

*Kabelo Sebolai*

# Validating a test of academic literacy at a South African university of technology

First submission: 20 September 2012

Acceptance: 8 August 2013

The advent of a democratic dispensation in South Africa in 1994 officially made it possible for historically disadvantaged groups, Black students in particular, to gain access to higher education at historically White universities. These universities, however, use English for teaching and learning, a second language to the majority of these students. Moreover, such students are products of Bantu Education, a tool used by the apartheid government to ensure that they leave secondary school with poor English skills. Both these factors constitute an obstacle to the students' capacity to handle the demands of university education in English. South African universities have responded to this challenge by introducing academic language programmes to help the students bridge the language gap between high school and university education. Most of these programmes mainly focus on the teaching of reading, writing and thinking in English, a combination of skills known as academic literacy in the South African higher education context. This study explores the validity of a test of academic literacy used for summative assessment at a university of technology. Evidence suggests that the test possessed an acceptable degree of validity.

## Die geldigheid van 'n toets vir akademiese geletterdheid by 'n Suid-Afrikaanse universiteit van tegnologie

Die totstandkoming van 'n demokratiese bestel in Suid-Afrika in 1994 het dit amptelik moontlik gemaak vir historiesbenadeelde groepe, Swart studente in die besonder, om toegang te verkry tot hoër onderwys by histories Wit universiteite. Hierdie universiteite gebruik Engels vir onderrig en leer, 'n tweede taal vir die meerderheid van die betrokke studente. Boonop is hierdie studente produkte van Bantu Onderwys, 'n middel wat aangewend is deur die Apartheidsregering om te verseker dat hulle hoërskool verlaat met gebrekkige Engelse vaardighede. Genoemde faktore behels 'n struikelblok in die studente se vermoë om te voldoen aan die eise van Universiteitsopleiding in Engels. Suid-Afrikaanse universiteite het gereageer op hierdie uitdaging deur akademiese taalprogramme in te stel om die studente te help om die taalgaping tussen hoërskool- en universiteitsonderrig te oorbrug. Hierdie programme fokus hoofsaaklik op die onderrig van lees, skryf en dink in Engels, 'n kombinasie van vaardighede wat bekend staan as akademiese geletterdheid in die Suid-Afrikaanse hoër onderwys konteks. Hierdie studie verken die geldigheid van 'n toets vir akademiese geletterdheid wat gebruik word vir summatiewe assessering by 'n universiteit van tegnologie. Die resultate dui daarop dat die toets oor 'n aanvaarbare vlak van geldigheid beskik.

*Mr K Sebolai, Academic Development and Support, Central University of Technology, Free State, Private Bag X20539, Bloemfontein, 9300; E-mail: ksebolai@cut.ac.za*



*Acta Academica*  
2013 45(3): 215-241  
ISSN 0587-2405  
© UV/UFS  
<<http://www.ufs.ac.za/ActaAcademica>>

SUN MØDIA  
BLOEMFONTEIN

The post-1994 period has opened doors for historically disadvantaged groups, Black students in particular, to gain admission to historically White universities where English is used as a medium of instruction. The Bantu Education Act of 1953 prevented such admission. Unfortunately, the majority of these students speak English as a second language and, as a result of the poor English tuition at high school, they are not adequately prepared to handle the demands of university education in English. South African universities have responded to this challenge by introducing academic literacy (AL) programmes to empower these students with the reading, writing and thinking skills required for success at university study. Logically, the curricula developed for such programmes have to be informed by how AL is currently defined within the higher education context in which such universities operate. Similarly, after the students have been taught in such programmes, the tests used to decide on the level of AL should be based on how AL is conceptualised and taught. The Central University of Technology's (CUT) AL programme was established at the beginning of 2007. This one-year course has, since 2009, been officially declared compulsory for all students entering CUT. In other words, students have to complete this course successfully before they are allowed to graduate. This can only be fair if the tests on the basis of which this decision is taken are truthful or possess what is technically known as validity. This study aims to investigate the degree to which the AL test used at CUT at the end of the first semester in 2012 truthfully measured what it was developed to test. The following section briefly explores the definition and classification of the term 'validity'.

## 1. Validity

The concept of validity is probably the most crucial and contested of all principles governing the design and development of tests. Traditionally, validity referred to the question of whether a test measures what it is intended to measure. Based on this definition, a test is valid if it measures what it purports to measure (Kelley 1927; Cattell 1964; Lado 1961). In other words, such a test restricts itself to "measuring only what it is intended to test and not extraneous

or unintended abilities” (Weir 1993: 19). In this sense, validity is a property of the test involved.

This definition has been challenged. For example, Messick (1989) associated validity with how test scores are interpreted and used, and not necessarily with the test yielding such scores. This is captured in his definition of the term:

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

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

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

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

A dimension of Messick’s (1989) framework, which introduced a new perspective to the way validity had been defined, was his

association of this concept with the consequences of how test scores are interpreted and used. Messick (1980: 1012) contended that “not only should tests be evaluated in terms of their measurement properties, but that testing applications should be evaluated in terms of their potential social consequences”. Considering these “extravalidity concerns” is in Gregory’s (2007: 139) view, a test designer’s way of acknowledging that testing has consequences that are unrelated to a test’s psychometric soundness. Bachman & Palmer (1996: 30) supported this view by arguing that,

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

These consequences relate directly to the stakes that accompany a decision taken on the basis of a particular test’s scores (Miller et al. 2009) and, therefore, determine the type of measurement instrument used and the quantity of the resources expended on the development of such a tool. The higher the stakes attached to a test, the more important it is that its consequences be taken into account in the assessment of its overall validity (Messick 1989). As the phrase implies, ‘high stakes’ decisions include those that will have a negative impact on a vast number of people (Bachman & Palmer 1996: 96-7). In most instances, such decisions cannot be rescinded and can, therefore, have a lifetime negative impact on the lives of those involved (Bachman & Palmer 1996: 97). It is thus important that high-stakes decisions taken on the basis of test scores be taken wisely and in harmony with the purpose for which a test was designed (Stoynoff & Chappelle 2005).

While Messick’s (1989) consequential dimension of validity has not generated any opposition from scholars in the fields of educational and psychological assessment, his inclination to associate validity solely with test scores and not the test whereby such scores are generated has been challenged by language-testing scholars, in particular. For example, Davies & Elder (2005: 279) argued that

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

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

Borsboom et al. (2004: 279) similarly argued that a test used many times for a similar purpose meets the psychometric requirement of validity if no evidence exists to show that it is used for purposes for which it was not designed. In Borsboom et al.'s (2004: 279) words, it should be possible to "speak of the validity of that particular test – as a characteristic of it". In addition, Weideman (2012) challenged Messick's (1989) insistence on associating validity with test scores and not the measurement instrument itself, by arguing that this drives attention away from the importance of the psychometric soundness of such an instrument. Weideman (2012) rightly points out that no matter how good the interpretation of a set of test scores is, if the measurement instrument is not technically sound, this interpretation is useless to the test user: "No amount of interpretation can improve the measurement result (score) obtained from an inadequate instrument that gives a faulty and untrustworthy reading" (Weideman 2012: 4). In the light of this, Weideman (2012: 6) argued for the need to distinguish between the objective effect of a test and the subjective interpretation of its scores. Weideman (2012) further argued that through his use of the word 'adequacy' in his definition of validity, Messick inadvertently attaches validity to the measurement instrument and not the interpretation of the scores from such an instrument as he claims he does. Adequacy is, in Weideman's (2012) thinking, a word conceptually appropriate to describe a test and not the interpretation of its scores: "... using validity as descriptive of a test therefore merely returns in another guise, that of adequacy ..." (Weideman 2012: 6). In other words, Weideman (2012) believes that Messick's definition of validity simply constitutes a circumlocution aimed at obfuscating the traditional definition of validity as a property of a test.

Traditionally, validity has been categorised into three types, namely the content, construct and criterion-related types. Scholars in the fields of educational and psychological measurement have viewed these concepts differently. The following section briefly explores how each of the concepts was traditionally defined and the current debate on what they mean.

## 1.1 Content validity

Content validity is a term traditionally used to refer to the degree to which tasks in a test adequately represent the universe of the content that allows test designers to capture ample construct or knowledge which they wish to measure (Cohen & Swerdlik 2010). Content validation is, in this sense, inherently a content sampling exercise that needs to be carried out with care if any claim is to be made that a test possesses content validity:

The essence of content consideration in validation, then, is determining the adequacy of the sampling of the content that the assessment results are interpreted to represent. More formally, the goal in the consideration of content validation is to determine the sample of the domain tasks about which interpretations of assessment results are made (Miller et al. 2009: 75).

In language testing, this sampling exercise involves considering the characteristics of the language tasks typical of what Bachman & Palmer (1996) call the Target Language Use (TLU) domain. TLU refers to the particular real-life situation in which the test taker will use language. This means that for the purpose of ensuring content validity, language test designers are obliged to ensure that the characteristics of their test tasks reflect those typical of the tasks inherent to a TLU domain. Bachman & Palmer (1996) refer to this correspondence between test tasks and the specified TLU domain as authenticity, namely “the degree of correspondence of the characteristics of a given language test task to the features of the TLU task” (Bachman & Palmer 1996: 23).

## 1.2 Construct validity

Construct validity is probably the most important of all traditional classifications of the concept of validity. This term refers to the degree to which a theory underpinning a test designed to measure an ability can be justified. According to Stoyneff & Chapelle (2005: 17), construct validity relates to the “extent to which evidence suggests that the test measures the construct it is intended to measure, in other words, that inference specified as one facet of test purpose is justified”. This means that testers “need to be precise about what a test is intended to measure” and should “develop the conceptual apparatus to do so” (Chapelle & Bridley 2002: 269). In other words, a construct first has to be defined and evidence subsequently produced

in order to demonstrate that a test measures the ability it purports to measure. In language testing, Bachman & Palmer's (1996) notion of a TLU is crucial to both the definition and the validation of a construct. Thus, not only is authenticity a function of content validity, it is also inherent to construct validity.

### 1.3 Criterion-related validity

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

Messick (1980) argues against this traditional categorisation of validity into the three types dealt with so far. Instead, he views validity as a single unifying concept that does not need to be compartmentalised in this manner. According to Messick (1980: 1014), the problem with this classification is that,

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

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

evidence includes “a consideration of the content measured, the ways in which students respond, the relationship of individual items to the test scores, the relationship of performance to other assessments, and the consequences of using and interpreting assessment results” (Miller et al. 2009: 73).

Weideman (2012) has contested Messick’s unitary approach to validity by arguing that it is a conflation of what he terms the regulative and constitutive concepts of responsible test design. Weideman’s (2009: 1) constitutive requirements include systematicity, reliability, the three traditional types of validity, and the meaningfulness of test results, while the regulative conditions are constituted by accessibility, transparency and accountability. Unlike Messick (1980; 1989), Weideman (2009; 2012) argues against subsuming all these conditions under a single concept such as construct validity. He believes that better conceptual clarity is achievable only if each of the constitutive and regulative conditions is recognisable as a critical factor in responsible test design and appraisal. Finally, Weideman (2012) observes that efforts by scholars such as Kane (1992), Bachman & Palmer (1996) and Kunnan (2000) to reinterpret Messick’s unitary concept of validity are so disunited that the need to distinguish between constitutive and regulative conditions of test design is inadvertently laid bare. Weideman (2012: 8) adds that these attempts at reinterpreting Messick’s framework are “far from helping us” achieve conceptual clarity and that instead, “they may help more to confuse” us.

In agreement with the traditional classification of validity into the content, construct and criterion-related types as well as the arguments currently advanced in favour of the need for testers to distinguish between the constitutive and regulative constituents of responsible test design, the present study focuses on determining the degree of concurrent validity possessed by a summative test of AL used at CUT at the end of the first semester in 2012.

## 2. Description of the sample

The sample used for this study consisted of male and female first-year students between the ages of 18 and 21, 55 of whom enrolled in the Hospitality Management, 44 in the Public Management and 43 in the



Mechanical Engineering programmes at CUT in Bloemfontein, South Africa. The sample consisted of 142 participants in total. The students had successfully completed their Grade 12 examination the previous year and had subsequently gained admission to the university. Of the participants, 122 were from Sotho, Tswana and Xhosa first-language backgrounds, 17 spoke Afrikaans as a first-language, and 3 spoke English as a first-language. The students were chosen for participation in this study because of the availability of their scores to the researcher. They were all taking the AL course offered at CUT. The Hospitality Management group took the class from the researcher himself, while the other two groups were taught by two other teachers under the supervision of the researcher.

### 3. Procedure

For the first time since their introduction at other South African universities at the beginning of 2011, the National Benchmark Tests (NBTs) were administered to a total of 2007 first-year students at CUT towards the end of March 2012. The timing of the administration of the tests indicates that the scores from these tests were not aimed at being used for making admission decisions. Instead, in compliance with the original idea behind the National Benchmark Test Project at CUT, the NBTs were administered with the aim to use the scores they would yield in order to estimate the general academic preparedness of the test takers in the three domains of interest, namely AL, quantitative literacy, and mathematical literacy. For the purpose of accomplishing this study, the AL test of the NBTs was used as the criterion for assessing the criterion-related validity of the summative AL test administered to students enrolled in the Academic Literacy Programme (ALP) at the end of the first semester at CUT in 2012. The criterion was chosen on the basis of its being a test of AL that was standardised, currently used countrywide for placement and admission decision-making purposes, and whose psychometric soundness had presumably been established.

Table 1 shows that the mean score obtained by the 2007 first-year students at CUT in 2012 in the AL test of the NBTs was 43.1 and the standard deviation was 10.2. These descriptive statistics attest to the low levels of AL among students admitted at CUT in 2012.

Table 1: Mean and standard deviation of the scores from the AL test of the NBTs for first-year students at CUT in 2012 (N=2007).

Variable	Mean score	Standard deviation	Maximum	Minimum
AL NBTs	43.1	10.2	83	22

Prior to this, two administrations of a standardised test of AL, called the Placement Test in English for Educational Purposes (PTEEP), in 2007 and 2008, revealed the low AL levels among first-year students at CUT. The PTEEP is no longer used and has been replaced by the AL test of the NBTs. Like the AL test of the NBTs, the PTEEP was developed by the AARP of the University of Cape Town (UCT) and designed on the basis of the same construct of AL as the AL test of the NBTs. Cliff & Yeld (2006) describe this construct as consisting of the test taker’s ability to

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

Tables 2 and 3 indicate the descriptive statistics of both administrations of the PTEEP referred to earlier.

Table 2: Mean and standard deviation of the scores from the PTEEP for first-year students at CUT in 2007 (N=408).

Variable	Mean score	Standard deviation	Maximum	Minimum
PTEEP	39.1	12.4	73.4	8.5

Table 3: Mean and standard deviation of the scores from the PTEEP for first-year students at CUT in 2008 (N=1549)

Variable	Mean score	Standard deviation	Maximum	Minimum
PTEEP	40.0	11.7	81.9	5.6

The consistently low AL levels of students admitted at CUT justify the existing ALP introduced by the then Unit for Academic Development after the first administration of the PTEEP at CUT referred to earlier. As its full name implies, the ALP was introduced with the aim to equip students with the ability to meet the reading, writing and thinking demands of higher education in the chosen language of teaching and learning. Logically, this ability would be constituted by the same skills that both the PTEEP and the AL test of the NBTs purported to measure. Whether the ALP has been doing this so far is an interesting subject of investigation for another study. However, the current study sought to establish whether a test of AL used at the end of the first semester at CUT in 2012 could be validated against the AL test of the NBTs. In other words, the nationally used and standardised AL test of the NBTs was used as a criterion against which the interpretations of the ALP end-of-first-semester summative test scores could be justified. This was, in the researcher's view, crucial because, since the introduction of the programme in 2007 at CUT, students who enrol in the programme at the beginning of the first semester and who obtain an average score of 75% and above from both the formative and summative assessment combined are exempted from continuing with the course in the second semester. In other words, students who are able to obtain this score are considered to be adequately academically literate and are consequently exempted from taking classes in the ALP. Similarly, students who do not complete the course successfully in any or all of the two semesters are required to

repeat it before they are allowed to graduate. These are the two critical decisions taken on the basis of the scores about the academic lives of the students involved. It is thus imperative that the adequacy or validity of the instruments used be ascertained.

As a test of AL, the end-of-first-semester summative test of the ALP used in 2012 mainly aimed to measure the reading, writing and thinking skills of students in academic English. The test consisted of two sections, the first of which mainly focused on testing the students' reading ability, while the second tested writing. The total mark allocated to reading comprehension was 80 and that allocated to writing was 20. The reading part of the test consisted of four questions using different formats in order to test reading. These were 10 multiple choice items; a close procedure of 20 items in which mainly content words were systematically removed from the text and had to be restored by the test taker; 20 scrambled sentences from two successive paragraphs of the same passage which also had to be restored to their original order, and a passage from which some sentences had been systematically removed and had to be restored to convey the original meaning of the passage. The Flesch Kincaid Grade Level of all the reading passages used ranged from 7.2 to 10.5. For the section on writing, the test takers were required to write one cohesive and coherent paragraph on one of the four topics provided (see Appendix A). For the purpose of assessing this piece of writing, Hughey et al.'s (1983) scoring rubric was used (see Appendix B). The break-down of the marks for all the language-related writing abilities assessed by means of the rubric was adjusted to suit the total mark of 20 that was allocated for the writing section of the ALP test. For example, content was holistically marked out of 8, organisation out of 6, and vocabulary, language and mechanics out of 2 marks each. For the purpose of accomplishing the aim of this study, however, the participants' scores from the writing part were not considered. The reason for this was to check the possible inter-scorer inconsistencies that could contaminate the finding of the study. As pointed out earlier, the three groups of participants in this study were taught by three teachers whose assessment of the paragraphs could be different. For the purpose of this study, the participants' average score on the ALP summative test was, therefore, only worked out from their performance in the reading part of the test, whose total constituted 80% of the entire test.

After the test was administered and marked, the descriptive statistics of the scores were computed. This is captured in Table 4.

Table 4: Mean and standard deviation of the scores from the ALP summative test administered to the Hospitality Management, Public Management and Mechanical Engineering students at CUT in May 2012 (N=142).

Variable	Mean score	Standard deviation	Maximum	Minimum
ALP AL test	41.8	11.8	73	14

These statistics closely resemble those yielded by the AL test of the NBTs for the same group of test takers earlier in the same year. This is evident in Table 5.

Table 5: Mean and standard deviation of the scores from the AL test of the NBTs administered to the Hospitality Management, Public Management and Mechanical Engineering students at CUT in March 2012 (N=142).

Variable	Mean score	Standard deviation	Maximum	Minimum
AL test NBTs	44.6	10.3	83	27

The standard procedure for establishing criterion-related validity involves running a correlation study of a test under study and the criterion measure used to validate it. Correlation is a statistical procedure that enables a researcher to “look at two variables and evaluate the strength and direction of their relationship or association with each other” (Dornyei 2007: 223). In this context, a correlation analysis is used to compute the test’s correlation coefficient which is also known as a validity coefficient. The correlation coefficient is, therefore, a statistical summary of the extent of the relationship or association between scores obtained on a test and the criterion measure (Miller et al. 2009). The validity coefficient can range from -1 to +1. A correlation coefficient of +1 signals a perfect positive association, whereas that of -1 indicates a perfect negative relationship between the variables involved (Miller et al. 2009). Mackey & Gass (2005: 286) explain the meaning of the difference between positive and negative correlation thus:

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

Lastly, a correlation coefficient of zero means that there is no relationship between the variables under study.

Salkind (2011) suggests that one should never expect perfect association between any two variables, especially in the behavioural and social sciences. As Salkind (2011: 85) points out, “values approaching .7 and .8 are just about the highest you will see”. Dornyei (2007: 223) similarly observed that “in applied linguistics research we can find meaningful correlations of as low as 0.3-0.5 [...] and if two tests correlate with each other in the order of .60, we can say that they measure more or less the same thing.”

#### 4. Findings

A correlation was computed in order to determine whether there was a statistically significant association between the scores on the ALP summative test administered at CUT in May 2012 and performance in the AL test of the NBTs administered at the same institution in March 2012. The Pearson Correlation ( $r$ ) statistic was calculated and found to be .67, and the probability ( $p$ ) value equalled 0.01. This correlation coefficient is above .60 and close to .70. As Dornyei (2007: 223) observed, “... if two tests correlate with each other in the order of 0.6, we can say that they measure more or less the same thing”. The  $p = 0.01$  value for statistical significance means that the probability that the findings of this study are due to chance alone is far below 5%. This is also known as internal validity (Mackey & Gass 2005). Finally, the correlation was positive, which means that students who performed well in the ALP summative test also tended to do well in the AL test of the NBTs, and that those who underperformed in any of the two tests tended to perform similarly in the other. This is graphically captured in Figure 1.

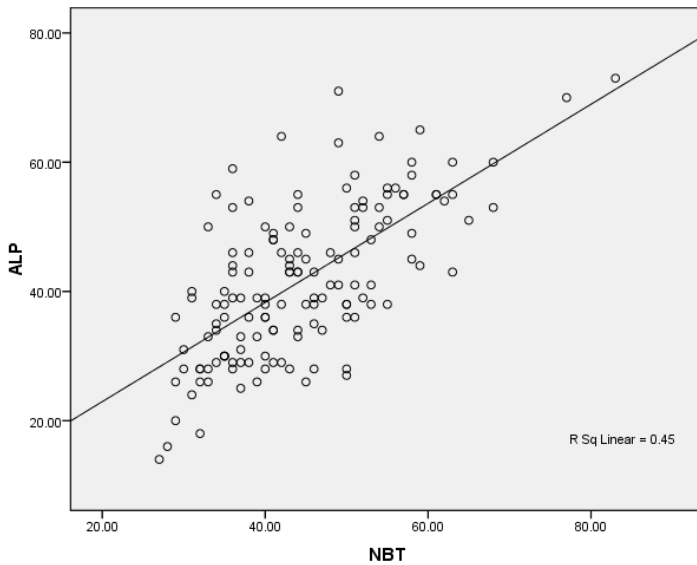


Figure 1: Correlations of the ALP summative test scores in May 2012 with scores on AL test of the NBTs in March 2012 (N=142)

## 5. Implications

The finding of this study has a number of implications for CUT and other universities in South Africa. First, the close association between the descriptive statistics of the two tests for the same group of students implies that the tests probably tested similar or related constructs and that they were pitched at the same level of difficulty for the targeted test takers. The nearly similar means and standard deviations yielded by the two tests also imply that the scores from these tests were reasonably widely and similarly distributed. This constitutes evidence that the tests probably possessed a reasonable degree of consistency, and the standard deviation in particular points to the probability that the tests could also be used for the placement of students into a programme such as the ALP. Placement tests should preferably have standard deviations of a similar magnitude. In fact, not only are the

descriptive statistics of the scores on the ALP test consistent with those of the scores from the AL test of the NBTs for the sample used in the study, the statistics also reflect the scores obtained by all the 2007 students who took the AL test of the NBTs at CUT in 2012. This is further testimony to the probable reliability and criterion-related validity of the ALP test of AL as the focus of this article. Table 1 presents the descriptive statistics for this entire cohort of test-takers.

Secondly, the high correlation between the two sets of scores attests to the efficiency of the four AL testing techniques employed in the ALP summative test. Three of these techniques are not used in the AL of the NBTs. The tasks in the latter test are mainly presented in multi-choice format. The fact that this variety of test tasks used in the two tests yielded similar descriptive statistics for the same group of test takers and that the correlation or validity coefficient of the ALP summative test is statistically significant has implications beyond CUT. Currently, AL is a very important issue at South African universities. Not only do these universities need to make research efforts to establish how effectively this skill can be taught, they also need to improve and diversify ways to test it. The finding of this research contributes to this last aspect.

The final implication of the main finding of this study for CUT, in particular, is that if the scores obtained by the sample from the two tests have shown so much evidence of validity, the ALP course taught to the students throughout the first semester did not make any significant difference in their growth as academic language learners. This is, undeniably, a very important and credible inference which is worth investigating in another study. One should hasten to point out, at the same time, however, that the administration of the AL test of the NBTs and the summative test of the ALP were barely two months apart and that it would be at odds with reality for anybody to expect the amount of teaching that took place within this period to impact on student learning in any significant manner. This would be at variance with evidence from second-language acquisition research. This is applicable to CUT especially, where only two hours are allocated for academic language development on the teaching schedule.



## 6. Conclusion



Being academically literate in the language of teaching and learning is, without question, critical for student success at institutions of higher learning in South Africa. This means that efforts should be made to understand the skills that constitute the reading, writing and thinking skills of AL. Standardised AL test designers and developers such as the Alternative Admissions Research Project at UCT and the Inter-Institutional Centre for Language Development and Assessment, a partnership of the Universities of Pretoria, Stellenbosch, North-West and Free State, have played their part in this regard. However, the AL tests these companies develop are mainly used for the placement of students into AL programmes by South African universities and are rarely used for assessing formative and summative AL achievement in such programmes. The finding of this study highlights the important need for the alternative and locally developed achievement tests used in such programmes to be validated against standardised and established tests of AL such as the AL test of the NBTs. This should, in turn, have positive feedback on the curricula that are used for AL intervention. In other words, such curricula interventions should be designed and developed on the basis of the same construct that informs the particular test of AL used for placing students. This is relevant to CUT, in particular, and other South African universities, in general.

## Bibliography

- BACHMAN L  
2004. *Statistical analysis for language assessment*. Cambridge: Cambridge University Press.
- BACHMAN L F & PALMER A S  
1996. *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press.
- BORSBOOM D, G J MELLENBERGH & J VAN HEERDEN  
2004. The concept of validity. *Psychological Review* 111(4): 1061-71.
- CATTELL R B  
1946. *Description and measurement of personality*. New York: World Book Company.
- CHAPELLE C A & G BRIDLEY  
2002. Assessment. Schmitt (ed) 2002: 267-88.
- CLIFF A F & N YELD  
2006. Test domains and constructs: academic literacy. Griesel (eds) 2006: 19-27.
- COHEN R J & M E SWERDLIK  
2010. *Psychological testing and assessment*. New York: McGraw-Hill.
- DAVIES A & C ELDER  
2005. Validity and validation in language testing. Hinkel (eds) 2005: 795-813.
- DAVIES A & C ELDER (eds)  
2004. *The handbook of applied linguistics*. Malden: Blackwell Publishing.
- DORNYEI Z  
2007. *Research methods in applied linguistics*. Oxford: Oxford University Press.
- GREGORY R J  
2007. *Psychological testing: history, principles and applications*. New York: Pearson.
- GRIESEL H (ed)  
2006. *Access and entry level benchmarks: the national benchmark tests project*. Pretoria: Higher Education South Africa.
- HINKEL E (ed)  
2005. *Handbook of research in second language teaching and learning*. Mahwah, NJ: Lawrence Erlbaum Associates.
- HUGHEY J B, D R WORMUTH, V F HATFIELD & H L JACOBS  
1983. *Teaching ESL composition: principles and techniques*. Rowley, MA: Newbury House.
- KANE M T  
1992. An argument-based approach to validity. *Psychological Bulletin* 112(3): 527-35.
- KELLY T L  
1927. *Interpretation of educational measurement*. New York: Macmillan.
- KUNNAN A J  
2000. Fairness and justice for all. Kunnan (ed) 2000: 1-14.
- KUNNAN A J (ed)  
2000. *Fairness and validation in language assessment: selected papers from the 19<sup>th</sup> Language Testing*

- Research Colloquium, Orlando, Florida.* Cambridge: University of Cambridge Local Examinations Syndicate.
- LADO R  
1961. *Language testing: the construction and use of foreign language tests.* New York: McGraw-Hill.
- LINN R L (ed)  
1989. *Educational measurement.* 3<sup>rd</sup> ed. New York: American Council of Education/Collier Macmillan.
- LYNCH B K  
2003. *Language assessment and program evaluation.* Edinburgh: Edinburgh University Press.
- MACKAY A & S M GASS  
2005. *Second language research: methodology and design.* New York: Routledge.
- MCNAMARA T  
2004. Language testing. Davies & Elder (eds) 2004: 763-83).
- MESSICK S  
1980. Test validity and the ethics of assessment. *American Psychologist* 35: 1012-27.  
1989. Validity. Linn (ed) 1989: 13-103.
- MILLER M D, LINN R L & N E GRONLUND  
2009. *Measurement and assessment in teaching.* Upper Saddle River, NJ: Pearson Education, Inc.
- SALKIND N J  
2011. *Statistics for people who think they hate statistics.* Los Angeles: Sage.
- SCHMITT N (ed)  
2002. *An introduction to applied linguistics.* London: Arnold.
- STOYNOFF S & C A CHAPELLE  
2005. *Esol tests and testing.* Alexandria, VA: TESOL.
- VAN ELS T, T BONGAERTS, G EXTRA, C VAN OS & A JANSSEN-VAN DITEN  
1984. *Applied linguistics and the learning and teaching of foreign languages.* London: Edward Arnold.
- WEIDEMAN A  
2009. Constitutive and regulative conditions for the assessment of academic literacy. *South African Linguistics and Applied Language Studies* 27: 1-26.  
2012. Validation and validity beyond Messick. *Per Linguam* 28(2): 1-14.
- WEIR C J  
1993. *Understanding and developing language tests.* New York: Prentice Hall.

## Appendix A

 Central University of Technology, Free State	<b>CENTRAL UNIVERSITY OF TECHNOLOGY, FREE STATE</b> <b>SENTRALE UNIVERSITEIT VIR TEGNOLOGIE, VRYSTAAT</b> <b>YUNIVESITHI E BOHARENG YA THEKENOLOJI, FOREISTATA</b>	
	<i>Academic Development and Support</i>	
<b>DATE: 14 – 18 May 2012</b>		<b>SESSION: In class time</b>
<b>SUBJECT: Academic Language Proficiency</b>		
<b>EXAMINER:</b>	<b>Mr K Sebolai</b>	
<b>MODERATOR:</b>	<b>Ms N Venter</b>	

MAY TEST: SEMESTER 1: 2012

**Initials and Surname:** \_\_\_\_\_

**Student Number:** \_\_\_\_\_

**Cell phone number:** \_\_\_\_\_

**Main Course at CUT e.g. Electrical Engineering:** \_\_\_\_\_

**INSTRUCTIONS:**

<b>Duration of paper:</b>	<b>2 hours</b>	<b>Maximum marks:</b>	<b>100</b>
---------------------------	----------------	-----------------------	------------

**ANSWER ALL THE QUESTIONS**

**THIS PAPER CONSISTS OF 12 PAGES. [edited for journal format – ed.]**

**SECTION A - READING**

**QUESTION 1**

**Read the passage below and answer the questions that follow. Circle either a, b, c, or d for your correct answer.**

**The Growth of Cities**

1. In 1800, only 3 percent of the world's population lived in cities or urban areas. In 2007, according to statistics from the United Nations, half of the world's population lived in urban areas. Recent urban growth has been in developing countries, and their cities are continuing to grow at astounding rates. The United Nations predicts that the urban population of developing countries will grow from 2.84 billion in 2000 to 4.9 billion in 2030. In the future, the United Nations predicts that almost all population growth will be in cities.

2. Cities have developed for many different reasons. The first cities grew up around marketplaces, where people traded food and goods. Because of this, major cities were established along large rivers or around harbours. Religion also played an important role in the development of urban areas. As religions became more organized, people built settlements around important religious buildings. Later, cities became the centres for government. They also provided security in a dangerous world. They were built on top of hills and often were surrounded by walls. Finally, and perhaps most importantly, cities attracted growing numbers of people with ideas about art and science. The cities then became the centres of culture.

3. By the nineteenth century, the largest cities were in Europe. Europeans planned cities around a central business district. The best shops and restaurants were often located in this sector. Wealthy people lived in an area that circled the central business district. Factories and low-quality housing were built away from this centre, so poorer people lived in the suburbs. This pattern continues today. In many European cities, the richest residents live close to the centre, while the poorest residents live in low-quality housing far from the centre.

4. In contrast, many North American cities have evolved differently. A well-known sociologist, Ernest Burgess, studied Chicago in the 1920s. He described the development of the city as a series of rings. The inner ring was the central business district. However, wealthy people did not live in the next ring, as they did in European cities. Instead, this ring had factories and poor housing. The third ring had better housing, but it was still for the working class. The fourth ring was the suburban area, where the wealthy lived in large houses with big yards. Not every North American city follows Burgess's model; however, it does explain why many cities in the United States have poor neighbourhoods close to the central business districts and wealthy suburbs far away from the central areas.

1. The main idea for paragraph one is

- a. By 2007, city population had grown by 47%.
- b. In the whole world, city population has grown tremendously.
- c. In developing countries, city population has grown rapidly.
- d. Everywhere, city population will continue to grow.

2. In 2007, city population was ..... of the whole population,
  - a. More than 3%
  - b. Almost 50%
  - b. Exactly 50%
  - d. More than 47
3. The United Nations thinks that the urban population will
  - a. probably grow in the future
  - b. possibly grow in the future
  - c. certainly increase in the future
  - d. logically increase in the future
4. The word “astounding” in paragraph 1 means
  - a. amazing
  - b. intriguing
  - c. tremendous
  - d. appalling
5. The word “recently” in paragraph 1 means
  - a. in the next few days
  - b. in the past few days
  - c. not a long time ago
  - d. Three days ago
6. The main idea for paragraph 2 is that
  - a. cities were a result of marketing
  - b. cities were a result of religion
  - c. cities were a result of government
  - c. cities were a result of several reasons
7. In the 19<sup>th</sup> century Europe, rich people lived near the CBD probably because
  - a. they had enough money to live there
  - b. they wanted to be apart from the poor
  - c. they wanted to live close to the shops
  - d. they owned the shops and restaurants
8. The word “evolved” in paragraph 3 means
  - a. developed naturally
  - b. developed rapidly
  - c. developed gradually
  - d. developed biologically

9. The word “however” in paragraph 3 is used to signal
- c. a predicament
  - a. a contrast
  - c. a postulation
  - d. a comparison
10. The word “instead” in paragraph 3 is used to indicate
- a. an explication
  - b. an elaboration
  - c. a contrast
  - a. a prediction

Mark = 20

---

## QUESTION 2

Read the passage below and provide the most meaningful word in the spaces provided. Write the word next to the corresponding numbers below the passage

### Testing in Education

Twelve-year-old Winston Lim of Singapore waits nervously for the results of an examination that will determine which secondary school he can attend. American Mark Saunders goes to class every Saturday to prepare for his university entrance examination. Like Winston and Mark, almost 1..... in school takes tests. They are a regular part of 2....., and they often have a significant impact on people’s 3..... Tests measure how much people know or what they can do. Examinations are 4..... at all levels of education, from primary school to university. In 5....., teachers write the tests and give them to their classes. Other kinds of 6..... are standardized. They are usually written by testing professionals. In some 7..... like Singapore, standardized testing begins very early. Every Singaporean takes a standardized test at the age of eleven. In many European countries, such as Italy, testing 8..... later, after two or three years of secondary school.

The score on these standardized tests can often 9..... a student’s educational future, especially outside of North America. Scores can also determine what subjects students can study. Students who want to study maths and science must have good math 10..... Students who prefer an arts curriculum must have good scores on reading and 11..... tests. Those with low scores may not get into an academic program at all. Instead, they may begin training for their future job immediately. Is this a good system? It is certainly efficient, yet some educators argue that it 12..... to the needs of the past. In the early twentieth century, growing economics needed many unskilled 13....., some skilled workers, and just a few very educated individuals. This system of 14..... and separating students was efficient then. Today, however, labour needs have 15....., and these economics need more skilled and educated workers. Therefore, these educators argue, a different kind of 16..... system may be needed.

Examinations in secondary school are just the beginning. The next important test comes when students 17..... to universities. Great Britain and many of its former colonies, including India, Tanzania, and Malaysia, use the A-level or a similar 18..... to

help determine who will get into universities. France, Switzerland, and some other countries in Europe use the international Baccalaureate examination. In the United States, most **19**..... take the Scholastic Achievement Test (SAT). Many parents and educators believe that these tests have become too important. They argue that **20**..... to universities should not depend on one examination. They say that it is important to include other factors, such as students' grades in school and their interests outside of school.

**Mark = 20**

---

**QUESTION 3**

The following sentences belong to two paragraphs of the same passage. The sentences are not in the right order. Rearrange the sentences so that the two paragraphs make meaning. Write the correct sentence numbers in the spaces provided.

**Supply and Demand in the Global Economy**

Paragraph 1

- \_\_\_\_\_ This shows that a natural disaster in one country can change the prices of a global product, such as oil, and can cause energy prices to increase around the world.
- \_\_\_\_\_ There is a saying that a butterfly flapping its wings in Japan can cause a hurricane in North America.
- \_\_\_\_\_ This happened in July 2007 when an earthquake closed several Japanese power plants.
- \_\_\_\_\_ As a result, oil prices around the world rose.
- \_\_\_\_\_ This saying illustrates what could happen in a global economy.
- \_\_\_\_\_ These plants produced energy.
- \_\_\_\_\_ An event in a business in one country can significantly affect businesses in other countries.
- \_\_\_\_\_ Because the Japanese could not use their own energy, they needed to buy more oil from other countries.



**Paragraph 2**

- \_\_\_\_\_ The law of demand states that as prices rise, demand falls.
  
- \_\_\_\_\_ In the example above, the Japanese demand for oil increased.
  
- \_\_\_\_\_ Every morning, over 166 million Americans wake up and drink their first cup for the day.
  
- \_\_\_\_\_ Coffee is an example of a global product that is in growing demand.
  
- \_\_\_\_\_ By the end of the day, the average American has drunk three cups of coffee.
  
- \_\_\_\_\_ It is important to note that a person will buy three cups a day at current prices
  
- \_\_\_\_\_ When people want more of a product, there is a growth in demand.
  
- \_\_\_\_\_ Therefore, the average demand for coffee at current prices in the United States is three cups per day per person.
  
- \_\_\_\_\_ Coffee is an example of a global product that is in growing demand
  
- \_\_\_\_\_ As they drive to work, they stop at a coffee shop and buy their second cup.
  
- \_\_\_\_\_ When someone wants or needs a product, it is called demand.
  
- \_\_\_\_\_ If the price of a cup of coffee increases, the person may decide to drink only two cups a day.

**Mark = 20**

**QUESTION 4**

**Read the passage below and fit the missing sentences into the spaces provided. The sentences are provided at the end of the passage. Write the letters A, B, C etc in the spaces, depending on the sentence you choose for each answer.**

**Education Around the World**

When you think of school, you may think of a classroom like the one you are sitting in now, but not all classrooms look like this. **1.....** Education comes in many different forms and has a long history. In early times, schools were available only to the elite, but this changed with the beginning of industrialization. **2.....** In order to meet this need, the number of schools expanded, and education became accessible to more children. Today, most nations want all of their children to go school. **3.....**

Most countries divide education into three levels: primary, secondary, and higher, or university education. **4.....** It is usually free and it is also generally compulsory. Primary instruction usually includes reading, writing, mathematics, and the nation's history. **5.....** In some parts of the world, children go to school even before they are five. For example, in Japan and the Czech Republic, the majority of children between the ages of three and five go to preschool. **6.....** This is because working parents cannot stay at home to take care of their children, and because parents believe that preschool can give their children an advantage.

**7.....** It begins when children are about twelve years old. In some countries, such as Germany and Hungary, secondary school children are put into groups based on their scores on a national test. **8.....** Children with lower test scores go to schools that teach more practical skills, such as car repair or cooking. **9.....** Sometimes secondary school students may go to different schools because of their grades in specific subjects. For example, some students are better at history and literature, and others are better at maths and science. **10 .....**

**C.** Recently, some nations have stopped separating their students in this way and now educate all children together.

**E.** Primary school begins when children are about five years old and lasts for six to nine years.

**F.** Secondary school lasts for three to six years.

**H.** One school may have dirty floors and no chalkboard; another may be in a modern building with computers.

**A.** Children with high test scores go to secondary schools that emphasize academic subjects, such as maths, science, languages, and literature.

**B.** It may also include religious and moral instruction.

**F.** New industries needed more educated workers.

**G.** These schools are becoming more popular, especially in Europe, North America, and Japan.

I. This is because educated citizens can contribute to their nation's development.

J. Or they may attend schools that teach technical skills, like how to use computers or other equipment.

**Mark = 20**

---

**SECTION B – WRITING**

**QUESTION 5**

**Write two paragraphs (introduction and second paragraph) on ONE of the topics below. Ensure that you adhere to the principles of COHESION AND COHERENCE in writing.**

1. What do you think is the main cause of crime in South Africa?
2. What step should be taken to reduce crime in South Africa?
3. What step should be taken to improve the hospitality industry in South Africa?
4. What do you think is the key difference between the traditional and modern hotel industry in South Africa?

**Mark = 20**