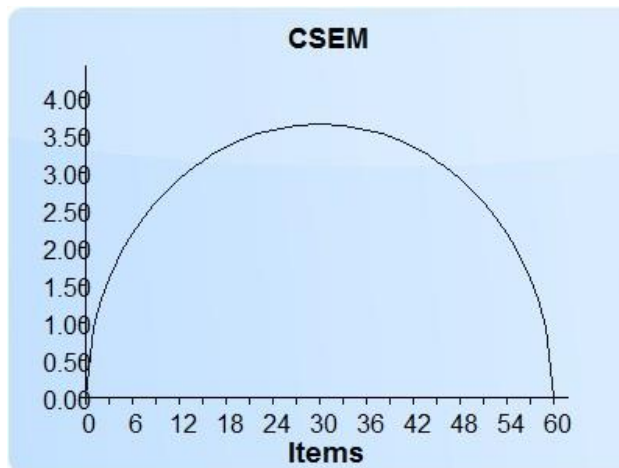


Figure 10 displays a graph of the Conditional Standard Error of Measurement (CSEM) Formula IV. The CSEM at the cutscore of 1.000 equaled 0.933.

**Figure 10: CSEM**





## ***Item-by-item results***

The following section presents the item-by-item results of the analysis. Each item has several tables and a figure. The figure, called a quantile plot, shows the proportion of examinees selecting each option, for consecutive segments of the examinees as ranked by score. The key thing to evaluate in this figure is that the line for the correct answer has a positive slope (goes up from left to right), which means that examinees with higher scores tend to answer correctly more often. Conversely, the lines for the incorrect options, called distractors, should have a negative slope. Note, however, that the use of a small number of groups (e.g., 3 or fewer) oversimplifies the graph, so that items which are very difficult or very easy (that is, discriminating in only the top or bottom 20% of examinees) might appear to have poor quantile plots and classical statistics. For such items, item response theory presents significant advantages in analysis

There are four tables presented for each item.

1. Item information table: records the information supplied by the control file (or Iteman 3 header) for this item.
2. Item statistics table: overall item statistics.
3. Option statistics: detailed statistics for each item, which helps diagnose issues in items with poor statistics.
4. Quantile plot data: the values used to create the quantile plot.

The item statistics table presents overall item statistics in the first row of numbers. The two most important item-level statistics for dichotomously scored (correct/incorrect) items are the P value and the point-biserial correlation, which represent the difficulty and discrimination of the item, respectively. For polytomously scored (rating scale or partial credit) items, the difficulty is represented by the mean (average) item score, while the discrimination is represented by a Pearson r correlation.

The P value is the proportion of examinees that answered an item in the keyed direction. P ranges from 0 to 1. A high value (0.95) means that an item is easy, a low value (0.25) means that the item is difficult. The point-biserial correlation (Rpbis) is a measure of the discriminating, or differentiating, power of the item. Rpbis ranges from -1 to 1. A negative Rpbis is indicative of a bad item as lower scoring examinees are more likely than higher scoring examinees to respond in the keyed direction.

For rating scale or partial credit items, the mean item score ranges from the minimum to the maximum of the scale. For example, if the item has a rating scale of 1 to 5, the possible range for the mean is 1 to 5. The Pearson r is similar to the Rpbis in that it ranges from -1 to 1, with a positive r indicating that the item correlates well with total score.

The option statistics table presents statistics for each individual option (alternative). The key thing to examine in this portion of the table is that no distractors have a higher Rpbis than the correct answer. That indicates that higher scoring examinees are selecting the incorrect answer, which therefore might be arguably correct.

The quantile plot data table simply presents the values calculated to create the quantile plot. Because it contains the same information, the quantile plot itself presents a useful picture of the item's performance, but this table can be used to examine that performance in detail to help diagnose possible issues.

### Item 1 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
1	1	D	Yes	5	Scrambled text	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.806	0.445	0.640	0.165	0.237	0.818

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	3	0.012	0.014	0.049	25.333	8.083	
B	31	0.128	-0.083	-0.132	22.581	6.956	
C	9	0.037	-0.149	-0.348	18.222	3.346	
D	195	0.806	0.165	0.237	25.954	8.296	**KEY**
E	2	0.008	-0.038	-0.151	21.000	2.828	
Omit	2	0.008	-0.049	-0.199	17.000	11.314	

### Item 2 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
2	2	C	Yes	5	Scrambled text	K, LR

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.636	0.487	0.624	-0.039	-0.051	0.823

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	17	0.070	0.014	0.026	24.882	8.580	
B	27	0.112	0.051	0.084	25.630	6.862	
C	154	0.636	-0.039	-0.051	25.234	8.487	**KEY**
D	17	0.070	-0.102	-0.194	21.471	6.186	
E	22	0.091	0.163	0.286	28.636	6.814	
Omit	5	0.021	-0.085	-0.243	16.200	5.718	

### Item 3 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
3	3	B	Yes	5	Scrambled text	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.343	0.529	0.684	0.179	0.231	0.818

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	55	0.227	-0.100	-0.139	23.291	8.025	
B	83	0.343	0.179	0.231	27.747	8.338	**KEY**
C	40	0.165	0.096	0.144	26.500	6.461	
D	10	0.041	-0.134	-0.302	19.600	5.441	
E	45	0.186	-0.051	-0.075	23.911	7.786	
Omit	9	0.037	-0.102	-0.238	17.889	8.388	

### Item 4 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
4	4	A	Yes	5	Scrambled text	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.409	0.444	0.562	0.158	0.200	0.818

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	99	0.409	0.158	0.200	27.222	8.052	**KEY**
B	26	0.107	-0.061	-0.101	23.308	6.938	
C	29	0.120	0.028	0.046	25.310	8.661	
D	12	0.050	-0.037	-0.078	23.417	8.262	
E	70	0.289	-0.136	-0.180	23.000	7.892	
Omit	6	0.025	0.007	0.018	25.167	9.087	

### Item 5 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
5	5	E	Yes	5	Scrambled text	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.347	0.367	0.474	0.194	0.251	0.817

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	60	0.248	-0.073	-0.100	23.750	7.697	
B	70	0.289	-0.027	-0.036	24.429	8.168	
C	8	0.033	-0.055	-0.134	22.375	7.050	
D	10	0.041	-0.129	-0.290	19.800	5.846	

E	84	0.347	0.194	0.251	27.893	8.111	**KEY**
Omit	10	0.041	-0.046	-0.103	22.200	7.772	

### **Item 6 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
6	6	B	Yes	4	Vocabulary knowledge	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.636	0.436	0.558	0.450	0.577	0.811

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	7	0.029	-0.111	-0.281	19.429	3.952	
B	154	0.636	0.450	0.577	28.149	7.656	**KEY**
C	48	0.198	-0.331	-0.473	19.250	6.299	
D	30	0.124	-0.197	-0.317	20.367	5.430	
Omit	3	0.012	-0.008	-0.028	23.667	12.014	

### **Item 7 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
7	7	D	Yes	4	Vocabulary knowledge	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.570	0.377	0.476	0.323	0.407	0.814

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	50	0.207	-0.378	-0.536	18.680	5.523	
B	27	0.112	0.035	0.059	25.333	7.791	
C	23	0.095	-0.065	-0.113	22.957	6.049	
D	138	0.570	0.323	0.407	27.761	8.034	**KEY**
Omit	4	0.017	0.005	0.016	25.000	7.439	

### **Item 8 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
8	8	C	Yes	4	Vocabulary knowledge	HP

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.909	0.281	0.494	0.194	0.341	0.817

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	9	0.037	-0.076	-0.176	21.111	12.077	
B	8	0.033	-0.114	-0.277	19.250	5.625	
C	220	0.909	0.194	0.341	25.695	7.876	**KEY**
D	4	0.017	-0.112	-0.348	17.250	8.180	
Omit	1	0.004	-0.049	-0.257	11.000	0.000	

### Item 9 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
9	9	C	Yes	4	Vocabulary knowledge	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.545	0.424	0.533	0.361	0.453	0.813

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	52	0.215	-0.281	-0.395	20.327	5.940	
B	12	0.050	-0.226	-0.479	16.750	6.341	
C	132	0.545	0.361	0.453	28.167	7.661	**KEY**
D	45	0.186	-0.057	-0.083	23.622	7.640	
Omit	1	0.004	0.050	0.261	38.000	0.000	

### Item 10 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
10	10	C	Yes	4	Vocabulary knowledge	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.401	0.411	0.521	0.396	0.502	0.812

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	38	0.157	-0.039	-0.059	24.000	8.107	
B	72	0.298	-0.218	-0.288	22.069	5.720	
C	97	0.401	0.396	0.502	29.526	7.537	**KEY**
D	32	0.132	-0.215	-0.340	20.375	8.541	
Omit	3	0.012	-0.044	-0.150	20.000	7.810	

### Item 11 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
11	11	A	Yes	4	Vocabulary knowledge	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.707	0.431	0.570	0.238	0.315	0.816

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	171	0.707	0.238	0.315	26.626	8.073	**KEY**
B	48	0.198	-0.080	-0.114	23.125	7.166	
C	8	0.033	-0.210	-0.508	15.375	3.998	
D	11	0.045	-0.103	-0.225	20.636	6.360	
Omit	4	0.017	-0.078	-0.240	16.000	7.394	

### Item 12 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
12	12	C	Yes	4	Vocabulary knowledge	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.277	0.296	0.396	0.317	0.423	0.815

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	76	0.314	0.140	0.183	26.474	7.422	
B	15	0.062	-0.096	-0.190	21.867	6.728	
C	67	0.277	0.317	0.423	29.896	8.306	**KEY**
D	81	0.335	-0.346	-0.448	20.963	5.995	
Omit	3	0.012	-0.084	-0.291	12.000	3.606	

### Item 13 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
13	13	B	Yes	4	Vocabulary knowledge	K, LR

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.293	0.186	0.246	0.102	0.134	0.819

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	26	0.107	-0.393	-0.658	15.731	4.522	
B	71	0.293	0.102	0.134	27.085	8.316	**KEY**
C	115	0.475	0.264	0.331	27.043	7.362	
D	28	0.116	-0.140	-0.229	21.714	5.563	
Omit	2	0.008	-0.065	-0.259	13.500	3.536	

### Item 14 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
14	14	D	Yes	4	Vocabulary knowledge	LR

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.372	0.142	0.182	0.092	0.118	0.820

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	51	0.211	-0.001	-0.001	24.725	7.278	
B	88	0.364	-0.003	-0.004	24.705	7.905	
C	10	0.041	-0.164	-0.369	18.400	5.232	
D	90	0.372	0.092	0.118	26.700	8.694	**KEY**
Omit	3	0.012	-0.055	-0.190	18.333	6.658	

### Item 15 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
15	15	A	Yes	4	Vocabulary knowledge	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.595	0.335	0.425	0.251	0.317	0.816

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	144	0.595	0.251	0.317	27.160	8.554	**KEY**
B	5	0.021	-0.035	-0.100	22.600	3.847	
C	46	0.190	-0.193	-0.279	21.348	5.755	
D	44	0.182	-0.083	-0.121	23.114	7.371	
Omit	3	0.012	-0.056	-0.194	18.000	6.083	

**Item 16 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
16	16	A	Yes	4	Interpreting graphs & visual information	K, LR

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.302	0.089	0.117	-0.009	-0.012	0.822

**Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	73	0.302	-0.009	-0.012	25.699	8.009	**KEY**
B	88	0.364	0.158	0.203	26.500	8.245	
C	20	0.083	-0.056	-0.101	23.300	8.196	
D	57	0.236	-0.117	-0.162	23.105	7.841	
Omit	4	0.017	-0.036	-0.112	21.500	7.594	

**Item 17 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
17	17	D	Yes	4	Interpreting graphs & visual information	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.318	0.271	0.354	0.276	0.360	0.815

**Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	39	0.161	-0.287	-0.432	19.590	5.609	
B	54	0.223	0.219	0.305	28.037	7.848	
C	71	0.293	-0.249	-0.330	21.718	6.234	
D	77	0.318	0.276	0.360	29.000	8.237	**KEY**
Omit	1	0.004	-0.005	-0.024	24.000	0.000	

**Item 18 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
18	18	D	Yes	4	Interpreting graphs & visual information	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.273	0.335	0.449	0.320	0.429	0.814



### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	61	0.252	-0.093	-0.126	23.574	6.299	
B	81	0.335	-0.122	-0.159	23.469	7.201	
C	34	0.140	-0.128	-0.200	22.324	7.551	
D	66	0.273	0.320	0.429	29.985	9.077	**KEY**

### Item 19 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
19	19	A	Yes	4	Interpreting graphs & visual information	LR

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.178	0.219	0.322	0.122	0.180	0.819

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	43	0.178	0.122	0.180	28.047	10.589	**KEY**
B	114	0.471	-0.012	-0.015	24.833	6.962	
C	66	0.273	-0.105	-0.140	23.561	7.817	
D	18	0.074	0.047	0.088	26.278	7.835	
Omit	1	0.004	-0.048	-0.255	12.000	0.000	

### Item 20 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
20	20	C	Yes	4	Interpreting graphs & visual information	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.488	0.417	0.523	0.338	0.424	0.814

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	48	0.198	-0.204	-0.293	21.375	6.862	
B	59	0.244	-0.160	-0.220	22.390	7.985	
C	118	0.488	0.338	0.424	28.364	7.618	**KEY**
D	15	0.062	-0.085	-0.168	22.000	7.041	
Omit	2	0.008	0.015	0.061	26.500	10.607	

### Item 21 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
21	21	C	Yes	4	Interpreting graphs & visual information	LR

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.335	0.262	0.339	0.108	0.140	0.819

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	86	0.355	-0.043	-0.055	24.314	7.774	
B	37	0.153	-0.011	-0.017	24.568	8.588	
C	81	0.335	0.108	0.140	27.000	8.244	**KEY**
D	33	0.136	-0.080	-0.126	23.152	7.492	
Omit	5	0.021	0.005	0.015	25.200	10.826	

### Item 22 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
22	22	B	Yes	4	Interpreting graphs & visual information	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.566	0.399	0.503	0.237	0.299	0.816

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	34	0.140	0.071	0.110	25.941	7.663	
B	137	0.566	0.237	0.299	27.197	8.243	**KEY**
C	43	0.178	-0.229	-0.336	20.628	6.484	
D	28	0.116	-0.171	-0.280	20.786	6.321	

### Item 23 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
23	23	D	Yes	4	Interpreting graphs & visual information	LR

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.306	0.234	0.307	0.130	0.171	0.819

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	36	0.149	-0.103	-0.158	22.833	8.392	
B	41	0.169	-0.077	-0.114	23.439	7.239	
C	87	0.360	-0.004	-0.006	24.759	7.893	
D	74	0.306	0.130	0.171	27.378	8.412	**KEY**
Omit	4	0.017	0.040	0.124	28.500	6.952	

### Item 24 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
24	24	B	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.405	0.310	0.392	0.299	0.379	0.815

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	62	0.256	0.006	0.008	24.790	7.182	
B	98	0.405	0.299	0.379	28.582	8.216	**KEY**
C	25	0.103	-0.099	-0.167	22.400	8.231	
D	56	0.231	-0.280	-0.388	20.661	6.299	
Omit	1	0.004	-0.015	-0.080	22.000	0.000	

### Item 25 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
25	25	B	Yes	4	Text comprehension	K, LR

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.318	0.064	0.084	0.022	0.029	0.821

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	16	0.066	-0.183	-0.354	19.250	6.588	
B	77	0.318	0.022	0.029	26.052	7.635	**KEY**
C	113	0.467	0.145	0.182	26.044	8.674	
D	34	0.140	-0.097	-0.152	22.853	6.845	
Omit	2	0.008	-0.025	-0.102	21.500	0.707	

### **Item 26 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
26	26	A	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.479	0.319	0.399	0.264	0.331	0.816

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	116	0.479	0.264	0.331	27.819	7.799	**KEY**
B	43	0.178	-0.139	-0.204	22.256	7.737	
C	54	0.223	-0.080	-0.112	23.444	8.357	
D	28	0.116	-0.144	-0.237	21.464	6.310	
Omit	1	0.004	0.008	0.041	26.000	0.000	

### **Item 27 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
27	27	B	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.612	0.318	0.404	0.278	0.353	0.815

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	45	0.186	-0.239	-0.347	20.533	8.030	
B	148	0.612	0.278	0.353	27.257	7.834	**KEY**
C	29	0.120	-0.118	-0.191	21.966	6.367	
D	17	0.070	-0.050	-0.095	23.059	7.327	
Omit	3	0.012	0.049	0.170	30.000	8.185	

### **Item 28 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
28	28	D	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.384	0.319	0.406	0.268	0.341	0.816

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	44	0.182	-0.258	-0.376	20.386	7.314	
B	84	0.347	-0.013	-0.017	24.583	7.119	
C	19	0.079	-0.094	-0.173	22.158	8.859	
D	93	0.384	0.268	0.341	28.419	7.985	<b>**KEY**</b>
Omit	2	0.008	0.006	0.025	25.500	3.536	

### Item 29 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
29	29	A	Yes	4	Text comprehension	K, LR

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.306	0.125	0.164	0.033	0.043	0.821

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	74	0.306	0.033	0.043	26.203	7.467	<b>**KEY**</b>
B	54	0.223	-0.225	-0.313	21.426	7.980	
C	74	0.306	0.079	0.104	25.770	8.264	
D	39	0.161	0.079	0.119	26.256	7.279	
Omit	1	0.004	0.059	0.309	50.000	0.000	

### Item 30 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
30	30	C	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.244	0.239	0.327	0.240	0.329	0.816

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	84	0.347	-0.119	-0.153	23.571	7.955	
B	69	0.285	-0.118	-0.157	23.377	7.331	
C	59	0.244	0.240	0.329	29.237	8.368	<b>**KEY**</b>
D	27	0.112	0.004	0.007	24.963	7.547	
Omit	3	0.012	0.033	0.113	28.333	5.508	

### Item 31 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
31	31	D	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.293	0.390	0.517	0.408	0.540	0.812

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	90	0.372	-0.253	-0.323	22.222	6.416	
B	33	0.136	-0.182	-0.285	21.212	6.609	
C	45	0.186	0.037	0.053	25.422	7.638	
D	71	0.293	0.408	0.540	30.817	7.846	**KEY**
Omit	3	0.012	-0.074	-0.255	15.000	8.718	

### Item 32 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
32	32	B	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.760	0.420	0.577	0.348	0.479	0.814

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	7	0.029	-0.182	-0.461	16.000	3.317	
B	184	0.760	0.348	0.479	26.902	7.708	**KEY**
C	18	0.074	-0.173	-0.324	19.500	7.579	
D	30	0.124	-0.156	-0.251	21.067	6.705	
Omit	3	0.012	-0.087	-0.299	10.667	1.528	

### Item 33 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
33	33	A	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.194	0.394	0.566	0.407	0.586	0.813

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	47	0.194	0.407	0.586	32.489	7.423	**KEY**
B	55	0.227	-0.067	-0.092	23.945	5.914	
C	77	0.318	-0.324	-0.424	21.156	6.600	
D	56	0.231	0.105	0.145	26.429	7.823	
Omit	7	0.029	-0.093	-0.237	17.714	9.340	

### Item 34 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
34	34	D	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.727	0.432	0.579	0.355	0.476	0.814

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	26	0.107	-0.165	-0.277	20.615	7.278	
B	20	0.083	-0.215	-0.389	18.700	5.459	
C	17	0.070	-0.130	-0.246	20.647	6.123	
D	176	0.727	0.355	0.476	27.108	7.790	**KEY**
Omit	3	0.012	-0.072	-0.248	15.000	11.533	

### Item 35 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
35	35	D	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.355	0.221	0.284	0.166	0.214	0.818

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	73	0.302	0.046	0.061	25.315	8.079	
B	37	0.153	0.047	0.072	25.649	8.826	
C	42	0.174	-0.264	-0.391	20.143	6.111	
D	86	0.355	0.166	0.214	27.547	7.469	**KEY**
Omit	4	0.017	-0.078	-0.242	16.250	10.468	

**Item 36 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
36	36	B	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.293	0.218	0.289	0.174	0.230	0.818

**Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	46	0.190	-0.206	-0.298	21.413	6.466	
B	71	0.293	0.174	0.230	27.972	8.270	**KEY**
C	39	0.161	-0.050	-0.076	23.897	7.055	
D	81	0.335	0.092	0.119	25.852	8.130	
Omit	5	0.021	-0.089	-0.255	16.000	9.874	

**Item 37 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
37	37	D	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.157	0.109	0.165	0.158	0.239	0.818

**Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	112	0.463	0.054	0.068	25.420	7.279	
B	48	0.198	-0.093	-0.133	23.458	7.377	
C	38	0.157	-0.108	-0.163	22.947	7.867	
D	38	0.157	0.158	0.239	28.895	10.224	**KEY**
Omit	6	0.025	-0.036	-0.096	22.333	9.223	

**Item 38 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
38	38	A	Yes	4	Text comprehension	K, LR

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.335	0.074	0.096	0.063	0.082	0.820



**Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	81	0.335	0.063	0.082	26.494	8.276	**KEY**
B	62	0.256	-0.059	-0.080	23.968	7.624	
C	37	0.153	-0.114	-0.173	22.622	7.395	
D	58	0.240	0.117	0.160	26.448	8.090	
Omit	4	0.017	-0.063	-0.194	18.500	13.000	

**Item 39 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
39	39	D	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.236	0.277	0.382	0.275	0.380	0.816

**Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	78	0.322	-0.237	-0.309	22.141	7.882	
B	69	0.285	0.044	0.059	25.435	7.074	
C	35	0.145	-0.032	-0.049	24.257	6.437	
D	57	0.236	0.275	0.380	29.825	8.263	**KEY**
Omit	3	0.012	-0.073	-0.250	15.333	11.372	

**Item 40 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
40	40	D	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.277	0.416	0.555	0.435	0.581	0.812

**Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	108	0.446	-0.177	-0.223	23.278	7.202	
B	41	0.169	-0.175	-0.260	21.780	5.944	
C	23	0.095	-0.081	-0.140	22.870	7.288	
D	67	0.277	0.435	0.581	31.373	7.518	**KEY**
Omit	3	0.012	-0.078	-0.269	14.000	9.165	

### **Item 41 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
41	41	D	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.541	0.472	0.593	0.513	0.645	0.809

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	48	0.198	-0.216	-0.310	21.167	5.854	
B	29	0.120	-0.251	-0.409	19.241	6.081	
C	29	0.120	-0.193	-0.313	20.483	5.829	
D	131	0.541	0.513	0.645	29.267	7.361	**KEY**
Omit	5	0.021	-0.095	-0.272	15.000	7.778	

### **Item 42 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
42	42	C	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.281	0.173	0.230	0.160	0.214	0.818

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	43	0.178	-0.173	-0.253	21.860	7.269	
B	110	0.455	0.073	0.092	25.473	7.248	
C	68	0.281	0.160	0.214	27.882	8.334	**KEY**
D	19	0.079	-0.097	-0.179	22.158	9.634	
Omit	2	0.008	-0.074	-0.298	9.000	4.243	

### **Item 43 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
43	43	B	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.459	0.273	0.343	0.223	0.281	0.817

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	70	0.289	0.013	0.017	24.814	8.559	
B	111	0.459	0.223	0.281	27.586	7.792	**KEY**
C	11	0.045	-0.140	-0.305	19.545	5.087	
D	47	0.194	-0.191	-0.275	21.553	6.801	
Omit	3	0.012	-0.065	-0.223	16.667	5.033	

### Item 44 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
44	44	B	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.554	0.355	0.447	0.303	0.382	0.815

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	26	0.107	-0.026	-0.044	23.962	8.469	
B	134	0.554	0.303	0.382	27.716	7.788	**KEY**
C	18	0.074	-0.260	-0.486	17.278	4.127	
D	61	0.252	-0.139	-0.190	22.656	7.176	
Omit	3	0.012	-0.070	-0.240	15.667	7.234	

### Item 45 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
45	45	B	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.393	0.268	0.340	0.202	0.257	0.817

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	70	0.289	-0.167	-0.222	22.629	6.499	
B	95	0.393	0.202	0.257	27.726	8.001	**KEY**
C	40	0.165	-0.095	-0.141	23.025	8.894	
D	32	0.132	0.083	0.131	26.406	7.894	
Omit	5	0.021	-0.069	-0.198	18.600	11.393	

**Item 46 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
46	46	A	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.537	0.509	0.639	0.451	0.566	0.811

**Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	130	0.537	0.451	0.566	28.862	7.611	**KEY**
B	94	0.388	-0.421	-0.536	20.426	6.408	
C	11	0.045	-0.018	-0.040	23.909	5.856	
D	3	0.012	-0.118	-0.405	16.333	5.508	
Omit	4	0.017	-0.015	-0.048	23.250	7.632	

**Item 47 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
47	47	A	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.252	0.279	0.379	0.266	0.362	0.816

**Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	61	0.252	0.266	0.362	29.508	8.777	**KEY**
B	127	0.525	-0.147	-0.184	23.748	7.483	
C	22	0.091	-0.094	-0.165	22.500	7.366	
D	29	0.120	-0.032	-0.052	24.172	6.985	
Omit	3	0.012	-0.031	-0.105	21.667	8.505	

**Item 48 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
48	48	D	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.558	0.378	0.476	0.335	0.421	0.814

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	35	0.145	-0.094	-0.146	22.743	7.362	
B	26	0.107	-0.166	-0.278	20.769	5.458	
C	43	0.178	-0.203	-0.299	21.093	7.910	
D	135	0.558	0.335	0.421	27.911	7.817	**KEY**
Omit	3	0.012	-0.025	-0.086	22.000	8.888	

### Item 49 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
49	49	C	Yes	4	Grammar & text relations	K, LR

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.190	0.054	0.078	0.062	0.090	0.820

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
A	107	0.442	0.191	0.240	26.645	7.560	Maroon	
B	52	0.215	-0.200	-0.281	21.846	7.188	Green	
C	46	0.190	0.062	0.090	26.957	9.430	Blue	**KEY**
D	30	0.124	-0.032	-0.052	24.233	6.912	Olive	
Omit	7	0.029	-0.094	-0.238	17.571	8.904		

### Item 50 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
50	50	D	Yes	4	Grammar & text relations	LP

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.124	0.187	0.301	0.177	0.285	0.817

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	45	0.186	-0.040	-0.059	24.311	6.175	
B	89	0.368	0.042	0.053	25.427	7.602	
C	73	0.302	-0.097	-0.127	23.808	8.363	
D	30	0.124	0.177	0.285	29.767	9.594	**KEY**
Omit	5	0.021	-0.078	-0.222	17.800	9.706	

**Item 51 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
51	51	B	Yes	4	Grammar & text relations	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.488	0.378	0.474	0.234	0.294	0.816

**Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	54	0.223	-0.134	-0.187	22.630	7.517	
B	118	0.488	0.234	0.294	27.534	8.429	**KEY**
C	26	0.107	-0.135	-0.225	21.538	8.110	
D	40	0.165	-0.022	-0.033	24.225	6.023	
Omit	4	0.017	-0.056	-0.173	19.250	6.292	

**Item 52 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
52	52	B	Yes	4	Grammar & text relations	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.620	0.416	0.531	0.328	0.418	0.814

**Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	37	0.153	-0.119	-0.182	22.270	6.590	
B	150	0.620	0.328	0.418	27.527	7.994	**KEY**
C	37	0.153	-0.231	-0.352	20.189	5.456	
D	13	0.054	-0.056	-0.117	22.615	10.308	
Omit	5	0.021	-0.084	-0.239	16.600	4.980	

**Item 53 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
53	53	C	Yes	4	Grammar & text relations	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.612	0.368	0.468	0.377	0.480	0.813

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	31	0.128	-0.173	-0.276	20.935	6.455	
B	34	0.140	-0.194	-0.303	20.706	7.872	
C	148	0.612	0.377	0.480	27.872	7.406	**KEY**
D	20	0.083	-0.152	-0.275	20.500	8.389	
Omit	9	0.037	-0.059	-0.137	21.000	7.433	

### Item 54 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
54	54	D	Yes	4	Grammar & text relations	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.293	0.369	0.488	0.310	0.410	0.815

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	43	0.178	-0.247	-0.362	20.605	6.994	
B	77	0.318	0.053	0.069	25.429	7.797	
C	42	0.174	-0.141	-0.208	22.381	6.317	
D	71	0.293	0.310	0.410	29.634	7.940	**KEY**
Omit	9	0.037	-0.063	-0.148	21.000	7.433	

### Item 55 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
55	55	B	Yes	4	Grammar & text relations	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.322	0.373	0.486	0.238	0.310	0.816

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	40	0.165	-0.147	-0.220	22.150	7.413	
B	78	0.322	0.238	0.310	28.538	7.754	**KEY**
C	84	0.347	-0.001	-0.002	24.774	7.995	
D	24	0.099	-0.068	-0.116	23.167	7.346	
Omit	16	0.066	-0.095	-0.183	20.500	7.941	

### **Item 56 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
56	56	A	Yes	4	Grammar & text relations	LR

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.413	0.365	0.461	0.124	0.156	0.819

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	100	0.413	0.124	0.156	26.880	8.061	**KEY**
B	36	0.149	-0.047	-0.071	23.806	8.720	
C	50	0.207	-0.009	-0.012	24.560	6.923	
D	42	0.174	-0.032	-0.047	24.143	8.498	
Omit	14	0.058	-0.082	-0.166	20.714	8.278	

### **Item 57 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
57	57	B	Yes	4	Grammar & text relations	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.467	0.385	0.483	0.250	0.314	0.816

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	44	0.182	-0.081	-0.119	23.273	7.190	
B	113	0.467	0.250	0.314	27.770	8.093	**KEY**
C	44	0.182	-0.116	-0.170	22.682	7.618	
D	21	0.087	-0.075	-0.133	22.714	7.128	
Omit	20	0.083	-0.069	-0.124	22.000	8.633	

### **Item 58 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
58	58	A	Yes	4	Grammar & text relations	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.306	0.251	0.330	0.179	0.235	0.818



### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	74	0.306	0.179	0.235	27.959	6.939	**KEY**
B	48	0.198	0.098	0.140	26.375	8.462	
C	75	0.310	-0.135	-0.177	23.200	8.190	
D	25	0.103	-0.099	-0.167	22.480	7.611	
Omit	20	0.083	-0.072	-0.131	22.000	8.633	

### Item 59 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
59	59	A	Yes	4	Grammar & text relations	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.310	0.361	0.473	0.281	0.368	0.815

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	75	0.310	0.281	0.368	29.133	8.192	**KEY**
B	46	0.190	-0.196	-0.283	21.587	6.490	
C	43	0.178	-0.088	-0.129	23.302	7.265	
D	58	0.240	-0.005	-0.008	24.724	7.860	
Omit	20	0.083	-0.044	-0.079	23.150	8.456	

### Item 60 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
60	60	D	Yes	4	Grammar & text relations	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
242	0.405	0.376	0.476	0.209	0.265	0.817

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	39	0.161	0.018	0.026	25.026	8.229	
B	44	0.182	-0.121	-0.177	22.659	6.799	
C	39	0.161	-0.108	-0.162	22.744	7.133	
D	98	0.405	0.209	0.265	27.724	8.361	**KEY**
Omit	22	0.091	-0.055	-0.096	22.727	8.276	

## **Annexure E**

### **Iteman 4.3 analysis of the second test**



# ***Classical Item and Test Analysis Report***

## ***TALA-like test***

***Report created on 2014/08/28***

***Iteman: Software for Classical Analysis***

***Copyright © 2013 - Assessment Systems Corporation***



## Introduction

This report provides the results of a classical item and test analysis by the computer program Iteman Version 4.3 (Assessment Systems Corporation, 2013) for User Test 1. The output is divided into three sections:

1. Specifications
2. Summary statistics
3. Item-by-item results.

The statistical output is also recorded in a comma-separated value (CSV) file of the same name.

## Specifications

The Windows paths for the input files used in this analysis were:

C:\AlbertDocuments\ICELDA\ice01171.txt  
C:\AlbertDocuments\ICELDA\ice01172.txt

The Windows paths for the output files produced by this analysis were:

C:\AlbertDocuments\ICELDA\ice01172.rtf  
C:\AlbertDocuments\ICELDA\ice01172.csv  
C:\AlbertDocuments\ICELDA\ice01172 Scores.csv

Table 1 presents the specifications and basic information concerning the analysis. This provides important documentation of the setup of the program for historical purposes.

**Table 1: Specifications**

Specification	Value	Specification	Value
Number of examinees	240	Total Items	60
Scored Items	60	Pretest Items	0
Multiple Choice Items	60	Polytomous Items	0
Number of domains	5	External scores	No
Minimum P	0.15	Maximum P	0.84
Minimum item mean	0.00	Maximum item mean	15.00
Minimum item correlation	0.15	Maximum item correlation	1.00
Responses begin in column	1	Omit character	X
Not Admin character	N	Produce quantile tables	Yes
Correct for spuriousness	Yes	Produce quantile plots	Yes
Save data matrix	No	Include omit codes in matrix	N/A
Scaled score setting 2	N/A	Dichotomous Classification	Yes
Classify based on	Total Score	Cutpoint	1.000

## Summary statistics

Table 2 presents the summary statistics of the test, for all items, scored items only, and for each domain (content area). Definitions of these statistics are found in the Iteman manual.

**Table 2: Summary statistics**

Score	Items	Mean	SD	Min Score	Max Score	Mean P	Mean Rpbis
All items	60	33.233	10.556	9	54	0.554	0.334
Scored Items	60	33.233	10.556	9	54	0.554	0.334
Scrambled text	5	2.513	1.998	0	5	0.502	0.483
Vocabulary knowledge	10	6.233	1.749	0	9	0.623	0.250
Interpreting graphs & visual information	8	5.004	2.438	0	8	0.626	0.493
Text comprehension	25	13.654	4.445	3	23	0.546	0.293
Grammar & text relations	12	5.829	2.839	0	12	0.486	0.319

Table 3 presents a reliability analysis of the tests. Alpha (also known as KR-20) is the most commonly used index of reliability, and is therefore used to calculate the standard error of measurement (SEM) on the raw score scale. Also presented are three configurations of split-half reliability, first as uncorrected correlations, and then as Spearman-Brown (S-B) corrected correlations. This is because an uncorrected split-half correlation is referenced to a "test" that only contains half as many items as the full test, and therefore underestimates reliability.

The cutscore on this exam was 1.000, producing a pass rate of 100.0%. The Livingston index of classification consistency at the cut-score was 0.990.

**Table 3: Reliability**

Score	Alpha	SEM	Split-Half (Random)	Split-Half (First-Last)	Split-Half (Odd-Even)	S-B Random	S-B First-Last	S-B Odd-Even
Scored items	0.897	3.394	0.846	0.683	0.843	0.916	0.812	0.915
Scrambled text	0.865	0.734	0.660	0.773	0.769	0.795	0.872	0.869
Vocabulary knowledge	0.444	1.304	0.413	0.299	0.354	0.585	0.461	0.523
Interpreting graphs & visual information	0.801	1.087	0.616	0.659	0.671	0.763	0.794	0.803
Text comprehension	0.751	2.216	0.576	0.617	0.602	0.731	0.763	0.751
Grammar & text relations	0.707	1.538	0.463	0.357	0.779	0.633	0.526	0.876

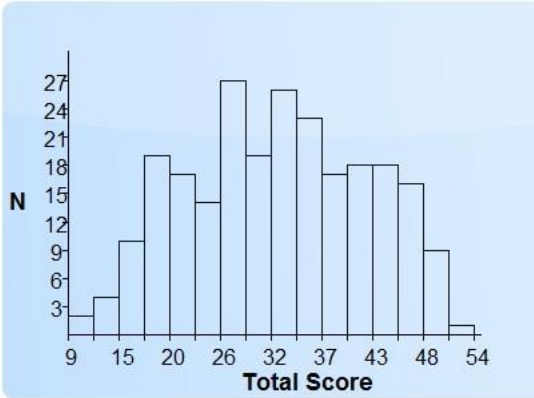
Table 4 presents the item statistics and flags for the item(s) that were flagged during the analysis

**Table 4: Summary Statistics for the Flagged Items**

Item ID	P / Item Mean	R	Flag(s)
6	0.596	-0.434	K, LR
7	0.904	0.235	HP
12	0.929	0.320	HP
13	0.129	-0.118	K, LP, LR
25	0.517	0.111	LR
27	0.658	0.135	LR
28	0.179	-0.045	K, LR
32	0.146	-0.129	K, LP, LR
45	0.429	0.114	K, LR
52	0.383	0.091	K, LR

Figure 1 displays the distribution of the raw scores for the scored items across all domains. Table 5 displays the frequency distribution for total score shown in Figure 1.

**Figure 1: Total score for the scored items**

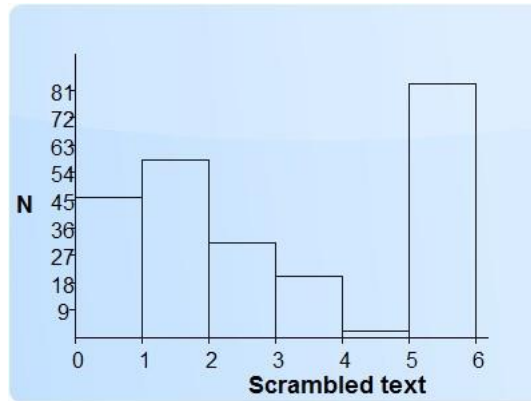


**Table 5: Frequency Distribution for Total Score**

Range	Frequency
8 to 11	2
12 to 14	4
15 to 17	10
18 to 20	19
21 to 23	17
24 to 26	14
27 to 29	27
30 to 32	19
33 to 35	26
36 to 38	23
39 to 41	17
42 to 44	18
45 to 47	18
48 to 50	16
51 to 53	9
54	1

Figure 2 displays the distribution of the raw scores for Scrambled text. Table 6 displays the frequency distribution of domain scores shown in Figure 2.

**Figure 2: Raw scores for Scrambled text**

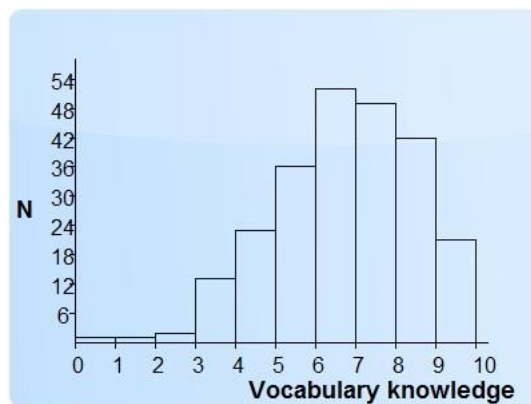


**Table 6: Frequency Distribution for Scrambled text**

Score	Frequency
0	46
1	58
2	31
3	20
4	2
5	83

Figure 3 displays the distribution of the raw scores for Vocabulary knowledge. Table 7 displays the frequency distribution of domain scores shown in Figure 3.

**Figure 3: Raw scores for Vocabulary knowledge**

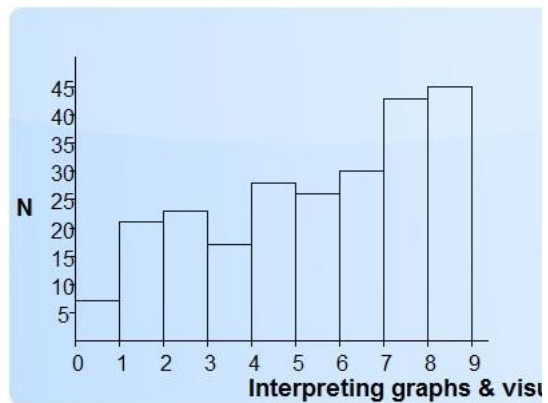


**Table 7: Frequency Distribution for Vocabulary knowledge**

Score	Frequency
0	1
1	1
2	2
3	13
4	23
5	36
6	52
7	49
8	42
9	21

Figure 4 displays the distribution of the raw scores for Interpreting graphs & visual information. Table 8 displays the frequency distribution of domain scores shown in Figure 4.

**Figure 4: Raw scores for Interpreting graphs & visual information**



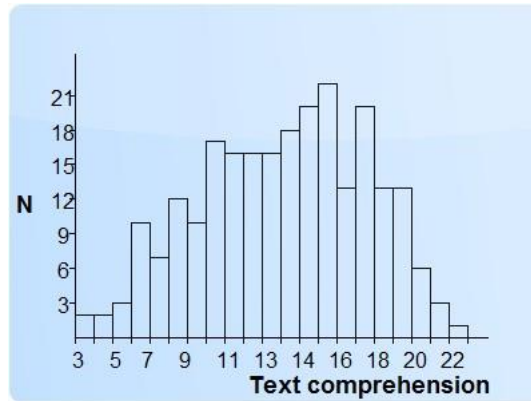
**Table 8: Frequency Distribution for Interpreting graphs & visual information**

Score	Frequency
0	7
1	21
2	23
3	17
4	28
5	26
6	30
7	43
8	45



Figure 5 displays the distribution of the raw scores for Text comprehension. Table 9 displays the frequency distribution of domain scores shown in Figure 5.

**Figure 5: Raw scores for Text comprehension**

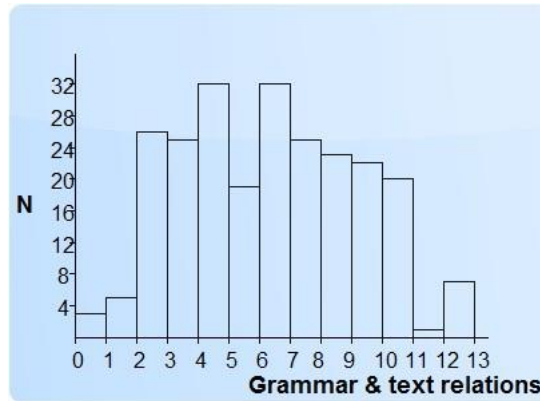


**Table 9: Frequency Distribution for Text comprehension**

Range	Frequency
2 to 3	2
4	2
5	3
6	10
7	7
8	12
9	10
10	17
11	16
12	16
13	16
14	18
15	20
16	22
17	13
18	20
19	13
20	13
21	6
22	3
23	1

Figure 6 displays the distribution of the raw scores for Grammar & text relations. Table 10 displays the frequency distribution of domain scores shown in Figure 6.

**Figure 6: Raw scores for Grammar & text relations**



**Table 10: Frequency Distribution for Grammar & text relations**

Score	Frequency
0	3
1	5
2	26
3	25
4	32
5	19
6	32
7	25
8	23
9	22
10	20
11	1
12	7

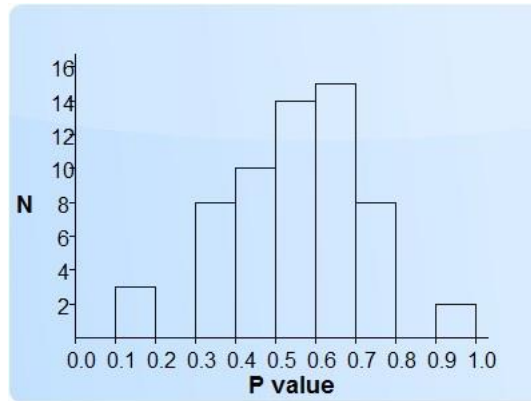
Table 11 displays the correlations of domain scores.

**Table 11: Correlations for Domain Scores**

Domain	Scrambled text	Vocabulary knowledge	Interpreting graphs & visual information	Text comprehension	Grammar & text relations
Scrambled text	1.000	0.418	0.497	0.430	0.331
Vocabulary knowledge	0.418	1.000	0.490	0.572	0.366
Interpreting graphs & visual information	0.497	0.490	1.000	0.636	0.521
Text comprehension	0.430	0.572	0.636	1.000	0.545
Grammar & text relations	0.331	0.366	0.521	0.545	1.000

Figure 7 displays the distribution of the P values for the dichotomously scored items (correct/incorrect). Table 12 displays the frequency distribution of the P values shown in Figure 7.

**Figure 7: P values for the scored items**

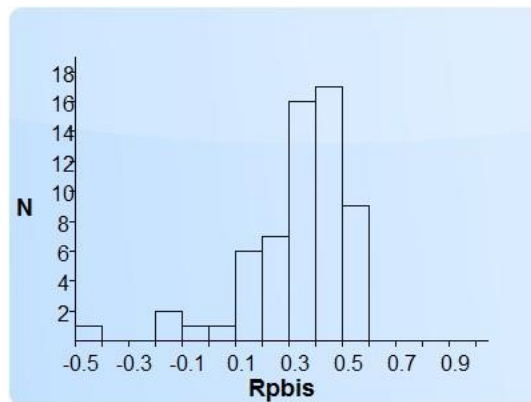


**Table 12: Frequency Distribution for the P values**

Score	Frequency
0.0 to 0.1	0
0.1 to 0.2	3
0.2 to 0.3	0
0.3 to 0.4	8
0.4 to 0.5	10
0.5 to 0.6	14
0.6 to 0.7	15
0.7 to 0.8	8
0.8 to 0.9	0
0.9 to 1.0	2

Figure 8 displays the distribution of the Point-Biserial Correlations for the dichotomously scored items (correct/incorrect). Table 13 displays the frequency distribution of the Point-Biserial correlations shown in Figure 8.

**Figure 8: Rpbis for the scored items**



**Table 13: Frequency Distribution for the Rpbis**

Score	Frequency
-0.5 to -0.4	1
-0.4 to -0.3	0
-0.3 to -0.2	0
-0.2 to -0.1	2
-0.1 to 0.0	1
0.0 to 0.1	1
0.1 to 0.2	6
0.2 to 0.3	7
0.3 to 0.4	16
0.4 to 0.5	17
0.5 to 0.6	9
0.6 to 0.7	0
0.7 to 0.8	0
0.8 to 0.9	0
0.9 to 1.0	0

Figure 9 displays the scatterplot of P (difficulty) by Rpbis (discrimination) for the dichotomously scored items (correct/incorrect).

**Figure 9: P by Rpbis**

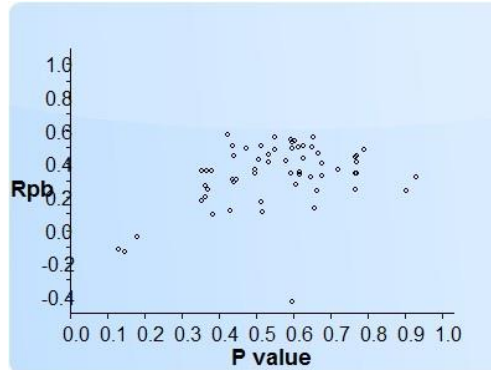
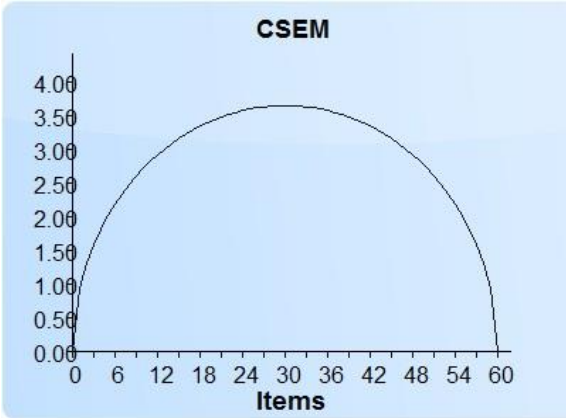


Figure 10 displays a graph of the Conditional Standard Error of Measurement (CSEM) Formula IV. The CSEM at the cutscore of 1.000 equaled 0.935.



## ***Item-by-item results***

The following section presents the item-by-item results of the analysis. Each item has several tables and a figure. The figure, called a quantile plot, shows the proportion of examinees selecting each option, for consecutive segments of the examinees as ranked by score. The key thing to evaluate in this figure is that the line for the correct answer has a positive slope (goes up from left to right), which means that examinees with higher scores tend to answer correctly more often. Conversely, the lines for the incorrect options, called distractors, should have a negative slope. Note, however, that the use of a small number of groups (e.g., 3 or fewer) oversimplifies the graph, so that items which are very difficult or very easy (that is, discriminating in only the top or bottom 20% of examinees) might appear to have poor quantile plots and classical statistics. For such items, item response theory presents significant advantages in analysis

There are four tables presented for each item.

1. Item information table: records the information supplied by the control file (or Iteman 3 header) for this item.
2. Item statistics table: overall item statistics.
3. Option statistics: detailed statistics for each item, which helps diagnose issues in items with poor statistics.
4. Quantile plot data: the values used to create the quantile plot.

The item statistics table presents overall item statistics in the first row of numbers. The two most important item-level statistics for dichotomously scored (correct/incorrect) items are the P value and the point-biserial correlation, which represent the difficulty and discrimination of the item, respectively. For polytomously scored (rating scale or partial credit) items, the difficulty is represented by the mean (average) item score, while the discrimination is represented by a Pearson r correlation.

The P value is the proportion of examinees that answered an item in the keyed direction. P ranges from 0 to 1. A high value (0.95) means that an item is easy, a low value (0.25) means that the item is difficult. The point-biserial correlation (Rpbis) is a measure of the discriminating, or differentiating, power of the item. Rpbis ranges from -1 to 1. A negative Rpbis is indicative of a bad item as lower scoring examinees are more likely than higher scoring examinees to respond in the keyed direction.

For rating scale or partial credit items, the mean item score ranges from the minimum to the maximum of the scale. For example, if the item has a rating scale of 1 to 5, the possible range for the mean is 1 to 5. The Pearson r is similar to the Rpbis in that it ranges from -1 to 1, with a positive r indicating that the item correlates well with total score.

The option statistics table presents statistics for each individual option (alternative). The key thing to examine in this portion of the table is that no distractors have a higher Rpbis than the correct answer. That indicates that higher scoring examinees are selecting the incorrect answer, which therefore might be arguably correct.

The quantile plot data table simply presents the values calculated to create the quantile plot. Because it contains the same information, the quantile plot itself presents a useful picture of the item's performance, but this table can be used to examine that performance in detail to help diagnose possible issues.

### Item 1 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
1	1	B	Yes	5	Scrambled text	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.550	0.676	0.850	0.558	0.702	0.892

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	17	0.071	-0.217	-0.411	24.647	6.873	
B	132	0.550	0.558	0.702	38.848	8.852	**KEY**
C	7	0.029	-0.150	-0.379	23.857	5.273	
D	2	0.008	-0.046	-0.186	27.500	3.536	
E	80	0.333	-0.389	-0.504	27.063	8.789	
Omit	2	0.008	-0.058	-0.231	21.000	4.243	

### Item 2 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
2	2	D	Yes	5	Scrambled text	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.508	0.691	0.866	0.428	0.537	0.894

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	53	0.221	-0.258	-0.361	27.736	8.700	
B	16	0.067	-0.178	-0.343	25.875	7.848	
C	16	0.067	-0.137	-0.265	27.438	7.321	
D	122	0.508	0.428	0.537	38.057	10.269	**KEY**
E	28	0.117	-0.052	-0.085	31.250	8.249	
Omit	5	0.021	-0.053	-0.152	27.000	9.274	

### Item 3 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
3	3	E	Yes	5	Scrambled text	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
	0.425	0.741	0.935	0.579	0.730	0.892

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	28	0.117	-0.219	-0.359	26.643	6.988	
B	27	0.113	-0.117	-0.194	29.444	7.638	
C	33	0.138	-0.215	-0.337	27.303	9.892	
D	40	0.167	-0.233	-0.348	27.475	8.575	
E	102	0.425	0.579	0.730	40.686	8.251	**KEY**
Omit	10	0.042	-0.060	-0.134	28.500	10.763	

### Item 4 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
4	4	A	Yes	5	Scrambled text	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.438	0.722	0.909	0.507	0.638	0.893

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	105	0.438	0.507	0.638	39.686	9.567	**KEY**
B	41	0.171	-0.331	-0.491	25.317	8.353	
C	38	0.158	-0.076	-0.115	31.000	8.917	
D	41	0.171	-0.175	-0.259	28.854	7.767	
E	12	0.050	-0.130	-0.274	27.000	7.508	
Omit	3	0.013	-0.031	-0.106	28.667	1.528	

### Item 5 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
5	5	C	Yes	5	Scrambled text	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.592	0.578	0.731	0.342	0.433	0.895



### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	32	0.133	-0.022	-0.035	32.063	9.270	
B	23	0.096	-0.219	-0.379	25.696	8.536	
C	142	0.592	0.342	0.433	36.577	10.270	**KEY**
D	30	0.125	-0.195	-0.314	27.300	8.856	
E	9	0.037	-0.135	-0.314	25.556	6.839	
Omit	4	0.017	-0.031	-0.097	29.000	11.402	

### Item 6 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
6	6	B	Yes	4	Vocabulary knowledge	K, LR

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.596	-0.164	-0.208	-0.434	-0.550	0.903

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	12	0.050	0.218	0.461	42.833	5.638	
B	143	0.596	-0.434	-0.550	29.804	9.643	**KEY**
C	66	0.275	0.390	0.522	39.424	8.658	
D	14	0.058	0.027	0.054	33.786	11.033	
Omit	5	0.021	-0.066	-0.187	25.000	16.628	

### Item 7 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
7	7	C	Yes	4	Vocabulary knowledge	HP

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.904	0.442	0.765	0.235	0.407	0.896

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	3	0.013	-0.115	-0.395	21.667	9.866	
B	11	0.046	-0.194	-0.422	23.091	6.008	
C	217	0.904	0.235	0.407	34.129	10.398	**KEY**
D	5	0.021	0.040	0.114	35.200	7.155	
Omit	4		-0.088	-0.272	18.750	7.500	

### **Item 8 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
8	8	C	Yes	4	Vocabulary knowledge	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.767	0.504	0.696	0.438	0.606	0.894

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	29	0.121	-0.205	-0.332	26.759	8.605	
B	10	0.042	-0.199	-0.447	22.600	7.397	
C	184	0.767	0.438	0.606	35.962	9.560	**KEY**
D	14	0.058	-0.275	-0.553	21.071	9.856	
Omit	3	0.013	-0.071	-0.243	20.667	0.577	

### **Item 9 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
9	9	D	Yes	4	Vocabulary knowledge	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.767	0.475	0.657	0.341	0.471	0.895

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
A	36	0.150	-0.339	-0.519	24.111	8.106	Maroon	
B	9	0.037	-0.151	-0.350	24.556	8.575	Green	
C	10	0.042	0.051	0.114	35.000	8.919	Blue	
D	184	0.767	0.341	0.471	35.413	10.050	Olive	**KEY**
Omit	1	0.004	-0.040	-0.209	21.000	0.000		

### **Item 10 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
10	10	A	Yes	4	Vocabulary knowledge	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.625	0.464	0.593	0.432	0.552	0.894

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	150	0.625	0.432	0.552	37.053	9.487	**KEY**
B	50	0.208	-0.304	-0.431	26.500	8.491	
C	19	0.079	-0.089	-0.163	29.474	10.002	
D	16	0.067	-0.123	-0.237	27.875	10.012	
Omit	5	0.021	-0.105	-0.301	17.400	4.930	

### Item 11 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
11	11	C	Yes	4	Vocabulary knowledge	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.446	0.323	0.407	0.301	0.379	0.895

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	84	0.350	-0.171	-0.221	30.369	10.321	
B	19	0.079	-0.230	-0.420	24.684	7.903	
C	107	0.446	0.301	0.379	37.262	10.101	**KEY**
D	28	0.117	-0.036	-0.060	31.750	7.820	
Omit	2	0.008	0.041	0.162	40.000	18.385	

### Item 12 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
12	12	B	Yes	4	Vocabulary knowledge	HP

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.929	0.381	0.721	0.320	0.605	0.896

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
A	7	0.029	-0.238	-0.602	18.000	3.958	Maroon	
B	223	0.929	0.320	0.605	34.224	10.220	Green	**KEY**
C	10	0.042	-0.210	-0.471	21.800	6.844	Blue	
D	0	0.000	--	--	--	--	Olive	

**Item 13 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
13	13	B	Yes	4	Vocabulary knowledge	K, LP, LR

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.129	-0.112	-0.179	-0.118	-0.188	0.899

**Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	156	0.650	0.454	0.584	36.615	9.704	
B	31	0.129	-0.118	-0.188	30.871	10.489	**KEY**
C	23	0.096	-0.192	-0.333	26.870	8.699	
D	29	0.121	-0.353	-0.572	23.069	6.100	
Omit	1	0.004	-0.043	-0.224	20.000	0.000	

**Item 14 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
14	14	D	Yes	4	Vocabulary knowledge	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.596	0.382	0.483	0.496	0.628	0.893

**Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	53	0.221	-0.377	-0.527	25.377	8.828	
B	11	0.046	-0.113	-0.245	27.364	7.646	
C	31	0.129	-0.166	-0.264	28.226	8.590	
D	143	0.596	0.496	0.628	37.825	9.289	**KEY**
Omit	2	0.008	-0.051	-0.203	23.000	4.243	

**Item 15 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
15	15	A	Yes	4	Vocabulary knowledge	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.475	0.450	0.565	0.490	0.615	0.893

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	114	0.475	0.490	0.615	39.044	8.698	**KEY**
B	3	0.013	-0.166	-0.568	17.667	7.024	
C	82	0.342	-0.235	-0.303	29.415	9.504	
D	41	0.171	-0.305	-0.453	25.854	8.332	

### Item 16 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
16	16	C	Yes	4	Interpreting graphs & visual information	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.512	0.533	0.668	0.504	0.632	0.893

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	18	0.075	-0.196	-0.365	25.667	7.941	
B	42	0.175	-0.221	-0.325	27.810	8.454	
C	123	0.512	0.504	0.632	38.764	9.528	**KEY**
D	57	0.237	-0.274	-0.377	27.684	8.337	

### Item 17 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
17	17	C	Yes	4	Interpreting graphs & visual information	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.625	0.582	0.743	0.508	0.649	0.893

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	30	0.125	-0.245	-0.393	25.967	7.476	
B	46	0.192	-0.335	-0.484	25.543	7.788	
C	150	0.625	0.508	0.649	37.647	9.617	**KEY**
D	12	0.050	-0.118	-0.249	27.333	8.489	
Omit	2	0.008	-0.049	-0.195	23.500	4.950	

### **Item 18 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
18	18	C	Yes	4	Interpreting graphs & visual information	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.654	0.651	0.840	0.560	0.722	0.893

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	13	0.054	-0.094	-0.194	28.538	11.450	
B	11	0.046	-0.042	-0.090	30.636	10.519	
C	157	0.654	0.560	0.722	37.745	8.682	**KEY**
D	57	0.237	-0.545	-0.750	22.579	5.892	
Omit	2	0.008	-0.030	-0.121	27.500	17.678	

### **Item 19 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
19	19	B	Yes	4	Interpreting graphs & visual information	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.613	0.610	0.777	0.496	0.632	0.893

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	48	0.200	-0.212	-0.303	28.271	8.152	
B	147	0.613	0.496	0.632	37.673	9.431	**KEY**
C	20	0.083	-0.239	-0.430	24.500	7.359	
D	23	0.096	-0.291	-0.504	23.435	8.717	
Omit	2	0.008	-0.038	-0.152	26.000	7.071	

### **Item 20 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
20	20	B	Yes	4	Interpreting graphs & visual information	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.771	0.566	0.784	0.411	0.570	0.894

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	9	0.037	-0.136	-0.316	25.333	7.000	
B	185	0.771	0.411	0.570	35.778	10.247	**KEY**
C	31	0.129	-0.320	-0.510	23.871	6.103	
D	14	0.058	-0.149	-0.299	26.286	7.258	
Omit	1	0.004	-0.040	-0.209	21.000	0.000	

### Item 21 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
21	21	B	Yes	4	Interpreting graphs & visual information	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.792	0.596	0.843	0.487	0.689	0.894

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	12	0.050	-0.179	-0.377	24.417	10.587	
B	190	0.792	0.487	0.689	36.016	9.489	**KEY**
C	29	0.121	-0.336	-0.544	23.103	5.966	
D	9	0.037	-0.259	-0.603	18.889	6.528	

### Item 22 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
22	22	A	Yes	4	Interpreting graphs & visual information	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.442	0.458	0.576	0.446	0.562	0.894

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	106	0.442	0.446	0.562	38.953	10.359	**KEY**
B	65	0.271	-0.179	-0.240	29.769	9.046	
C	50	0.208	-0.223	-0.316	28.320	7.670	
D	19	0.079	-0.191	-0.349	26.105	7.187	

**Item 23 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
23	23	C	Yes	4	Interpreting graphs & visual information	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.596	0.595	0.753	0.532	0.674	0.893

**Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	29	0.121	-0.333	-0.539	23.448	8.534	
B	34	0.142	-0.236	-0.367	26.676	8.271	
C	143	0.596	0.532	0.674	38.126	9.014	**KEY**
D	31	0.129	-0.193	-0.307	27.516	8.156	
Omit	3	0.013	-0.034	-0.118	28.000	9.644	

**Item 24 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
24	24	C	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.767	0.266	0.368	0.246	0.340	0.896

**Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	13	0.054	-0.190	-0.390	24.231	9.400	
B	23	0.096	-0.064	-0.110	30.435	9.834	
C	184	0.767	0.246	0.340	34.880	10.177	**KEY**
D	19	0.079	-0.135	-0.247	27.684	10.781	
Omit	1	0.004	-0.047	-0.246	17.000	0.000	

**Item 25 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
25	25	B	Yes	4	Text comprehension	LR



N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.517	0.216	0.271	0.111	0.139	0.898

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	82	0.342	0.024	0.031	33.061	10.327	
B	124	0.517	0.111	0.139	34.839	10.461	**KEY**
C	25	0.104	-0.115	-0.194	29.200	9.743	
D	8	0.033	-0.142	-0.342	24.750	10.754	
Omit	1	0.004	-0.047	-0.247	17.000	0.000	

### Item 26 information and statistics

Seq. ID	Key	Scored	Num Options	Domain	Flags
26 26	A	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.771	0.487	0.675	0.445	0.618	0.894

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	185	0.771	0.445	0.618	35.968	9.872	**KEY**
B	19	0.079	-0.252	-0.462	23.579	8.662	
C	10	0.042	-0.256	-0.574	19.800	4.442	
D	24	0.100	-0.215	-0.368	25.792	6.627	
Omit	2	0.008	-0.024	-0.096	28.500	3.536	

### Item 27 information and statistics

Seq. ID	Key	Scored	Num Options	Domain	Flags
27 27	C	Yes	4	Text comprehension	LR

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.658	0.165	0.213	0.135	0.174	0.897

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	7	0.029	-0.166	-0.421	22.571	7.185	
B	36	0.150	-0.118	-0.181	29.639	11.455	
C	158	0.658	0.135	0.174	34.589	9.854	**KEY**
D	37	0.154	0.032	0.048	33.351	11.617	
Omit	2	0.008	-0.037	-0.149	26.000	11.314	

**Item 28 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
28	28	D	Yes	4	Text comprehension	K, LR

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.179	-0.044	-0.065	-0.045	-0.066	0.898

**Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	92	0.383	0.095	0.121	34.326	9.652	
B	66	0.275	0.033	0.044	33.621	11.178	
C	38	0.158	-0.114	-0.172	30.289	10.384	
D	43	0.179	-0.045	-0.066	33.047	11.542	**KEY**
Omit	1	0.004	-0.024	-0.128	27.000	0.000	

**Item 29 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
29	29	C	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.646	0.394	0.507	0.321	0.412	0.895

**Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	6	0.025	-0.184	-0.493	20.667	9.245	
B	34	0.142	-0.297	-0.461	25.029	8.178	
C	155	0.646	0.321	0.412	36.045	10.249	**KEY**
D	45	0.188	-0.054	-0.078	31.422	8.748	

**Item 30 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
30	30	D	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.675	0.438	0.570	0.403	0.525	0.894

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	20	0.083	-0.298	-0.538	22.350	6.746	
B	31	0.129	-0.195	-0.311	27.323	8.972	
C	27	0.113	-0.129	-0.214	28.815	9.115	
D	162	0.675	0.403	0.525	36.444	9.876	**KEY**

### Item 31 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
31	31	C	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.604	0.563	0.714	0.541	0.686	0.893

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	30	0.125	-0.225	-0.361	26.533	9.123	
B	47	0.196	-0.359	-0.516	25.170	7.968	
C	145	0.604	0.541	0.686	38.110	8.846	**KEY**
D	12	0.050	-0.126	-0.266	27.000	8.697	
Omit	6	0.025	-0.078	-0.209	24.500	10.654	

### Item 32 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
32	32	D	Yes	4	Text comprehension	K, LP, LR

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.146	-0.110	-0.170	-0.129	-0.199	0.899

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	159	0.662	0.254	0.329	35.000	10.480	
B	23	0.096	-0.162	-0.281	27.826	7.785	
C	19	0.079	-0.046	-0.085	31.421	10.112	
D	35	0.146	-0.129	-0.199	30.800	11.260	**KEY**
Omit	4	0.017	-0.068	-0.209	24.000	4.082	

### Item 33 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
33	33	B	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.650	0.495	0.637	0.497	0.640	0.893

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	33	0.138	-0.381	-0.597	22.788	6.740	
B	156	0.650	0.497	0.640	37.327	9.217	**KEY**
C	40	0.167	-0.159	-0.238	28.925	9.250	
D	5	0.021	-0.187	-0.534	19.400	7.537	
Omit	6	0.025	-0.078	-0.208	24.500	6.775	

### Item 34 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
34	34	A	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.496	0.344	0.432	0.363	0.455	0.895

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	119	0.496	0.363	0.455	37.513	10.258	**KEY**
B	51	0.212	-0.188	-0.265	29.000	9.786	
C	19	0.079	-0.233	-0.427	24.526	8.978	
D	47	0.196	-0.058	-0.083	31.532	7.824	
Omit	4	0.017	-0.081	-0.249	21.250	3.403	

### **Item 35 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
35	35	D	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.608	0.378	0.481	0.278	0.353	0.896

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	11	0.046	-0.215	-0.467	22.455	8.287	
B	53	0.221	-0.073	-0.102	31.208	10.723	
C	23	0.096	-0.168	-0.292	27.261	9.992	
D	146	0.608	0.278	0.353	35.938	9.702	**KEY**
Omit	7	0.029	-0.045	-0.113	28.714	11.101	

### **Item 36 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
36	36	B	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.667	0.472	0.612	0.465	0.603	0.894

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	15	0.063	-0.272	-0.535	21.733	6.250	
B	160	0.667	0.465	0.603	36.950	9.782	**KEY**
C	31	0.129	-0.250	-0.397	25.903	7.846	
D	32	0.133	-0.177	-0.279	27.938	8.234	
Omit	2	0.008	-0.059	-0.234	20.500	0.707	

### **Item 37 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
37	37	C	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.371	0.198	0.253	0.245	0.314	0.896

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	59	0.246	-0.147	-0.201	30.186	10.357	
B	53	0.221	0.033	0.046	33.509	11.015	
C	89	0.371	0.245	0.314	37.180	10.292	**KEY**
D	37	0.154	-0.172	-0.262	28.676	7.424	
Omit	2	0.008	-0.045	-0.182	24.500	4.950	

### Item 38 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
38	38	D	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.579	0.453	0.572	0.419	0.529	0.894

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	22	0.092	-0.204	-0.357	26.045	9.353	
B	31	0.129	-0.075	-0.120	30.645	8.894	
C	44	0.183	-0.283	-0.413	26.500	8.746	
D	139	0.579	0.419	0.529	37.331	9.659	**KEY**
Omit	4	0.017	-0.064	-0.197	24.500	11.790	

### Item 39 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
39	39	B	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.617	0.386	0.492	0.338	0.430	0.895

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	18	0.075	-0.292	-0.544	22.000	6.499	
B	148	0.617	0.338	0.430	36.372	9.824	**KEY**
C	43	0.179	-0.171	-0.250	28.837	9.808	
D	30	0.125	-0.047	-0.075	31.333	10.005	
Omit	1	0.004	-0.047	-0.248	17.000	0.000	

### **Item 40 information**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
40	40	D	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.617	0.421	0.536	0.348	0.443	0.895

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	16	0.067	-0.310	-0.598	20.625	6.152	
B	15	0.063	-0.222	-0.437	23.733	6.442	
C	54	0.225	-0.047	-0.065	31.722	9.279	
D	148	0.617	0.348	0.443	36.453	10.063	**KEY**
Omit	7	0.029	-0.072	-0.181	26.000	6.583	

### **Item 41 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
41	41	C	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.496	0.314	0.394	0.341	0.427	0.895

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	28	0.117	-0.294	-0.481	24.393	4.833	
B	21	0.087	-0.282	-0.501	23.333	8.071	
C	119	0.496	0.341	0.427	37.286	9.816	**KEY**
D	70	0.292	0.028	0.037	33.186	10.022	
Omit	2	0.008	-0.056	-0.224	21.500	6.364	

### **Item 42 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
42	42	D	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.367	0.348	0.445	0.355	0.455	0.895

**Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	28	0.117	-0.294	-0.482	24.500	10.024	
B	72	0.300	-0.197	-0.259	29.764	9.404	
C	48	0.200	0.027	0.038	33.417	9.906	
D	88	0.367	0.355	0.455	38.693	8.979	**KEY**
Omit	4	0.017	0.014	0.045	34.500	12.477	

**Item 43 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
43	43	B	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.512	0.264	0.331	0.168	0.211	0.897

**Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	57	0.237	0.052	0.071	33.684	10.128	
B	123	0.512	0.168	0.211	35.431	10.746	**KEY**
C	28	0.117	-0.109	-0.178	29.607	7.809	
D	31	0.129	-0.198	-0.316	27.355	10.131	
Omit	1	0.004	-0.040	-0.211	21.000	0.000	

**Item 44 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
44	44	C	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.354	0.311	0.400	0.354	0.456	0.895

**Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	31	0.129	-0.258	-0.410	25.968	6.785	
B	43	0.179	-0.267	-0.391	26.977	9.249	
C	85	0.354	0.354	0.456	38.824	10.347	**KEY**
D	80	0.333	0.051	0.066	33.625	9.148	
Omit	1	0.004	-0.041	-0.214	21.000	0.000	



### **Item 45 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
45	45	A	Yes	4	Text comprehension	K, LR

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.429	0.183	0.230	0.114	0.143	0.898

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	103	0.429	0.114	0.143	35.175	10.265	**KEY**
B	77	0.321	0.141	0.184	34.948	10.877	
C	34	0.142	-0.206	-0.321	27.500	8.853	
D	25	0.104	-0.150	-0.254	28.200	8.935	
Omit	1	0.004	-0.038	-0.200	22.000	0.000	

### **Item 46 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
46	46	C	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.721	0.458	0.611	0.362	0.483	0.895

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	13	0.054	-0.234	-0.483	22.385	7.859	
B	22	0.092	-0.199	-0.348	26.045	9.302	
C	173	0.721	0.362	0.483	35.844	9.710	**KEY**
D	32	0.133	-0.153	-0.242	28.469	10.562	

### **Item 47 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
47	47	B	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.438	0.346	0.435	0.308	0.388	0.895

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	76	0.317	0.034	0.045	33.316	8.930	
B	105	0.438	0.308	0.388	37.410	10.366	**KEY**
C	20	0.083	-0.275	-0.496	23.350	10.075	
D	39	0.163	-0.251	-0.377	26.897	7.783	

### Item 48 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
48	48	D	Yes	4	Text comprehension	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.771	0.404	0.560	0.342	0.474	0.895

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	25	0.104	-0.135	-0.228	28.360	9.831	
B	10	0.042	-0.146	-0.328	25.200	8.404	
C	17	0.071	-0.260	-0.492	22.706	8.275	
D	185	0.771	0.342	0.474	35.395	10.037	**KEY**
Omit	3	0.013	-0.039	-0.135	27.000	7.937	

### Item 49 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
49	49	B	Yes	4	Grammar & text relations	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.362	0.317	0.407	0.266	0.341	0.896

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	80	0.333	-0.041	-0.054	32.263	10.271	
B	87	0.362	0.266	0.341	37.529	10.685	**KEY**
C	51	0.212	-0.204	-0.287	28.804	8.699	
D	18	0.075	-0.076	-0.141	30.111	9.190	
Omit	4	0.017	-0.027	-0.084	29.750	12.121	

### **Item 50 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
50	50	D	Yes	4	Grammar & text relations	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.442	0.350	0.441	0.289	0.364	0.896

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	34	0.142	-0.161	-0.251	28.676	7.764	
B	45	0.188	-0.245	-0.355	27.511	9.080	
C	49	0.204	0.052	0.074	33.857	9.312	
D	106	0.442	0.289	0.364	37.160	10.865	**KEY**
Omit	6	0.025	-0.054	-0.145	27.500	10.330	

### **Item 51 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
51	51	A	Yes	4	Grammar & text relations	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.354	0.213	0.274	0.175	0.225	0.897

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	85	0.354	0.175	0.225	36.341	9.705	**KEY**
B	46	0.192	-0.261	-0.376	27.304	7.656	
C	87	0.362	0.102	0.131	34.287	11.427	
D	21	0.087	-0.091	-0.162	29.810	10.073	
Omit	1	0.004	-0.038	-0.202	22.000	0.000	

### **Item 52 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
52	52	C	Yes	4	Grammar & text relations	K, LR

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.383	0.163	0.207	0.091	0.116	0.898

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	80	0.333	0.173	0.225	35.413	11.887	
B	23	0.096	-0.105	-0.182	29.478	7.786	
C	92	0.383	0.091	0.116	35.054	9.750	**KEY**
D	43	0.179	-0.230	-0.337	27.698	8.568	
Omit	2	0.008	-0.045	-0.180	24.500	3.536	

### Item 53 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
53	53	D	Yes	4	Grammar & text relations	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.362	0.327	0.419	0.197	0.252	0.897

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	71	0.296	-0.069	-0.091	31.761	10.708	
B	66	0.275	-0.059	-0.079	31.879	9.173	
C	13	0.054	-0.140	-0.289	26.769	9.816	
D	87	0.362	0.197	0.252	36.586	10.857	**KEY**
Omit	3	0.013	-0.031	-0.106	28.667	5.859	

### Item 54 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
54	54	C	Yes	4	Grammar & text relations	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.379	0.454	0.579	0.359	0.457	0.895

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	74	0.308	-0.124	-0.163	30.932	9.440	
B	24	0.100	-0.157	-0.268	28.000	9.245	
C	91	0.379	0.359	0.457	38.593	10.591	**KEY**
D	50	0.208	-0.161	-0.227	29.620	8.706	
Omit	1	0.004	-0.039	-0.202	22.000	0.000	

### **Item 55 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
55	55	D	Yes	4	Grammar & text relations	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.533	0.541	0.678	0.454	0.569	

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	48	0.200	-0.099	-0.141	30.667	8.816	
B	35	0.146	-0.264	-0.407	26.143	8.565	
C	26	0.108	-0.270	-0.452	24.731	7.887	
D	128	0.533	0.454	0.569	38.063	9.671	**KEY**
Omit	3	0.013	-0.054	-0.187	24.667	3.055	

### **Item 56 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
56	56	C	Yes	4	Grammar & text relations	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.533	0.505	0.634	0.411	0.515	0.894

### **Option statistics**

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	31	0.129	-0.225	-0.359	26.677	8.987	
B	52	0.217	-0.142	-0.199	29.923	9.098	
C	128	0.533	0.411	0.515	37.656	10.019	**KEY**
D	26	0.108	-0.198	-0.331	26.846	7.993	
Omit	3	0.013	-0.053	-0.181	25.000	7.000	

### **Item 57 information and statistics**

Seq.	ID	Key	Scored	Num Options	Domain	Flags
57	57	A	Yes	4	Grammar & text relations	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.675	0.408	0.531	0.324	0.421	0.895

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	162	0.675	0.324	0.421	35.883	10.129	**KEY**
B	40	0.167	-0.081	-0.121	30.675	9.141	
C	16	0.067	-0.319	-0.616	20.188	4.199	
D	21	0.087	-0.132	-0.235	28.143	9.901	
Omit	1	0.004	-0.038	-0.198	22.000	0.000	

### Item 58 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
58	58	B	Yes	4	Grammar & text relations	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.662	0.424	0.549	0.240	0.311	0.896

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	24	0.100	-0.139	-0.237	28.250	8.789	
B	159	0.662	0.240	0.311	35.352	10.762	**KEY**
C	38	0.158	-0.084	-0.127	30.553	9.328	
D	17	0.071	-0.125	-0.236	27.882	8.373	
Omit	2	0.008	-0.057	-0.227	21.000	1.414	

### Item 59 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
59	59	D	Yes	4	Grammar & text relations	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.550	0.518	0.651	0.481	0.605	0.893

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	32	0.133	-0.203	-0.320	27.375	8.805	
B	23	0.096	-0.133	-0.231	28.478	9.244	
C	52	0.217	-0.308	-0.432	26.673	8.269	
D	132	0.550	0.481	0.605	38.152	9.498	**KEY**
Omit	1	0.004	-0.038	-0.201	22.000	0.000	

### Item 60 information and statistics

Seq.	ID	Key	Scored	Num Options	Domain	Flags
60	60	A	Yes	4	Grammar & text relations	

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
240	0.592	0.542	0.685	0.541	0.685	0.893

### Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	
A	142	0.592	0.541	0.685	38.246	9.281	**KEY**
B	18	0.075	-0.217	-0.404	24.833	8.590	
C	48	0.200	-0.266	-0.381	27.188	7.737	
D	30	0.125	-0.281	-0.451	25.033	7.536	
Omit	2	0.008	-0.057	-0.230	21.000	1.414	

## **Annexure F**

### **TiaPlus analysis of TALA**



TiaPlus® Test and Item Analysis Build 303  
 Cito, Measurement and Research Department. Arnhem, the Netherlands. © 2007.  
 Population : Eunice and Heidedal Grade 10s  
 Test : **TALA test for academic literacy at secondary school**  
 Date : 28 August 2014  
 Time : 16:33  
 Data file : C:\AlbertDocuments\ICELDA\ice01176.txt  
 Missing handling : Missing as Zero  
 Persons: All persons Items: All items

Test and Item Analysis																	%		#	
Item	Item	----- P- and A- values -----								Mis-	----- Weighted -----									
Label	nr.	Weight	Key	A	B	C	D	E	F	O/D	sing	Max	Mean	P	Sd	RSK	Rit	Rir	AR	
	1	1	D	1	13	4	81*	1		1	2	1	0.81	81	0.40	0.40	21	16	82	
	2	1	C	7	11	64*	7	9		2	5	1	0.64	64	0.48	0.48	2	-4	82	
	3	1	B	23	34*	17	4	19		4	9	1	0.34	34	0.47	0.47	23	18	82	
	4	1	A	41*	11	12	5	29		2	6	1	0.41	41	0.49	0.49	22	16	82	
	5	1	E	25	29	3	4	35*		4	10	1	0.35	35	0.48	0.48	25	19	82	
	6	1	B	3	64*	20	12			1	3	1	0.64	64	0.48	0.48	50	45	81	
	7	1	D	21	11	10	57*			2	4	1	0.57	57	0.50	0.50	38	32	81	
	8	1	C	4	3	91*	2			0	1	1	0.91	91	0.29	0.29	23	19	82	
	9	1	C	21	5	55*	19			0	1	1	0.55	55	0.50	0.50	41	36	81	
	10	1	C	16	30	40*	13			1	3	1	0.40	40	0.49	0.49	45	40	81	
	11	1	A	71*	20	3	5			2	4	1	0.71	71	0.46	0.46	29	24	82	
	12	1	C	31	6	28*	33			1	3	1	0.28	28	0.45	0.45	37	32	81	
	13	1	B	11	29*	48	12			1	2	1	0.29	29	0.46	0.46	16	10	82	
	14	1	D	21	36	4	37*			1	3	1	0.37	37	0.48	0.48	15	9	82	
	15	1	A	60*	2	19	18			1	3	1	0.60	60	0.49	0.49	31	25	82	
	16	1	A	30*	36	8	24			2	4	1	0.30	30	0.46	0.46	5	-1	82	
	17	1	D	16	22	29	32*			0	1	1	0.32	32	0.47	0.47	33	28	81	
	18	1	D	25	33	14	27*			0	0	1	0.27	27	0.45	0.45	37	32	81	
	19	1	A	18*	47	27	7			0	1	1	0.18	18	0.38	0.38	17	12	82	
	20	1	C	20	24	49*	6			1	2	1	0.49	49	0.50	0.50	39	34	81	
	21	1	C	36	15	33*	14			2	5	1	0.33	33	0.47	0.47	17	11	82	
	22	1	B	14	57*	18	12			0	0	1	0.57	57	0.50	0.50	29	24	82	
	23	1	D	15	17	36	31*			2	4	1	0.31	31	0.46	0.46	19	13	82	
	24	1	B	26	40*	10	23			0	1	1	0.40	40	0.49	0.49	35	30	81	
	25	1	B	7	32*	47	14			1	2	1	0.32	32	0.47	0.47	8	2	82	
	26	1	A	48*	18	22	12			0	1	1	0.48	48	0.50	0.50	32	26	81	
	27	1	B	19	61*	12	7			1	3	1	0.61	61	0.49	0.49	33	28	81	
	28	1	D	18	35	8	38*			1	2	1	0.38	38	0.49	0.49	32	27	81	
	29	1	A	31*	22	31	16			0	1	1	0.31	31	0.46	0.46	9	3	82	

Test and Item Analysis										% #									
Item Label	Item nr.	Weight	Key	P- and A- values						Mis-	Weighted								
				A	B	C	D	E	F	O/D	sing	Max	Mean	P	Sd	RSK	Rit	Rir	AR
	30	1	C	35	29	24*	11			1	3	1	0.24	24	0.43	0.43	29	24	82
	31	1	D	37	14	19	29*			1	3	1	0.29	29	0.46	0.46	45	41	81

Test and Item Analysis										% #									
Item Label	Item nr.	Weight	Key	P- and A- values						Mis-	Weighted								
				A	B	C	D	E	F	O/D	sing	Max	Mean	P	Sd	RSK	Rit	Rir	AR
	32	1	B	3	76*	7	12			1	3	1	0.76	76	0.43	0.43	39	35	81
	33	1	A	19*	23	32	23			3	7	1	0.19	19	0.40	0.40	45	41	81
	34	1	D	11	8	7	73*			1	3	1	0.73	73	0.45	0.45	40	35	81
	35	1	D	30	15	17	36*			2	4	1	0.36	36	0.48	0.48	22	17	82
	36	1	B	19	29*	16	33			2	5	1	0.29	29	0.46	0.46	23	17	82
	37	1	D	46	20	16	16*			2	6	1	0.16	16	0.36	0.36	20	16	82
	38	1	A	33*	26	15	24			2	4	1	0.33	33	0.47	0.47	12	6	82
	39	1	D	32	29	14	24*			1	3	1	0.24	24	0.42	0.42	32	27	81
	40	1	D	45	17	10	28*			1	3	1	0.28	28	0.45	0.45	48	43	81
	41	1	D	20	12	12	54*			2	5	1	0.54	54	0.50	0.50	56	51	81
	42	1	C	18	45	28*	8			1	2	1	0.28	28	0.45	0.45	21	16	82
	43	1	B	29	46*	5	19			1	3	1	0.46	46	0.50	0.50	28	22	82
	44	1	B	11	55*	7	25			1	3	1	0.55	55	0.50	0.50	36	30	81
	45	1	B	29	39*	17	13			2	5	1	0.39	39	0.49	0.49	26	20	82
	46	1	A	54*	39	5	1			2	4	1	0.54	54	0.50	0.50	50	45	81
	47	1	A	25*	52	9	12			1	3	1	0.25	25	0.43	0.43	31	27	81
	48	1	D	14	11	18	56*			1	3	1	0.56	56	0.50	0.50	39	33	81
	49	1	C	44	21	19*	12			3	7	1	0.19	19	0.39	0.39	11	6	82
	50	1	D	19	37	30	12*			2	5	1	0.12	12	0.33	0.33	22	18	82
	51	1	B	22	49*	11	17			2	4	1	0.49	49	0.50	0.50	29	23	82
	52	1	B	15	62*	15	5			2	5	1	0.62	62	0.49	0.49	38	33	81
	53	1	C	13	14	61*	8			4	9	1	0.61	61	0.49	0.49	43	38	81
	54	1	D	18	32	17	29*			4	9	1	0.29	29	0.46	0.46	36	31	81
	55	1	B	17	32*	35	10			7	16	1	0.32	32	0.47	0.47	29	24	82
	56	1	A	41*	15	21	17			6	14	1	0.41	41	0.49	0.49	18	12	82
	57	1	B	18	47*	18	9			8	20	1	0.47	47	0.50	0.50	31	25	82
	58	1	A	31*	20	31	10			8	20	1	0.31	31	0.46	0.46	23	18	82
	59	1	A	31*	19	18	24			8	20	1	0.31	31	0.46	0.46	33	28	81
	60	1	D	16	18	16	40*			9	22	1	0.40	40	0.49	0.49	27	21	82

```

SubGroup number      : 0
Number of persons in test : 242
Minimum test score   : 0
Average test score   : 25.11
Average P-value      : 41.85
Average Rit          : 0.30
Coefficient Alpha     : 0.82
GLB                  : 0.94
Items used in GLB proc : 60
Cut-off score        : 19.5

SubTest number       : 0
Number of selected items : 60
Maximum test score   : 60
Standard deviation    : 8.10
Std. Error of Measurement : 3.46

SE Coeff. Alpha      : 0.02
Asymptotic GLB coef : 0.91

Percentage failing    : 26.03

```

Misclassifications:

Alpha based				GLB based			
-Rxx' case	Percentage	:	15.2	Percentage	:	10.6	
	Number	:	37	Number	:	26	
-Rxt case	Percentage	:	11	Percentage	:	7.6	
	Number	:	27	Number	:	18	

90% Confidence limits for Coefficient Alpha: (0.79 =< 0.82 =< 0.84)

Estimated Coefficient Alpha if this test had a standard norm length of 40 items: 0.75 (Spearman-Brown)

Persons: All persons

Items: All items

Test and Item Analysis										% #																	
Item	----- Rir and Rar values -----					Mis-	Rel. Score Frequencies (unweighted, %) -----																				
Label	nr.	Weight	A	B	C	D	E	O/D	sing	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Code	
1	1	1	-8	-15	16*	-4	1	2		19	81																
2	1	1	5	-4*	-10	16	2	5		36	64																
3	1	-10	18*	10	-13	-5	4	9		66	34																
4	1	16*	-6	3	-4	-14	2	6		59	41																
5	1	-7	-3	-6	-13	19*	4	10		65	35																
6	1	-11	45*	-33	-20		1	3		36	64																
7	1	-38	4	-6	32*		2	4		43	57																
8	1	-8	-11	19*	-11		0	1		9	91																
9	1	-28	-23	36*	-6		0	1		45	55																
10	1	-4	-22	40*	-21		1	3		60	40																
11	1	24*	-8	-21	-10		2	4		29	71																
12	1	14	-10	32*	-35		1	3		72	28																

ABC



Test and Item Analysis

Item Label	nr.	Weight	Rir and Rar values				Mis-	Rel. Score			Frequencies (unweighted, %)															
			A	B	C	D	E	O/D	sing	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Code
46	1		45*	-42	-2	-12		2	4		46	54														
47	1		27*	-15	-9	-3		1	3		75	25														
48	1		-9	-17	-20	33*		1	3		44	56														
49	1		19	-20	6*	-3		3	7		81	19														
50	1		-4	4	-10	18*		2	5		88	12														
51	1		-13	23*	-13	-2		2	4		51	49														
52	1		-12	33*	-23	-6		2	5		38	62														
53	1		-17	-19	38*	-15		4	9		39	61														
54	1		-25	5	-14	31*		4	9		71	29														
55	1		-15	24*	0	-7		7	16		68	32														
56	1		12*	-5	-1	-3		6	14		59	41														
57	1		-8	25*	-12	-7		8	20		53	47														
58	1		18*	10	-13	-10		8	20		69	31														
59	1		28*	-20	-9	-1		8	20		69	31														
60	1		2	-12	-11	21*		9	22		60	40														

AC

Code legend: A: Rar >= Rir B: Rir <= 0 C: Rar >= 10

SubGroup number	: 0	SubTest number	: 0
Number of persons in test	: 242	Number of selected items	: 60
Minimum test score	: 0	Maximum test score	: 60
Average test score	: 25.11	Standard deviation	: 8.10
Average P-value	: 41.85	Std. Error of Measurement	: 3.46
Average Rit	: 0.30		
Coefficient Alpha	: 0.82	SE Coeff. Alpha	: 0.02
GLB	: 0.94	Asymptotic GLB coef	: 0.91
Items used in GLB proc	: 60		
Cut-off score	: 19.5	Percentage failing	: 26.03

Misclassifications:

Alpha based

GLB based

-Rxx' case	Percentage	: 15.2	Percentage	: 10.6
	Number	: 37	Number	: 26
-Rxt case	Percentage	: 11	Percentage	: 7.6
	Number	: 27	Number	: 18

90% Confidence limits for Coefficient Alpha: (0.79 =< 0.82 =< 0.84)

Estimated Coefficient Alpha if this test had a standard norm length of 40 items: 0.75 (Spearman-Brown)

Persons: All persons

Subtest(1): Scrambled text  
1-5

Test and Item Analysis

Item Label	Item nr.	Weight	Key	P- and A- values						Mis-	% # Weighted									
				A	B	C	D	E	F	O/D	sing	Max	Mean	P	Sd	RSK	Rit	Rir	AR	
	1	1	D	1	13	4	81*	1		1	2	1	0.81	81	0.40	0.40	50	25	59	
	2	1	C	7	11	64*	7	9		2	5	1	0.64	64	0.48	0.48	61	32	56	
	3	1	B	23	34*	17	4	19		4	9	1	0.34	34	0.47	0.47	75	53	44	
	4	1	A	41*	11	12	5	29		2	6	1	0.41	41	0.49	0.49	65	37	53	
	5	1	E	25	29	3	4	35*		4	10	1	0.35	35	0.48	0.48	58	30	57	

```

SubGroup number      : 0          SubTest number      : 1
Number of persons in test : 242      Number of selected items : 5
Minimum test score    : 0          Maximum test score    : 5
Average test score    : 2.54      Standard deviation    : 1.44
Average P-value       : 50.83     Std. Error of Measurement : 0.91
Average Rit          : 0.62
Coefficient Alpha      : 0.60      SE Coeff. Alpha      : 0.04
GLB                   : 0.75     Asymptotic GLB coef  : 0.63
Items used in GLB proc : 5

```

90% Confidence limits for Coefficient Alpha: (0.52 =< 0.60 =< 0.66)

Estimated Coefficient Alpha if this test had a standard norm length of 40 items: 0.92 (Spearman-Brown)

Persons: All persons

Subtest(1): Scrambled text  
1-5

Test and Item Analysis

Item Label	Item nr.	Weight	Rir and Rar values					Mis-	Rel. Score Frequencies (unweighted, %)																	
			A	B	C	D	E	O/D	sing	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Code
	1	1	-9	-17	-16	25*	-2		1	2	19	81														
	2	1	-25	2	32*	-22	-15		2	5	36	64														
	3	1	-21	53*	-20	-13	-13		4	9	66	34														
	4	1	37*	-13	-30	-15	-3		2	6	59	41														
	5	1	-4	-14	-14	-13	30*		4	10	65	35														

Code legend: A: Rar >= Rir B: Rir <= 0 C: Rar >= 10

```

SubGroup number      : 0          SubTest number      : 1
Number of persons in test : 242      Number of selected items : 5

```

```

Minimum test score      : 0
Average test score     : 2.54
Average P-value        : 50.83
Average Rit            : 0.62
Coefficient Alpha       : 0.60
GLB                    : 0.75
Items used in GLB proc : 5

Maximum test score     : 5
Standard deviation     : 1.44
Std. Error of Measurement : 0.91
SE Coeff. Alpha       : 0.04
Asymptotic GLB coef   : 0.63

```

-----  
90% Confidence limits for Coefficient Alpha: (0.52 =< 0.60 =< 0.66)  
-----

Estimated Coefficient Alpha if this test had a standard  
norm length of 40 items: 0.92 (Spearman-Brown)  
-----

Persons: All persons

Subtest(2): Vocabulary knowledge  
6-15

Test and Item Analysis			Subtest(2): Vocabulary knowledge																	
Item	Item		P- and A- values						Mis-	% #										
Label	nr.	Weight	Key	A	B	C	D	E	F	O/D	sing	Max	Mean	P	Sd	RSK	Rit	Rir	AR	
	6	1	B	3	64*	20	12			1	3		1	0.64	64	0.48	0.48	52	32	49
	7	1	D	21	11	10	57*			2	4		1	0.57	57	0.50	0.50	48	26	51
	8	1	C	4	3	91*	2			0	1		1	0.91	91	0.29	0.29	29	16	54
	9	1	C	21	5	55*	19			0	1		1	0.55	55	0.50	0.50	53	32	49
	10	1	C	16	30	40*	13			1	3		1	0.40	40	0.49	0.49	55	35	48
	11	1	A	71*	20	3	5			2	4		1	0.71	71	0.46	0.46	50	30	50
	12	1	C	31	6	28*	33			1	3		1	0.28	28	0.45	0.45	45	25	51
	13	1	B	11	29*	48	12			1	2		1	0.29	29	0.46	0.46	34	13	55
	14	1	D	21	36	4	37*			1	3		1	0.37	37	0.48	0.48	29	6	57
	15	1	A	60*	2	19	18			1	3		1	0.60	60	0.49	0.49	43	21	53

```

SubGroup number      : 0
Number of persons in test : 242
Minimum test score   : 0
Average test score   : 5.31
Average P-value      : 53.06
Average Rit          : 0.44
Coefficient Alpha     : 0.54
GLB                  : 0.64
Items used in GLB proc : 10

SubTest number       : 2
Number of selected items : 10
Maximum test score   : 10
Standard deviation   : 2.05
Std. Error of Measurement : 1.38
SE Coeff. Alpha     : 0.04
Asymptotic GLB coef : 0.62

```

-----  
90% Confidence limits for Coefficient Alpha: (0.47 =< 0.54 =< 0.61)  
-----

Estimated Coefficient Alpha if this test had a standard norm length of 40 items: 0.83 (Spearman-Brown)

Persons: All persons

Subtest(2): Vocabulary knowledge  
6-15

Test and Item Analysis

Item Label	nr.	Weight	Rir and Rar values				Mis-	Rel. Score Frequencies (unweighted, %)															Code		
			A	B	C	D	O/D sing	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
6	1		-10	32*	-28	-9	1	3		36	64														
7	1		-20	-12	-7	26*	2	4		43	57														
8	1		-7	-4	16*	-14	0	1		9	91														
9	1		-20	-15	32*	-14	0	1		45	55														
10	1		-9	-18	35*	-13	1	3		60	40														
11	1		30*	-15	-8	-22	2	4		29	71														
12	1		10	-11	25*	-24	1	3		72	28														
13	1		-35	13*	19	-12	1	2		71	29														
14	1		-3	1	-5	6*	1	3		63	37														
15	1		21*	-7	-20	-1	1	3		40	60														

AC

Code legend: A: Rar >= Rir B: Rir <= 0 C: Rar >= 10

SubGroup number	: 0	SubTest number	: 2
Number of persons in test	: 242	Number of selected items	: 10
Minimum test score	: 0	Maximum test score	: 10
Average test score	: 5.31	Standard deviation	: 2.05
Average P-value	: 53.06	Std. Error of Measurement	: 1.38
Average Rit	: 0.44		
Coefficient Alpha	: 0.54	SE Coeff. Alpha	: 0.04
GLB	: 0.64	Asymptotic GLB coef	: 0.62
Items used in GLB proc	: 10		

90% Confidence limits for Coefficient Alpha: (0.47 =< 0.54 =< 0.61)

Estimated Coefficient Alpha if this test had a standard norm length of 40 items: 0.83 (Spearman-Brown)



Persons: All persons

Subtest(3): Interpreting graphs & visual information  
16-23

Test and Item Analysis

Item Label	Item nr.	Weight	Key	P- and A- values					Mis-	% # Weighted									
				A	B	C	D	E	F	O/D	sing	Max	Mean	P	Sd	RSK	Rit	Rir	AR
	16	1	A	30*	36	8	24			2	4	1	0.30	30	0.46	0.46	28	1	50
	17	1	D	16	22	29	32*			0	1	1	0.32	32	0.47	0.47	46	20	42
	18	1	D	25	33	14	27*			0	0	1	0.27	27	0.45	0.45	53	29	38
	19	1	A	18*	47	27	7			0	1	1	0.18	18	0.38	0.38	41	19	43
	20	1	C	20	24	49*	6			1	2	1	0.49	49	0.50	0.50	57	31	37
	21	1	C	36	15	33*	14			2	5	1	0.33	33	0.47	0.47	45	18	43
	22	1	B	14	57*	18	12			0	0	1	0.57	57	0.50	0.50	53	26	39
	23	1	D	15	17	36	31*			2	4	1	0.31	31	0.46	0.46	42	16	44

```

SubGroup number      : 0
Number of persons in test : 242
Minimum test score   : 0
Average test score   : 2.76
Average P-value      : 34.56
Average Rit          : 0.46
Coefficient Alpha     : 0.46
GLB                  : 0.57
Items used in GLB proc : 8

SubTest number       : 3
Number of selected items : 8
Maximum test score   : 8
Standard deviation    : 1.68
Std. Error of Measurement : 1.24
SE Coeff. Alpha      : 0.05
Asymptotic GLB coef : 0.55

```

90% Confidence limits for Coefficient Alpha: (0.36 =< 0.46 =< 0.54)

Estimated Coefficient Alpha if this test had a standard norm length of 40 items: 0.81 (Spearman-Brown)

Persons: All persons

Subtest(4): Text comprehension  
24-48

Test and Item Analysis

Item Label	Item nr.	Weight	Key	----- P- and A- values -----						Mis-		% # Weighted -----							
				A	B	C	D	E	F	O/D	sing	Max	Mean	P	Sd	RSK	Rit	Rir	AR
	24	1	B	26	40*	10	23			0	1   1	0.40	40	0.49	0.49	38	27	71	
	25	1	B	7	32*	47	14			1	2   1	0.32	32	0.47	0.47	14	3	73	
	26	1	A	48*	18	22	12			0	1   1	0.48	48	0.50	0.50	38	27	71	
	27	1	B	19	61*	12	7			1	3   1	0.61	61	0.49	0.49	36	25	71	
	28	1	D	18	35	8	38*			1	2   1	0.38	38	0.49	0.49	39	28	71	
	29	1	A	31*	22	31	16			0	1   1	0.31	31	0.46	0.46	20	9	72	
	30	1	C	35	29	24*	11			1	3   1	0.24	24	0.43	0.43	32	22	71	
	31	1	D	37	14	19	29*			1	3   1	0.29	29	0.46	0.46	47	37	70	
	32	1	B	3	76*	7	12			1	3   1	0.76	76	0.43	0.43	44	36	71	
	33	1	A	19*	23	32	23			3	7   1	0.19	19	0.40	0.40	47	39	70	
	34	1	D	11	8	7	73*			1	3   1	0.73	73	0.45	0.45	46	37	70	
	35	1	D	30	15	17	36*			2	4   1	0.36	36	0.48	0.48	29	18	72	
	36	1	B	19	29*	16	33			2	5   1	0.29	29	0.46	0.46	29	19	72	
	37	1	D	46	20	16	16*			2	6   1	0.16	16	0.36	0.36	18	10	72	
	38	1	A	33*	26	15	24			2	4   1	0.33	33	0.47	0.47	15	4	73	
	39	1	D	32	29	14	24*			1	3   1	0.24	24	0.42	0.42	35	26	71	
	40	1	D	45	17	10	28*			1	3   1	0.28	28	0.45	0.45	49	40	70	
	41	1	D	20	12	12	54*			2	5   1	0.54	54	0.50	0.50	53	43	70	
	42	1	C	18	45	28*	8			1	2   1	0.28	28	0.45	0.45	25	15	72	
	43	1	B	29	46*	5	19			1	3   1	0.46	46	0.50	0.50	34	23	71	
	44	1	B	11	55*	7	25			1	3   1	0.55	55	0.50	0.50	41	30	71	
	45	1	B	29	39*	17	13			2	5   1	0.39	39	0.49	0.49	34	23	71	
	46	1	A	54*	39	5	1			2	4   1	0.54	54	0.50	0.50	56	47	69	
	47	1	A	25*	52	9	12			1	3   1	0.25	25	0.43	0.43	36	26	71	
	48	1	D	14	11	18	56*			1	3   1	0.56	56	0.50	0.50	43	33	71	

SubGroup number	: 0	SubTest number	: 4
Number of persons in test	: 242	Number of selected items	: 25
Minimum test score	: 0	Maximum test score	: 25
Average test score	: 9.95	Standard deviation	: 4.17
Average P-value	: 39.80	Std. Error of Measurement	: 2.20
Average Rit	: 0.36		
Coefficient Alpha	: 0.72	SE Coeff. Alpha	: 0.03
GLB	: 0.83	Asymptotic GLB coef	: 0.81
Items used in GLB proc	: 25		

90% Confidence limits for Coefficient Alpha: (0.68 =< 0.72 =< 0.76)

-----  
 Estimated Coefficient Alpha if this test had a standard  
 norm length of 40 items: 0.80 (Spearman-Brown)  
 -----

Persons: All persons

Subtest(4): Text comprehension  
 24-48

Test and Item Analysis

Item Label	nr.	Weight	Rir and Rar values				Mis- E	O/D	Rel. Score	% # Frequencies (unweighted, %)															Code	
			A	B	C	D			0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
24	1		-1	27*	-10	-23		0	1	60	40															
25	1		-19	3*	14	-9		1	2	68	32															AC
26	1		27*	-13	-10	-15		0	1	52	48															
27	1		-25	25*	-6	-3		1	3	39	61															
28	1		-27	-1	-9	28*		1	2	62	38															
29	1		9*	-26	7	6		0	1	69	31															
30	1		-14	-8	22*	3		1	3	76	24															
31	1		-27	-13	5	37*		1	3	71	29															
32	1		-19	36*	-10	-23		1	3	24	76															
33	1		39*	-7	-30	10		3	7	81	19															C
34	1		-17	-20	-16	37*		1	3	27	73															
35	1		2	0	-22	18*		2	4	64	36															
36	1		-19	19*	-4	4		2	5	71	29															
37	1		4	-5	-7	10*		2	6	84	16															
38	1		4*	-1	-13	10		2	4	67	33															AC
39	1		-25	10	-8	26*		1	3	76	24															C
40	1		-15	-21	-6	40*		1	3	72	28															
41	1		-17	-25	-13	43*		2	5	46	54															
42	1		-12	4	15*	-9		1	2	72	28															
43	1		0	23*	-8	-22		1	3	54	46															
44	1		-2	30*	-24	-16		1	3	45	55															
45	1		-18	23*	-9	5		2	5	61	39															
46	1		47*	-44	-4	-15		2	4	46	54															
47	1		26*	-15	-10	-3		1	3	75	25															
48	1		-8	-15	-23	33*		1	3	44	56															

-----  
 Code legend: A: Rar >= Rir B: Rir <= 0 C: Rar >= 10  
 SubGroup number : 0 SubTest number : 4  
 Number of persons in test : 242 Number of selected items : 25  
 Minimum test score : 0 Maximum test score : 25  
 Average test score : 9.95 Standard deviation : 4.17  
 -----

```

Average P-value          : 39.80          Std. Error of Measurement : 2.20
Average Rit              : 0.36
Coefficient Alpha        : 0.72          SE Coeff. Alpha          : 0.03
GLB                      : 0.83          Asymptotic GLB coef     : 0.81
Items used in GLB proc   : 25

```

-----  
90% Confidence limits for Coefficient Alpha: (0.68 =< 0.72 =< 0.76)  
-----

Estimated Coefficient Alpha if this test had a standard  
norm length of 40 items: 0.80 (Spearman-Brown)  
-----

```

Persons: All persons                Subtest(5): Grammar & text relations
                                      49-60

```

Test and Item Analysis			Subtest(5): Grammar & text relations 49-60																	
Item Label	Item nr.	Weight	Key	P- and A- values						Mis-	% # Weighted									
				A	B	C	D	E	F	O/D	sing	Max	Mean	P	Sd	RSK	Rit	Rir	AR	
	49	1	C	44	21	19*	12			3	7	1	0.19	19	0.39	0.39	18	2	64	
	50	1	D	19	37	30	12*			2	5	1	0.12	12	0.33	0.33	30	18	62	
	51	1	B	22	49*	11	17			2	4	1	0.49	49	0.50	0.50	48	30	60	
	52	1	B	15	62*	15	5			2	5	1	0.62	62	0.49	0.49	49	32	60	
	53	1	C	13	14	61*	8			4	9	1	0.61	61	0.49	0.49	44	26	61	
	54	1	D	18	32	17	29*			4	9	1	0.29	29	0.46	0.46	50	34	59	
	55	1	B	17	32*	35	10			7	16	1	0.32	32	0.47	0.47	50	34	59	
	56	1	A	41*	15	21	17			6	14	1	0.41	41	0.49	0.49	48	31	60	
	57	1	B	18	47*	18	9			8	20	1	0.47	47	0.50	0.50	49	31	60	
	58	1	A	31*	20	31	10			8	20	1	0.31	31	0.46	0.46	38	20	62	
	59	1	A	31*	19	18	24			8	20	1	0.31	31	0.46	0.46	49	33	60	
	60	1	D	16	18	16	40*			9	22	1	0.40	40	0.49	0.49	49	32	60	

```

SubGroup number          : 0          SubTest number          : 5
Number of persons in test : 242       Number of selected items : 12
Minimum test score       : 0          Maximum test score      : 12
Average test score       : 4.55       Standard deviation      : 2.46
Average P-value          : 37.91      Std. Error of Measurement : 1.50
Average Rit              : 0.44
Coefficient Alpha         : 0.63          SE Coeff. Alpha        : 0.04
GLB                      : 0.79          Asymptotic GLB coef    : 0.80
Items used in GLB proc   : 12

```

-----  
90% Confidence limits for Coefficient Alpha: (0.57 =< 0.63 =< 0.68)  
-----



Persons: All persons

Items: All items

Table of Subtest Intercorrelations

	Subtest	Total test	Subtest(s)					
			1	2	3	4	5	
Scrambled text	1	0.30						
Vocabulary know	2	0.74	0.10					
Interpreting gr	3	0.54	0.05	0.30				
Text comprehens	4	0.88	0.15	0.56	0.38			
Grammar & text	5	0.65	0.04	0.39	0.17	0.40		
-----								
Number of testees :		242	242	242	242	242	242	
Number of items :		60	5	10	8	25	12	
Average test score:		25.11	2.54	5.31	2.76	9.95	4.55	
Standard deviation:		8.10	1.44	2.05	1.68	4.17	2.46	
SEM :		3.46	0.91	1.38	1.24	2.20	1.50	
Average P-value :		41.85	50.83	53.06	34.56	39.80	37.91	
Coefficient Alpha :		0.82	0.60	0.54	0.46	0.72	0.63	
GLB :		0.94	0.75	0.64	0.57	0.83	0.79	
Asymptotic GLB :		0.91	0.63	0.62	0.55	0.81	0.80	
-----								

TiaPlus® Test and Item Analysis Build 303

Cito, Measurement and Research Department. Arnhem, the Netherlands. © 2007.

Population : Eunice and Heidedal Grade 10s  
Test : TALA re-pilot  
Date : 02 September 2014  
Time : 11:50  
Data file : C:\AlbertDocuments\ICELDA\ice01196.txt  
Missing handling : Missing as Zero  
Mantel-Haenszel DIF statistics  
Information on Subtest: All items  
Comparing subgroups 1 vs 2 (Eunice - Heidedal)

Label	Item	DIF stat	z (stand)
	1	0.0584	-2.0649
	2	0.0470	-2.5978 *
	3	0.2030	-1.7471
	4	0.7538	-0.3675
	5	1.5407	0.5352
	6	1.7119	0.6859
	7	1.8767	0.8491
	8	2.1985	0.6377
	9	1.6676	0.6815
	10	1.5304	0.5057
	11	2.9260	1.3794
	12	0.7285	-0.3497
	13	0.6857	-0.4707
	14	1.5826	0.5819
	15	1.1128	0.1467
	16	1.2567	0.2899
	17	2.1031	0.8100
	18	0.8390	-0.1829
	19	0.6271	-0.4326
	20	2.2986	1.0886
	21	1.6365	0.6064
	22	1.1640	0.2070
	23	0.7351	-0.3799
	24	1.7150	0.6730
	25	1.0060	0.0076
	26	0.8906	-0.1542
	27	1.1620	0.2011
	28	1.1362	0.1585
	29	1.7852	0.7014
	30	2.7467	0.9476
	31	1.3789	0.3211
	32	1.2299	0.2511
	33	3.0362	0.7429
	34	1.3067	0.3458
	35	1.4140	0.4522
	36	0.8405	-0.2058
	37	0.3765	-0.9110
	38	0.9388	-0.0807
	39	1.8323	0.5976
	40	1.6183	0.4321
	41	2.9485	1.3551

Label	Item	DIF stat	z (stand)
	42	1.0589	0.0679
	43	2.0366	0.9341
	44	1.0599	0.0784
	45	0.9774	-0.0297
	46	3.1431	1.4626
	47	0.5231	-0.6994
	48	1.7906	0.7851
	49	1.0768	0.0748
	50	0.8817	-0.0929
	51	0.8562	-0.2089
	52	1.3962	0.4495
	53	2.6610	1.2856
	54	1.0101	0.0109
	55	0.7068	-0.3995
	56	1.2322	0.2798
	57	1.8155	0.7874
	58	1.0724	0.0840
	59	1.0612	0.0677
	60	1.4295	0.46

Interpretation:

If the DIF statistic is  $< 1$  then the studied item is more difficult in the first subgroup.

If the DIF statistic is approx. 1 then the studied item has equal difficulty for both subgroups.

If the DIF statistic is  $> 1$  then the studied item is more difficult in the second subgroup.

Significance (at alpha level = 1%):

Differences between subgroups are significant when the absolute value of  $z(\text{stand}) \geq 2.58$

('--' is shown if TiaPlus can not calculate the statistic).

Note that in case of subtest processing the result for an item will change as a subtest has a total score that differs from the total test total score. The group of persons therefore will be partitioned differently.

Cito, Measurement and Research Department. Arnhem, the Netherlands. © 2007.



## **Annexure G**

### **TiaPlus analysis of the second test**



33	1	B	14	65*	17	2	3	6		1	0.65	65	0.48	0.48	53	50	89
34	1	A	50*	21	8	20	2	4		1	0.50	50	0.50	0.50	40	36	89
35	1	D	5	22	10	61*	3	7		1	0.61	61	0.49	0.49	32	28	90
36	1	B	6	67*	13	13	1	2		1	0.67	67	0.47	0.47	50	46	89
37	1	C	25	22	37*	15	1	2		1	0.37	37	0.48	0.48	29	24	90
38	1	D	9	13	18	58*	2	4		1	0.58	58	0.49	0.49	46	42	89
39	1	B	8	62*	18	13	0	1		1	0.62	62	0.49	0.49	38	34	89
40	1	D	7	6	23	62*	3	7		1	0.62	62	0.49	0.49	39	35	89
41	1	C	12	9	50*	29	1	2		1	0.50	50	0.50	0.50	38	34	89
42	1	D	12	30	20	37*	2	4		1	0.37	37	0.48	0.48	39	35	89
43	1	B	24	51*	12	13	0	1		1	0.51	51	0.50	0.50	21	17	90
44	1	C	13	18	35*	33	0	1		1	0.35	35	0.48	0.48	39	35	89
45	1	A	43*	32	14	10	0	1		1	0.43	43	0.49	0.49	16	11	90
46	1	C	5	9	72*	13	0	0		1	0.72	72	0.45	0.45	40	36	89
47	1	B	32	44*	8	16	0	0		1	0.44	44	0.50	0.50	35	31	89
48	1	D	10	4	7	77*	1	3		1	0.77	77	0.42	0.42	38	34	89
49	1	B	33	36*	21	8	2	4		1	0.36	36	0.48	0.48	31	27	90
50	1	D	14	19	20	44*	3	6		1	0.44	44	0.50	0.50	33	29	90
51	1	A	35*	19	36	9	0	1		1	0.35	35	0.48	0.48	22	17	90
52	1	C	33	10	38*	18	1	2		1	0.38	38	0.49	0.49	14	9	90
53	1	D	30	28	5	36*	1	3		1	0.36	36	0.48	0.48	24	20	90
54	1	C	31	10	38*	21	0	1		1	0.38	38	0.49	0.49	40	36	89
55	1	D	20	15	11	53*	1	3		1	0.53	53	0.50	0.50	49	45	89
56	1	C	13	22	53*	11	1	3		1	0.53	53	0.50	0.50	45	41	89
57	1	A	68*	17	7	9	0	1		1	0.68	68	0.47	0.47	36	32	89
58	1	B	10	66*	16	7	1	2		1	0.66	66	0.47	0.47	28	24	90
59	1	D	13	10	22	55*	0	1		1	0.55	55	0.50	0.50	52	48	89
60	1	A	59*	8	20	13	1	2		1	0.59	59	0.49	0.49	57	54	89

```

-----
SubGroup number          : 0
Number of persons in test : 240
Minimum test score      : 0
Average test score      : 33.23
Average P-value         : 55.39
Average Rit             : 0.38
Coefficient Alpha        : 0.90
GLB                     : 0.97
Items used in GLB proc  : 60
Cut-off score           : 0.5
Misclassifications:
    Alpha based
SubTest number          : 0
Number of selected items : 60
Maximum test score      : 60
Standard deviation       : 10.53
Std. Error of Measurement : 3.40
SE Coeff. Alpha         : 0.01
Asymptotic GLB coef     : 0.96
Percentage failing       : 0
    GLB based
-----

```



Test and Item Analysis

Item	Rir and Rar values						Mis-	Rel. Score Frequencies (unweighted, %)																	Code			
Label	nr.	Weight	A	B	C	D	E	O/D	sing	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Code
29	1		-18	-30	32*	-5		0	0		35	65																
30	1		-30	-20	-13	40*		0	0		33	68																
31	1		-22	-36	54*	-13		3	6		40	60																
32	1		25	-16	-5	-13*		2	4		85	15																ABC
33	1		-38	50*	-16	-19		3	6		35	65																
34	1		36*	-19	-23	-6		2	4		50	50																
35	1		-21	-7	-17	28*		3	7		39	61																
36	1		-27	46*	-25	-18		1	2		33	67																
37	1		-15	3	24*	-17		1	2		63	37																
38	1		-20	-7	-28	42*		2	4		42	58																
39	1		-29	34*	-17	-5		0	1		38	62																
40	1		-31	-22	-5	35*		3	7		38	62																
41	1		-29	-28	34*	3		1	2		50	50																
42	1		-29	-20	3	35*		2	4		63	37																
43	1		5	17*	-11	-20		0	1		49	51																
44	1		-26	-27	35*	5		0	1		65	35																
45	1		11*	14	-21	-15		0	1		57	43																AC
46	1		-23	-20	36*	-15		0	0		28	72																
47	1		3	31*	-27	-25		0	0		56	44																
48	1		-13	-15	-26	34*		1	3		23	77																
49	1		-4	27*	-20	-8		2	4		64	36																
50	1		-16	-24	5	29*		3	6		56	44																
51	1		17*	-26	10	-9		0	1		65	35																C
52	1		17	-10	9*	-23		1	2		62	38																AC
53	1		-7	-6	-14	20*		1	3		64	36																
54	1		-12	-16	36*	-16		0	1		62	38																
55	1		-10	-26	-27	45*		1	3		47	53																
56	1		-22	-14	41*	-20		1	3		47	53																
57	1		32*	-8	-32	-13		0	1		33	68																
58	1		-14	24*	-8	-12		1	2		34	66																
59	1		-20	-13	-31	48*		0	1		45	55																
60	1		54*	-22	-27	-28		1	2		41	59																

Code legend: A: Rar >= Rir B: Rir <= 0 C: Rar >= 10

SubGroup number	: 0	SubTest number	: 0
Number of persons in test	: 240	Number of selected items	: 60
Minimum test score	: 0	Maximum test score	: 60
Average test score	: 33.23	Standard deviation	: 10.53
Average P-value	: 55.39	Std. Error of Measurement	: 3.40
Average Rit	: 0.38		

Coefficient Alpha : 0.90 SE Coeff. Alpha : 0.01  
 GLB : 0.97 Asymptotic GLB coef : 0.96  
 Items used in GLB proc : 60  
 Cut-off score : 0.5 Percentage failing : 0

Misclassifications:

Alpha based			GLB based		
-Rxx' case	Percentage	: 0.1	Percentage	:	0.1
	Number	: 0	Number	:	0
-Rxt case	Percentage	: 0.1	Percentage	:	0
	Number	: 0	Number	:	0

90% Confidence limits for Coefficient Alpha: (0.88 =< 0.90 =< 0.91)

Estimated Coefficient Alpha if this test had a standard norm length of 40 items: 0.85 (Spearman-Brown)

Persons: All persons

Subtest(1): Scrambled text  
1-5

Test and Item Analysis		P- and A- values										Mis-		% #					
Item Label	Item nr.	Weight	Key	A	B	C	D	E	F	O/D	sing	Max	Mean	P	Sd	RSK	Rit	Rir	AR
	1	1	B	7	55*	3	1	33		1	2	1	0.55	55	0.50	0.50	79	66	84
	2	1	D	22	7	7	51*	12		2	5	1	0.51	51	0.50	0.50	81	69	83
	3	1	E	12	11	14	17	43*		4	10	1	0.42	43	0.49	0.49	88	80	80
	4	1	A	44*	17	16	17	5		1	3	1	0.44	44	0.50	0.50	86	77	81
	5	1	C	13	10	59*	13	4		2	4	1	0.59	59	0.49	0.49	68	50	88

SubGroup number	: 0	SubTest number	: 1
Number of persons in test	: 240	Number of selected items	: 5
Minimum test score	: 0	Maximum test score	: 5
Average test score	: 2.51	Standard deviation	: 1.99
Average P-value	: 50.25	Std. Error of Measurement	: 0.74
Average Rit	: 0.81		
Coefficient Alpha	: 0.86	SE Coeff. Alpha	: 0.01
GLB	: 0.90	Asymptotic GLB coef	: 0.90
Items used in GLB proc	: 5		

90% Confidence limits for Coefficient Alpha: (0.84 =< 0.86 =< 0.89)

Estimated Coefficient Alpha if this test had a standard norm length of 40 items: 0.98 (Spearman-Brown)

Persons: All persons

Subtest(1): Scrambled text  
1-5

Test and Item Analysis

Item Label	nr.	Weight	Rir and Rar values					Mis-	Rel. Score Frequencies (unweighted, %)																	Code	
			A	B	C	D	E	O/D sing	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	
1	1		-25	66*	-12	-5	-49	1	2		45	55															
2	1		-44	-20	-23	69*	-15	2	5		49	51															
3	1		-20	-23	-39	-28	80*	4	10		58	43															
4	1		77*	-40	-20	-33	-13	1	3		56	44															
5	1		-6	-27	50*	-37	-9	2	4		41	59															

Code legend: A: Rar >= Rir B: Rir <= 0 C: Rar >= 10

SubGroup number	: 0	SubTest number	: 1
Number of persons in test	: 240	Number of selected items	: 5
Minimum test score	: 0	Maximum test score	: 5
Average test score	: 2.51	Standard deviation	: 1.99
Average P-value	: 50.25	Std. Error of Measurement	: 0.74
Average Rit	: 0.81		
Coefficient Alpha	: 0.86	SE Coeff. Alpha	: 0.01
GLB	: 0.90	Asymptotic GLB coef	: 0.90
Items used in GLB proc	: 5		

90% Confidence limits for Coefficient Alpha: (0.84 =< 0.86 =< 0.89)

Estimated Coefficient Alpha if this test had a standard norm length of 40 items: 0.98 (Spearman-Brown)

Persons: All persons

Subtest(2): Vocabulary knowledge  
6-15

Test and Item Analysis

Item Label	Item nr.	Weight	Key	P- and A- values					Mis-	Weighted										
				A	B	C	D	E	F	O/D sing	Max	Mean	P	Sd	RSK	Rit	Rir	AR		
6	1		B	5	60*	28	6			2	5		1	0.60	60	0.49	0.49	-5	-32	59
7	1		C	1	5	90*	2			2	4		1	0.90	90	0.29	0.29	46	31	39
8	1		C	12	4	77*	6			1	3		1	0.77	77	0.42	0.42	56	36	35
9	1		D	15	4	4	77*			0	1		1	0.77	77	0.42	0.42	53	32	36
10	1		A	63*	21	8	7			2	5		1	0.63	63	0.48	0.48	57	33	35
11	1		C	35	8	45*	12			1	2		1	0.45	45	0.50	0.50	48	22	40

Test and Item Analysis										% #										
Item Label	Item nr.	Weight	Key	P- and A- values						Mis-	Weighted									
				A	B	C	D	E	F	O/D	sing	Max	Mean	P	Sd	RSK	Rit	Rir	AR	
	12	1	B	3	93*	4	0			0	0   1	0.93	93	0.26	0.26	39	26	40		
	13	1	B	65	13*	10	12			0	1   1	0.13	13	0.34	0.34	6	-14	50		
	14	1	D	22	5	13	60*			1	2   1	0.60	60	0.49	0.49	49	24	39		
	15	1	A	48*	1	34	17			0	0   1	0.47	48	0.50	0.50	60	36	33		

```

SubGroup number      : 0
Number of persons in test : 240
Minimum test score   : 0
Average test score   : 6.23
Average P-value      : 62.33
Average Rit          : 0.42
Coefficient Alpha     : 0.44
GLB                  : 0.64
Items used in GLB proc : 10

SubTest number       : 2
Number of selected items : 10
Maximum test score   : 10
Standard deviation    : 1.75
Std. Error of Measurement : 1.30
SE Coeff. Alpha      : 0.05
Asymptotic GLB coef  : 0.53

```

90% Confidence limits for Coefficient Alpha: (0.35 =< 0.44 =< 0.53)

Estimated Coefficient Alpha if this test had a standard norm length of 40 items: 0.76 (Spearman-Brown)

Persons: All persons

Subtest(2): Vocabulary knowledge  
6-15

Test and Item Analysis										% #																		
Item Label	Item nr.	Weight	Rir and Rar values					Mis-	Rel. Score Frequencies (unweighted, %)																			
			A	B	C	D	E	O/D	sing	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Code
	6	1	17	-32*	30	2			2	5   40	60																	ABC
	7	1	-21	-20	31*	-1			2	4   10	90																	
	8	1	-10	-16	36*	-28			1	3   23	77																	
	9	1	-30	-19	7	32*			0	1   23	77																	
	10	1	33*	-23	-2	-12			2	5   38	63																	
	11	1	-15	-17	22*	2			1	2   55	45																	
	12	1	-21	26*	-15	0			0	0   7	93																	
	13	1	33	-14*	-11	-23			0	1   87	13																	ABC
	14	1	-19	-9	-5	24*			1	2   40	60																	
	15	1	36*	-16	-18	-21			0	0   53	48																	

Code legend: A: Rar >= Rir B: Rir <= 0 C: Rar >= 10



```

SubGroup number      : 0                   SubTest number       : 2
Number of persons in test : 240              Number of selected items : 10
Minimum test score   : 0                   Maximum test score     : 10
Average test score   : 6.23                Standard deviation      : 1.75
Average P-value      : 62.33                Std. Error of Measurement : 1.30
Average Rit          : 0.42
Coefficient Alpha     : 0.44                 SE Coeff. Alpha       : 0.05
GLB                   : 0.64                 Asymptotic GLB coef   : 0.53
Items used in GLB proc : 10

```

```

-----
90% Confidence limits for Coefficient Alpha: (0.35 =< 0.44 =< 0.53)
-----

```

```

Estimated Coefficient Alpha if this test had a standard
norm length of 40 items: 0.76 (Spearman-Brown)
-----

```

```

Persons: All persons                               Subtest(3): Interpreting graphs & visual information
                                                16-23

```

Test and Item Analysis																					
Item Label	Item nr.	Weight	Key	P- and A- values							Mis-				Weighted						
				A	B	C	D	E	F	O/D	sing	Max	Mean	P	Sd	RSK	Rit	Rir	AR		
	16	1	C	8	18	51*	24			0	0	1	0.51	51	0.50	0.50	63	48	78		
	17	1	C	13	19	63*	5			1	2	1	0.63	63	0.48	0.48	66	52	78		
	18	1	C	5	5	65*	24			1	2	1	0.65	65	0.48	0.48	72	60	76		
	19	1	B	20	61*	8	10			1	2	1	0.61	61	0.49	0.49	69	56	77		
	20	1	B	4	77*	13	6			0	1	1	0.77	77	0.42	0.42	60	48	78		
	21	1	B	5	79*	12	4			0	0	1	0.79	79	0.41	0.41	63	51	78		
	22	1	A	44*	27	21	8			0	0	1	0.44	44	0.50	0.50	57	41	79		
	23	1	C	12	14	60*	13			1	3	1	0.60	60	0.49	0.49	67	54	77		

```

-----
SubGroup number      : 0                   SubTest number       : 3
Number of persons in test : 240              Number of selected items : 8
Minimum test score   : 0                   Maximum test score     : 8
Average test score   : 5.00                Standard deviation      : 2.43
Average P-value      : 62.55                Std. Error of Measurement : 1.09
Average Rit          : 0.65
Coefficient Alpha     : 0.80                 SE Coeff. Alpha       : 0.02
GLB                   : 0.85                 Asymptotic GLB coef   : 0.84
Items used in GLB proc : 8

```

```

-----
90% Confidence limits for Coefficient Alpha: (0.77 =< 0.80 =< 0.83)
-----

```

```

Estimated Coefficient Alpha if this test had a standard
norm length of 40 items: 0.95 (Spearman-Brown)

```

Persons: All persons

Subtest(3): Interpreting graphs & visual information  
16-23

Test and Item Analysis

Item Label	nr.	Weight	Rir and Rar values					Mis-	Rel. Score Frequencies (unweighted, %)																	#	
			A	B	C	D	E	O/D sing	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Code
16	1		-17	-16	48*	-32		0	0		49	51															
17	1		-25	-38	52*	-7		1	2		38	63															
18	1		-10	-3	60*	-61		1	2		35	65															
19	1		-24	56*	-28	-29		1	2		39	61															
20	1		-13	48*	-37	-19		0	1		23	77															
21	1		-16	51*	-40	-22		0	0		21	79															
22	1		41*	-19	-16	-20		0	0		56	44															
23	1		-35	-19	54*	-25		1	3		40	60															

Code legend: A: Rar >= Rir B: Rir <= 0 C: Rar >= 10

SubGroup number	: 0	SubTest number	: 3
Number of persons in test	: 240	Number of selected items	: 8
Minimum test score	: 0	Maximum test score	: 8
Average test score	: 5.00	Standard deviation	: 2.43
Average P-value	: 62.55	Std. Error of Measurement	: 1.09
Average Rit	: 0.65		
Coefficient Alpha	: 0.80	SE Coeff. Alpha	: 0.02
GLB	: 0.85	Asymptotic GLB coef	: 0.84
Items used in GLB proc	: 8		

90% Confidence limits for Coefficient Alpha: (0.77 =< 0.80 =< 0.83)

Estimated Coefficient Alpha if this test had a standard norm length of 40 items: 0.95 (Spearman-Brown)

Persons: All persons

Subtest(4): Text comprehension  
24-48

Test and Item Analysis

Item Label	Item nr.	Weight	Key	P- and A- values				Mis-	Weighted										#	
				A	B	C	D	E	F	O/D sing	Max	Mean	P	Sd	RSK	Rit	Rir	AR		
24	1		C	5	10	77*	8			0	1		1	0.77	77	0.42	0.42	29	20	75
25	1		B	34	52*	10	3			0	1		1	0.52	52	0.50	0.50	27	16	75
26	1		A	77*	8	4	10			1	2		1	0.77	77	0.42	0.42	51	43	73
27	1		C	3	15	66*	15			1	2		1	0.66	66	0.47	0.47	20	10	75
28	1		D	38	28	16	18*			0	1		1	0.18	18	0.38	0.38	3	-6	76
29	1		C	3	14	65*	19			0	0		1	0.65	65	0.48	0.48	43	34	74

Test and Item Analysis

Item Label	Item nr.	Weight	Key	P- and A- values						Mis- O/D	sing	% # Weighted							
				A	B	C	D	E	F			Max	Mean	P	Sd	RSK	Rit	Rir	AR
	30	1	D	8	13	11	68*			0	0	1	0.68	68	0.47	0.47	47	38	74
	31	1	C	13	20	60*	5			3	6	1	0.60	60	0.49	0.49	61	53	73
	32	1	D	66	10	8	15*			2	4	1	0.15	15	0.35	0.35	-4	-12	76
	33	1	B	14	65*	17	2			3	6	1	0.65	65	0.48	0.48	53	45	73
	34	1	A	50*	21	8	20			2	4	1	0.50	50	0.50	0.50	40	30	74
	35	1	D	5	22	10	61*			3	7	1	0.61	61	0.49	0.49	42	32	74
	36	1	B	6	67*	13	13			1	2	1	0.67	67	0.47	0.47	51	42	73
	37	1	C	25	22	37*	15			1	2	1	0.37	37	0.48	0.48	27	16	75
	38	1	D	9	13	18	58*			2	4	1	0.58	58	0.49	0.49	50	41	73
	39	1	B	8	62*	18	13			0	1	1	0.62	62	0.49	0.49	43	33	74
	40	1	D	7	6	23	62*			3	7	1	0.62	62	0.49	0.49	46	37	74
	41	1	C	12	9	50*	29			1	2	1	0.50	50	0.50	0.50	37	27	74
	42	1	D	12	30	20	37*			2	4	1	0.37	37	0.48	0.48	42	32	74
	43	1	B	24	51*	12	13			0	1	1	0.51	51	0.50	0.50	32	21	75
	44	1	C	13	18	35*	33			0	1	1	0.35	35	0.48	0.48	38	28	74
	45	1	A	43*	32	14	10			0	1	1	0.43	43	0.49	0.49	25	14	75
	46	1	C	5	9	72*	13			0	0	1	0.72	72	0.45	0.45	49	40	73
	47	1	B	32	44*	8	16			0	0	1	0.44	44	0.50	0.50	41	31	74
	48	1	D	10	4	7	77*			1	3	1	0.77	77	0.42	0.42	42	34	74

SubGroup number	: 0	SubTest number	: 4
Number of persons in test	: 240	Number of selected items	: 25
Minimum test score	: 0	Maximum test score	: 25
Average test score	: 13.65	Standard deviation	: 4.44
Average P-value	: 54.62	Std. Error of Measurement	: 2.22
Average Rit	: 0.38		
Coefficient Alpha	: 0.75	SE Coeff. Alpha	: 0.02
GLB	: 0.86	Asymptotic GLB coef	: 0.84
Items used in GLB proc	: 25		

90% Confidence limits for Coefficient Alpha: (0.71 =< 0.75 =< 0.79)

Estimated Coefficient Alpha if this test had a standard norm length of 40 items: 0.83 (Spearman-Brown)

Persons: All persons  
Subtest(4): Text comprehension  
24-48

Test and Item Analysis

Item Label	nr.	Weight	A	B	C	D	E	Mis-	O/D	sing	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Code
24	1		-16	-4	20*	-11		0	1		23	77																	
25	1		-2	16*	-12	-16		0	1		48	52																	
26	1		43*	-28	-26	-18		1	2		23	77																	
27	1		-16	-13	10*	9		1	2		34	66																	
28	1		16	-1	-14	-6*		0	1		82	18																	ABC
29	1		-18	-28	34*	-10		0	0		35	65																	
30	1		-19	-21	-18	38*		0	0		33	68																	
31	1		-25	-32	53*	-12		3	6		40	60																	
32	1		25	-17	-5	-12*		2	4		85	15																	ABC
33	1		-31	45*	-15	-17		3	6		35	65																	
34	1		30*	-15	-24	0		2	4		50	50																	
35	1		-23	-10	-16	32*		3	7		39	61																	
36	1		-26	42*	-21	-17		1	2		33	67																	
37	1		-12	3	16*	-10		1	2		63	37																	
38	1		-19	-8	-27	41*		2	4		42	58																	
39	1		-29	33*	-23	3		0	1		38	62																	
40	1		-30	-29	-2	37*		3	7		38	62																	
41	1		-24	-28	27*	8		1	2		50	50																	
42	1		-24	-16	-1	32*		2	4		63	37																	
43	1		6	21*	-15	-24		0	1		49	51																	
44	1		-24	-23	28*	8		0	1		65	35																	
45	1		14*	14	-27	-12		0	1		57	43																	AC
46	1		-27	-21	40*	-17		0	0		28	72																	
47	1		1	31*	-29	-20		0	0		56	44																	
48	1		-14	-16	-26	34*		1	3		23	77																	

Code legend: A: Rar >= Rir B: Rir <= 0 C: Rar >= 10

SubGroup number	: 0	SubTest number	: 4
Number of persons in test	: 240	Number of selected items	: 25
Minimum test score	: 0	Maximum test score	: 25
Average test score	: 13.65	Standard deviation	: 4.44
Average P-value	: 54.62	Std. Error of Measurement	: 2.22
Average Rit	: 0.38		
Coefficient Alpha	: 0.75	SE Coeff. Alpha	: 0.02
GLB	: 0.86	Asymptotic GLB coef	: 0.84
Items used in GLB proc	: 25		

90% Confidence limits for Coefficient Alpha: (0.71 =< 0.75 =< 0.79)

Estimated Coefficient Alpha if this test had a standard norm length of 40 items: 0.83 (Spearman-Brown)

Persons: All persons

Subtest(5): Grammar & text relations  
49-60

Test and Item Analysis

Item Label	Item nr.	Weight	Key	P- and A- values						Mis-	% # Weighted									
				A	B	C	D	E	F	O/D	sing	Max	Mean	P	Sd	RSK	Rit	Rir	AR	
	49	1	B	33	36*	21	8			2	4   1	0.36	36	0.48	0.48	42	27	70		
	50	1	D	14	19	20	44*			3	6   1	0.44	44	0.50	0.50	45	29	69		
	51	1	A	35*	19	36	9			0	1   1	0.35	35	0.48	0.48	32	16	71		
	52	1	C	33	10	38*	18			1	2   1	0.38	38	0.49	0.49	27	10	72		
	53	1	D	30	28	5	36*			1	3   1	0.36	36	0.48	0.48	43	28	69		
	54	1	C	31	10	38*	21			0	1   1	0.38	38	0.49	0.49	56	42	67		
	55	1	D	20	15	11	53*			1	3   1	0.53	53	0.50	0.50	62	49	66		
	56	1	C	13	22	53*	11			1	3   1	0.53	53	0.50	0.50	59	45	67		
	57	1	A	68*	17	7	9			0	1   1	0.68	68	0.47	0.47	46	32	69		
	58	1	B	10	66*	16	7			1	2   1	0.66	66	0.47	0.47	48	34	69		
	59	1	D	13	10	22	55*			0	1   1	0.55	55	0.50	0.50	60	46	67		
	60	1	A	59*	8	20	13			1	2   1	0.59	59	0.49	0.49	61	48	66		

```

SubGroup number      : 0
Number of persons in test : 240
Minimum test score   : 0
Average test score   : 5.83
Average P-value      : 48.58
Average Rit          : 0.49
Coefficient Alpha     : 0.70
GLB                  : 0.89
Items used in GLB proc : 12

SubTest number       : 5
Number of selected items : 12
Maximum test score   : 12
Standard deviation    : 2.83
Std. Error of Measurement : 1.54
SE Coeff. Alpha      : 0.03
Asymptotic GLB coef : 0.84

```

90% Confidence limits for Coefficient Alpha: (0.66 =< 0.70 =< 0.75)

Estimated Coefficient Alpha if this test had a standard norm length of 40 items: 0.89 (Spearman-Brown)

Persons: All persons

Subtest(5): Grammar & text relations  
49-60

Test and Item Analysis

Item Label	Item nr.	Weight	Rir and Rar values					Mis-	Rel. Score Frequencies (unweighted, %)																			
			A	B	C	D	E	O/D	sing	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Code
	49	1	-5	27*	-19	-6			2	4   64	36																	
	50	1	-15	-17	-2	29*			3	6   56	44																	
	51	1	16*	-31	15	-6			0	1   65	35																	

Test and Item Analysis

Item Label	nr.	Weight	Rir and Rar values				Mis-	Rel. Score Frequencies (unweighted, %)																						
			A	B	C	D	E	O/D	sing	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Code		
52	1		18	-13	10*	-21		1	2		62	38																		
53	1		-10	-10	-12	28*		1	3		64	36																		
54	1		-14	-21	42*	-17		0	1		62	38																		
55	1		-19	-22	-23	49*		1	3		47	53																		
56	1		-15	-20	45*	-25		1	3		47	53																		
57	1		32*	-11	-24	-13		0	1		33	68																		
58	1		-17	34*	-14	-16		1	2		34	66																		
59	1		-21	-22	-21	46*		0	1		45	55																		
60	1		48*	-22	-20	-25		1	2		41	59																		

AC

Code legend: A: Rar >= Rir B: Rir <= 0 C: Rar >= 10

SubGroup number	: 0	SubTest number	: 5
Number of persons in test	: 240	Number of selected items	: 12
Minimum test score	: 0	Maximum test score	: 12
Average test score	: 5.83	Standard deviation	: 2.83
Average P-value	: 48.58	Std. Error of Measurement	: 1.54
Average Rit	: 0.49		
Coefficient Alpha	: 0.70	SE Coeff. Alpha	: 0.03
GLB	: 0.89	Asymptotic GLB coef	: 0.84
Items used in GLB proc	: 12		

90% Confidence limits for Coefficient Alpha: (0.66 =< 0.70 =< 0.75)

Estimated Coefficient Alpha if this test had a standard norm length of 40 items: 0.89 (Spearman-Brown)

TiaPlus® Test and Item Analysis Build 303  
 Cito, Measurement and Research Department. Arnhem, the Netherlands. © 2007.  
 Population : Eunice and Heidedal Grade 10s  
 Test : **TALA-like test for academic literacy at secondary school**  
 Date : 28 August 2014  
 Time : 16:08  
 Data file : C:\AlbertDocuments\ICELDA\ice01171.txt  
 Missing handling : Missing as Zero  
 Persons: All persons Items: All items

Table of Subtest Intercorrelations

Subtest	Total test	Subtest(s) 1	2	3	4	5
Scrambled text	1	0.64				
Vocabulary know	2	0.70	0.42			
Interpreting gr	3	0.81	0.50	0.49		
Text comprehens	4	0.89	0.43	0.57	0.64	
Grammar & text	5	0.74	0.33	0.37	0.52	0.54
Number of testees :	240	240	240	240	240	240
Number of items :	60	5	10	8	25	12
Average test score:	33.23	2.51	6.23	5.00	13.65	5.83
Standard deviation:	10.53	1.99	1.75	2.43	4.44	2.83
SEM :	3.40	0.74	1.30	1.09	2.22	1.54
Average P-value :	55.39	50.25	62.33	62.55	54.62	48.58
Coefficient Alpha :	0.90	0.86	0.44	0.80	0.75	0.70
GLB :	0.97	0.90	0.64	0.85	0.86	0.89
Asymptotic GLB :	0.96	0.90	0.53	0.84	0.84	0.84

TiaPlus® Test and Item Analysis Build 303  
 Cito, Measurement and Research Department. Arnhem, the Netherlands. © 2007.  
 Population : Eunice and Heidedal Grade 10s  
 Test : TALA- like test (Gadgets... ) for academic literacy at secondary school  
 Date : 02 September 2014  
 Time : 11:41  
 Data file : C:\AlbertDocuments\ICELDA\ice01195.txt  
 Missing handling : Missing as Zero  
 Mantel-Haenszel DIF statistics  
 Information on Subtest: All items  
 Comparing subgroups 1 vs 2 (Eunice - Heidedal)

Label	Item	DIF stat	z (stand)
	1	0.5379	-0.6564
	2	0.6891	-0.4254
	3	0.7513	-0.2725
	4	0.4853	-0.7330
	5	0.4151	-1.0112
	6	0.4519	-0.8701
	7	1.2683	0.2017
	8	2.3667	0.9774
	9	1.1812	0.1919
	10	3.6466	1.6117
	11	1.3171	0.3357
	12	1.3362	0.2147
	13	0.7425	-0.2680
	14	2.9985	1.3601
	15	1.9088	0.7606
	16	1.5036	0.4755
	17	0.9062	-0.1166
	18	4.9379	1.9185
	19	2.2387	1.0020
	20	3.5790	1.3785
	21	1.7843	0.6231
	22	0.8916	-0.1258
	23	1.2386	0.2545
	24	1.2070	0.2140
	25	0.4883	-0.8806
	26	1.0745	0.0821
	27	0.7468	-0.3492
	28	0.5218	-0.6695
	29	1.6042	0.5809
	30	1.9921	0.8765
	31	2.5580	1.1555
	32	0.6095	-0.4405
	33	3.4398	1.5272
	34	2.3952	1.0690
	35	0.8332	-0.2261
	36	2.0159	0.8657
	37	0.7038	-0.4133
	38	0.6971	-0.4261
	39	0.6215	-0.5710
	40	1.1441	0.1686
	41	1.8611	0.7779

Label	Item	DIF stat	z (stand)
	42	0.3732	-0.9644
	43	0.6589	-0.5143
	44	0.3331	-1.0573
	45	0.5726	-0.7103
	46	0.7406	-0.3454
	47	1.6141	0.5858
	48	1.1646	0.1804
	49	1.1702	0.1813
	50	0.7603	-0.3295
	51	0.9006	-0.1200
	52	2.1752	0.9421
	53	2.6624	1.1094
	54	1.3715	0.3524
	55	2.5562	1.1498
	56	2.7057	1.2244
	57	1.0490	0.0618
	58	1.3302	0.3717
	59	1.5471	0.5376
	60	1.5202	0.5024



Interpretation:

If the DIF statistic is  $< 1$  then the studied item is more difficult in the first subgroup.

If the DIF statistic is approx. 1 then the studied item has equal difficulty for both subgroups.

If the DIF statistic is  $> 1$  then the studied item is more difficult in the second subgroup.

Significance (at alpha level = 1%):

Differences between subgroups are significant when the absolute value of  $z(\text{stand}) \geq 2.58$

('--' is shown if TiaPlus can not calculate the statistic).

Note that in case of subtest processing the result for an item will change as a subtest has a total score that differs from the total test total score. The group of persons therefore will be partitioned differently.

Cito, Measurement and Research Department. Arnhem, the Netherlands. © 2007.

## **Annexure H**

### **Correlational analysis**

Marginal Correlations (NOT adjusting for the School effect) between Test1, Test2, English and ave\_

## The CORR Procedure

4 Variables: ave\_without\_eng test1 test2 english

Pearson Correlation Coefficients, N = 238

Prob > |r| under H0: Rho=0

	ave_without_ eng	test1	test2	english
ave_without_eng Ave without ENG	1.00000	0.45512 <.0001	0.78491 <.0001	0.81810 <.0001
test1 Test1	0.45512 <.0001	1.00000	0.35253 <.0001	0.31814 <.0001
test2 Test2	0.78491 <.0001	0.35253 <.0001	1.00000	0.78408 <.0001
english English	0.81810 <.0001	0.31814 <.0001	0.78408 <.0001	1.00000

Partial Correlations (adjusting for the School effect) between Test1, Test2, English and ave\_witho

## The CORR Procedure

1 Partial Variables: school  
 4 Variables: ave\_without\_eng test1 test2 english

Pearson Partial Correlation Coefficients, N = 238  
 Prob > |r| under H0: Partial Rho=0

	ave_without_ eng	test1	test2	english
ave_without_eng Ave without ENG	1.00000	0.04069 0.5331	0.63060 <.0001	0.74331 <.0001
test1 Test1	0.04069 0.5331	1.00000	0.04284 0.5116	0.02039 0.7549
test2 Test2	0.63060 <.0001	0.04284 0.5116	1.00000	0.67106 <.0001
english English	0.74331 <.0001	0.02039 0.7549	0.67106 <.0001	1.00000

## **Annexure I**

### **Regression analysis**

Regression analysis

The REG Procedure

Model: MODEL1

Dependent Variable: ave\_without\_eng Ave without ENG

Number of Observations Read 238  
 Number of Observations Used 238

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	55660	13915	360.83	<.0001
Error	233	8985.28375	38.56345		
Corrected Total	237	64645			

Root MSE 6.20995 R-Square 0.8610  
 Dependent Mean 58.22881 Adj R-Sq 0.8586  
 Coeff Var 10.66473

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	39.96217	3.90967	10.22	<.0001
school	School	1	-16.50002	1.21402	-13.59	<.0001
test1	Test1	1	0.02577	0.05865	0.44	0.6607
test2	Test2	1	0.27200	0.06493	4.19	<.0001
english	English	1	0.58198	0.05689	10.23	<.0001

Regression analysis

The REG Procedure

Model: MODEL2

Dependent Variable: ave\_without\_eng Ave without ENG

Number of Observations Read 238  
 Number of Observations Used 238

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	54983	18328	443.87	<.0001
Error	234	9661.98496	41.29053		
Corrected Total	237	64645			

Root MSE 6.42577 R-Square 0.8505  
 Dependent Mean 58.22881 Adj R-Sq 0.8486  
 Coeff Var 11.03538

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	42.19211	4.00787	10.53	<.0001
school	School	1	-17.85858	1.21056	-14.75	<.0001
test1	Test1	1	0.03544	0.06064	0.58	0.5595
english	English	1	0.74188	0.04365	16.99	<.0001

Regression analysis

The REG Procedure

Model: MODEL3

Dependent Variable: ave\_without\_eng Ave without ENG

Number of Observations Read 238  
 Number of Observations Used 238

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	55652	18551	482.71	<.0001
Error	234	8992.73089	38.43047		
Corrected Total	237	64645			

Root MSE 6.19923 R-Square 0.8609  
 Dependent Mean 58.22881 Adj R-Sq 0.8591  
 Coeff Var 10.64633

Parameter Estimates

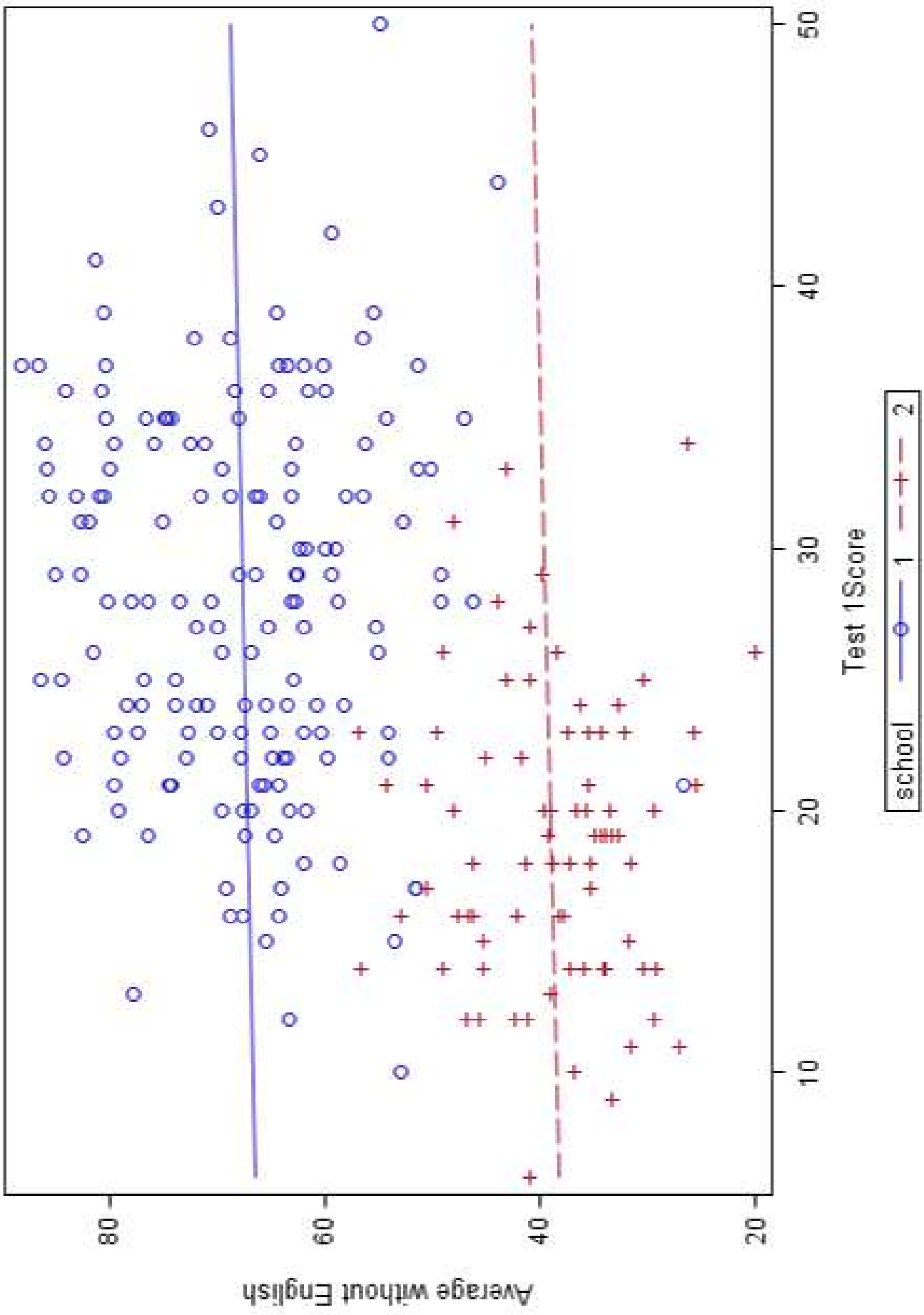
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	40.88988	3.28504	12.45	<.0001
school	School	1	-16.72666	1.09714	-15.25	<.0001
test2	Test2	1	0.27312	0.06477	4.22	<.0001
english	English	1	0.58169	0.05679	10.24	<.0001



## **Annexure J**

### **ANCOVA analysis (1)**

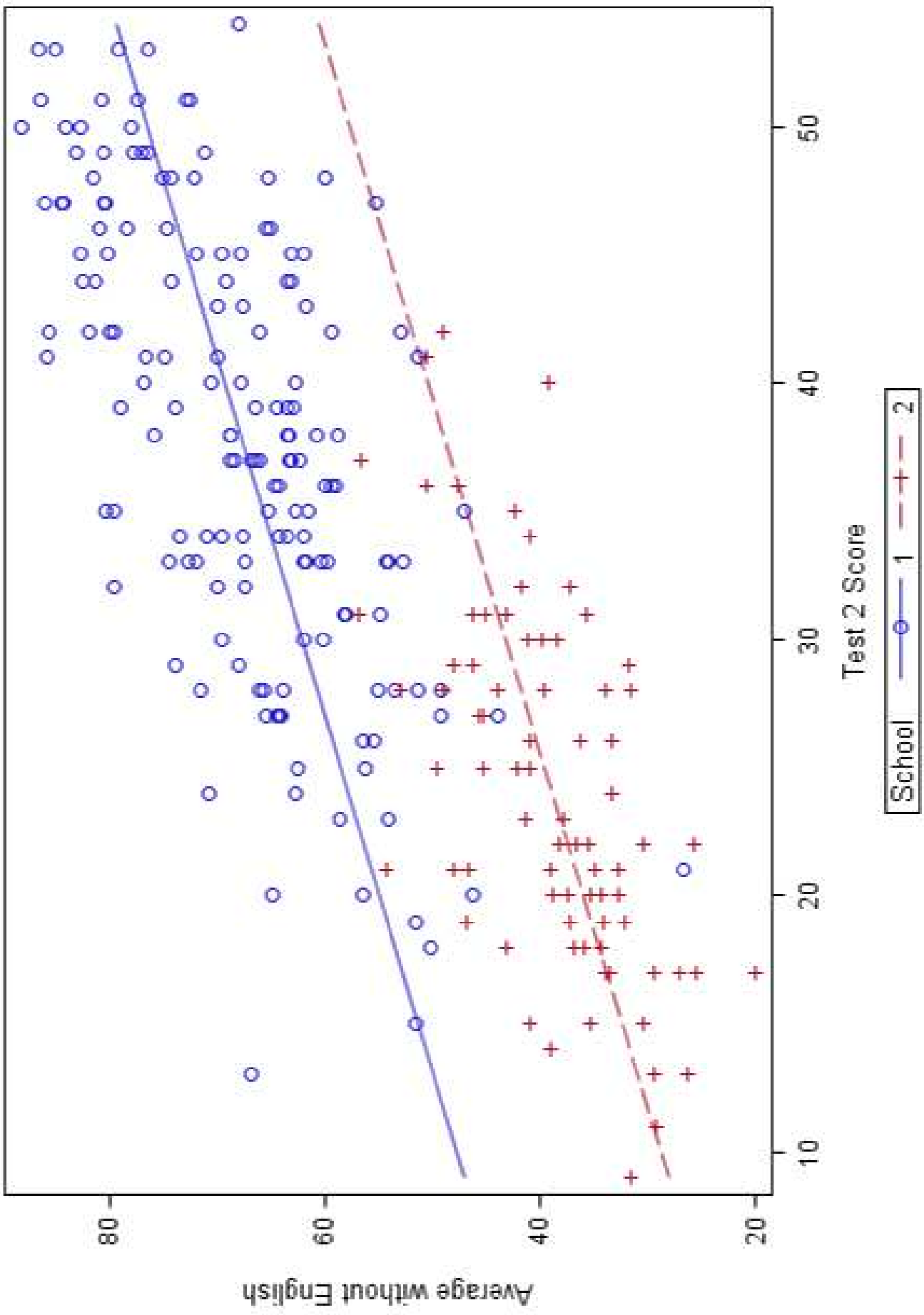
**Analysis of Covariance: Average without English versus Test 1 Score**



## **Annexure K**

### **ANCOVA analysis (2)**

**Analysis of Covariance: Average without English versus Test 2 Score**



## **Annexure L**

### **ANCOVA analysis (3)**

**Analysis of Covariance: Average without English versus English Score**

