Kabelo Sebolai
Central University of Technology

Do the Academic and Quantitative Literacy tests of the National Benchmark Tests have discriminant validity?

ABSTRACT

The massification of higher education since the last decade of the 20th century has resulted in an unprecedented influx of students to universities throughout the world. In South Africa, the advent of democracy in 1994 has added impetus to this phenomenon. As a result of the poor quality of education offered at the schools they attended, however, most of the students entering universities in the country are underprepared to handle the demands of higher education in English, the language of teaching and learning at these institutions and a second language to most students. This is the case even for students who obtain good results in their high school exit examinations. Higher Education South Africa (HESA) has introduced a set of tests known as the National Benchmark Tests (NBTs) to assess the level of academic preparedness of the students entering universities for the first time. The NBTs comprise three tests of Academic Literacy (AL), Quantitative Literacy (QL) and Maths Literacy (ML). As their names imply, the three tests are aimed at measuring three different knowledge/skills domains that are key to student success at university. It is important therefore that performance in these tests exhibits evidence of the difference that these domains entail. The aim of this study was to establish if two of these tests, namely, the AL and QL tests possessed discriminant validity. The results revealed that the tests lack discriminant validity.

Keywords: discriminant validity, academic literacy, quantitative literacy, test.

Journal for Language Teaching, Volume 48, Number 1, pp. 131 - 147. 2014.
ISSN 0259-9570. http://dx.doi.org/10.4314/jlt.v48i1.7

1. Introduction

In 2011, Higher Education South Africa (HESA), an association of 23 public universities in South Africa, introduced a battery of tests known as the National Benchmark Tests (NBTs). This was an outcome of the low levels of academic preparedness reported among the majority of the students graduating from high school and entering South African universities in the past 20 to 30 years (Van Dyk 2005). This is a challenge faced by the universities even regarding the students entering such universities with good scores on their Grade 12 exams. Indeed, Van Wyk and Yeld (2013) have pointed out that gaining access to university means that students have to acquire academic literacy which is in Bourdieu and Passeron's (1990: 66) view, nobody's native language. This means that they need to learn new ways of "saying (writing) – doing – being – valuing – believing combinations" (Gee 1996: 127). The aim of the NBTs is to provide diagnostic information regarding the level of academic preparedness among high school leavers and assisting in their proper placement at institutions of higher learning. The rationale for and context of the introduction of the NBTs have been described as follows:

> The NBTs are designed to provide criterion-referenced information to supplement school leaving results such as the National Senior Certificate (NSC). The NSC is of necessity norm-referenced, which means that its results yield sometimes difficult to interpret information about candidates' actual level of achievement. It is therefore challenging for institutions to use the NSC on its own to prepare in advance to meet the educational needs of incoming students as effectively as possible. (National Benchmark Tests Project (NBTP) 2013: 7)

As their name implies, and for the placement purpose for which they were introduced, scores on the NBTs are used to categorise test takers into three groups; basic, intermediate and proficient. The test takers classified as being at the 'basic' level of achievement are identified as those having serious learning problems and who are unlikely to succeed at university if they do not go through a bridging programme or go to a Further Education and Training (FET) college first before they start pursuing university education (NBTP 2013). Those categorized as being in the 'intermediate' rung would also need to receive academic foundational support which is shorter than that needed by those in the basic category before they start their first year university study (NBTP 2013). Lastly, those belonging in the 'proficient' category are those that can be admitted straight into regular university programmes without having to undergo any academic preparation prior to commencing their university studies (NBTP 2013). All the documents written on the NBTs do not associate them with access to university. These tests are, however, administered in the year preceding a test taker's admission to university. Furthermore, the defining characteristic of the 'proficient' category of the NBTs is 'admission'. This has created an ambivalence regarding the actual aim and purpose of the NBTs. In other words, while the NBTs have been described as placement tests, this ambivalence has opened them up for use for both placement and access. A total of 49 institutions, organizations and bodies currently use the NBTs and 38 of all these participants in the tests use them for admission and placement (NBTP 2013: 12).

As pointed out in the abstract, the NBTs consist of three different tests aimed at measuring the Academic Literacy (AL), Quantitative Literacy (QL) and Maths Literacy (ML) levels of these students to determine if they will cope with the demands of higher education related to these domains. Given the differences in how their domains and constructs are defined, it is necessary that the three tests possess discriminant validity. If two or more tests possess discriminant validity, the same group of test takers should perform differently on both, as a function of the difference in how the constructs underlying the tests are defined. The aim of this study was to establish the extent to which the AL and QL tests of the NBTs possessed discriminant validity. The study was initiated out of concern about the scant research on various aspects of the NBTs. This is in contrast with the numerous studies that have been published on the Test of Academic Literacy Levels, another local standardized test of academic literacy. In the absence of data from another standardized test of academic literacy for the participants in this study, as well as the inaccessibility of item level data from these two tests to outside researchers, the researcher could not use other possible procedures to investigate the construct validity of the two tests. Validating the construct of these tests through discrimination seemed the only option at the researcher's disposal.

2. Validity

The discriminant type of validity is rarely mentioned in scholarly discussions on the notion of validity. It is appropriate therefore that the literature on the term validity is comprehensively explored so that the relationship between discriminant and other types of validity is first clarified. So far, the term validity has mainly been defined in two ways. Firstly, the meaning of validity has traditionally been understood to relate to the extent to which a test measures what it was designed to measure. Put differently, validity has been interpreted as "an inherent attribute or characteristic of a test, that a psychologically real construct or attribute exists in the minds of the test taker – this implies that if something does not exist, it cannot be measured" (Van der Walt & Steyn 2007: 139). Secondly, validity has been interpreted to mean that the results obtained on a test mean what they are interpreted to mean. In the words of Messick (1989: 13),

> 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.

Thus, the first definition associates validity with a test itself while in the second one, it is viewed as a property of the results and not necessarily the test generating such results. The two interpretations, however, seem to be two different ways of expressing the same idea. A test that does not measure what it is designed to measure cannot produce results whose interpretation is valid. In other words, test scores cannot be interpreted validly if the instrument yielding such scores is not measuring what it purports to measure (Weideman 2009; 2012; Borsboom, Mellenbergh & Van Heerden 2004).

Traditionally, validity has been classified into three types; content, construct and criterion-related types. Firstly, the content type of validity refers to the adequacy of the content sampled for the purpose of measuring a particular domain of knowledge, skills or trait. In the words of Miller, Linn and Gronlund (2009: 75),

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

Two main procedures are used to assemble evidence for the validity of a test's content. The first of these involves a compilation of a table that specifies the content from which a test was developed. A table of specifications constitutes a framework detailing the information covered by the test items, the number of items that tap each content area covered and the way the items are organized in a test (Cohen & Swerdlik 2010). In educational measurement, the sources of this information include syllabi, course textbooks, teachers of a course and curriculum developers (Cohen & Swerdlik 2010). A table of specification is therefore very useful in helping test developers construct a test whose results will represent the content and objectives they wish to measure. It serves as a source of guidance to the test developer with regard to the relative degree of emphasis that each content area will be allocated in the test (Miller, Linn & Gronlund 2009). It is for this reason that Van Els, Bongaerts, Extra, Van Os and Janssen-van Dieten (1984: 318) have argued that content validation is typically "incorporated into the process of test construction itself, and therefore takes place before the test is used". The second procedure for investigating test content validity involves help by experts of the content at issue. Such experts are asked to assess the degree to which test items are relevant to the content covered by a test (Gregory 2007; Cohen & Swerdlik 2010). Typically, this involves the use of a scale by the experts to rate each item by indicating whether it is essential, useful and necessary (Cohen and Swerdlik 2010). This culminates in the computation of a content validity ratio (Cohen and Swerdlik 2010), a statistical index of the degree to which the experts agree on the content validity of an item. An item satisfies the content validity requirement if more than half of the raters agree that it is essential to a test (Cohen & Swerdlik 2010).

Construct validity refers to the extent to which evidence can be provided to prove that abstract knowledge, a skill or trait measured by a test exists. In other words, construct validity refers 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" (Stoynoff & Chapelle 2005: 17). A test's construct validity can be determined in a number of ways. The first of these involves determining the extent to which the test measures a single construct in that it consists of tasks that are homogeneous in the sense that they all function to elicit different dimensions of the same ability. Gregory (2007: 132) has pointed out that "if a test measures a single construct, then its component items (or subtests) likely will be homogeneous (also referred to as

internal consistency)." A common procedure for investigating test homogeneity involves computing statistical correlations between individual item scores and the total score. If all these correlations are high and the general trend is that high performers in the test as a whole get most of the items right as compared to those who perform poorly, it is very likely that all the items measure various aspects of the same construct and that they all contribute to the test's homogeneity (Cohen & Swerdlik 2010). Weideman (2009: 5) has argued, however, that for "for an ability as richly varied and potentially complex as academic language ability, one would expect and therefore have to tolerate, a more heterogeneous construct".

The second piece of evidence for construct validity relates to the relationship between age and test scores. Variance in performance on tests developed to measure most constructs has also been found to be a function of age. Research involving tests of vocabulary knowledge has revealed, for example, that performance in these tests change with age; the older one gets, the more one gains in vocabulary (Gregory 2007). In the words of Cohen and Swerdlik (2010: 195), "if a test score purports to be a measure of a construct that could be expected to change over time, then the test scores too should show the same progressive changes with age to be considered as a valid measure of the construct". Another piece of evidence for construct validity is one's ability to demonstrate that scores in pre- and post-testing are incrementally different as a result of some form of intervention or experience. Put differently, a test used for pre- and post-testing should be sensitive to intervening experiences or treatment in order for a construct validity argument to be maintained for it (Miller et al. 2009). Collecting evidence for construct validity in this manner involves setting up an experimental study in which the participants comprise several groups of individuals who are believed to have a degree of the ability being tested and whose possession of this ability is hypothesized to increase as a result of exposure to some treatment related to the construct involved (Bachman 2004). In the words of Cohen and Swerdlik (2010: 196), "such changes in scores in the predicted direction after the treatment program contribute to evidence for construct validity …" A vocabulary test that yields scores that are incrementally consistent with vocabulary instruction, for example, allows the test developer to use the intervention as a benchmark for ascertaining the validity of the claim that the test measures a defined construct of vocabulary knowledge. Conversely, if the test does not show an increase in performance after the intervention, it falls short of serving as evidence that it measures the construct.

The fourth procedure for collecting evidence for construct validity is what Cohen and Swerdlik (2010: 196) refer to as the "method of constructed groups" and which Gregory (2007: 133) calls "theory-consistent group differences". In this case, the test developer is required to demonstrate that two or more groups of test takers who are different regarding a specific characteristic perform differently in a test as a result of this difference. This method involves administering "the same test to several groups of individuals who are known, or who are believed, on the basis of some prior criterion, to differ in the ability to be assessed" (Bachman 2004: 290). For example, a test developer might want to demonstrate that a reading test has construct validity by showing that test takers who have a strong reading background perform better than those who do not. In

the words of Cohen and Swerdlik (2010: 196), "the rationale here is that if a test is a valid measure of a particular construct, then the test scores from groups of people who would be presumed to be different with respect to that construct should have correspondingly different scores". In other words, the difference in performance between two groups that differ in some specific way becomes the benchmark for validating the construct of a test. Statistically, this difference can be determined by computing a one-way Analysis of Variance (ANOVA) of the scores.

Finally, construct validity can be established through a statistical procedure called factor analysis. Factor analysis is used "to reduce a large number of variables (e.g. test or questionnaire items) to a smaller number (thought to represent the underlying abilities the test developer is seeking to measure) of variables" (Stoynoff & Chapelle 2005: 21). Factor analysis is a procedure premised on the understanding that a construct consists of a number of traits, some of which intercorrelate and can therefore be reduced into a single factor or dimension. It involves administering a battery of tests to the same group of test takers and investigating the degree of correlations between the scores yielded by such tests (Bachman 2004). In the words of Cohen and Swerdlik (2010: 332) "Factor analysis helps us discover the smaller number of psychological dimensions (or factors) that can account for various behaviours, symptoms, and test scores we observe". In other words, factor analysis is a statistical tool that test analysts can use to summarize the properties of a complex construct into fewer manageable ones. A discovery of such factors through the use of the factor analysis procedure suggests that the construct, which a test user is interested in, exists and that it can be measured (Cohen & Swerdlik 2010). The ultimate role of factor analysis is, however, to help the test developer determine the extent to which the construct underpinning a test is unidimensional. In the words of Weideman (2009: 4), "tests are assumed to measure a single, homogeneous ability. If they do not, but instead measure more than one (i.e. heterogeneous) ability, this shows up particularly well in one technical measure of consistency, a factor analysis."

Criterion-related validity relates to the degree to which scores on one test correlate with those on another test of a related construct for the same group of test takers. Criterion-related type of validity has further been classified into two types. These are the concurrent and predictive types. On the one hand, in a concurrent criterion-related validity study, scores obtained by one group of test takers from two different tests that are based on related constructs and that are administered around the same time are compared to determine the degree of their correlation. On the other hand, predictive validity involves a study of the degree to which test scores from a test can predict performance in some measurement variable in the future. The procedure for computing the two types of criterion-related validity involves running a correlation analysis of performance on the tests or criteria being validated. A high correlation of the scores signals high criterion-related validity which can be interpreted to mean a high degree of similarity or overlapping between the constructs underpinning the measurement variables being studied. In this sense, the ultimate outcome of criterion-related validation is to establish the construct validity of the tests involved.

A type of validity that has attracted interest in both educational and psychological testing in recent years relates to the consequences of testing. Measurement specialists now believe that in investigating the validity of a test, it should be determined whether the decisions taken on the basis of the test results are harmful or beneficial to the test takers. In other words, the belief is that testing should be carried out to promote the interests of those involved. Otherwise it has negative consequences that impact its consequential validity. Indeed, Messick (1989: 1012) has pointed out 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". This is particularly relevant to language testing because "language is rooted in social life and nowhere is this more apparent than in the ways in which knowledge of language is assessed" (McNamara & Roever 2006: xiv). Bachman and Palmer (1996: 30) have added to 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 have consequences". Typically, testing has consequences for test takers, teachers and educational systems (Bachman & Palmer 1996). The procedure for determining if a test possesses consequential validity is to investigate the degree to which the decisions taken on the basis of its scores promote the well being of those affected. The more harmful the consequences of these decisions, the less valid a test could be considered.

Discriminant validity is the opposite of the concurrent criterion-related type of validity. As pointed out earlier, the procedure for establishing the degree of a test's criterion-related validity is to correlate its scores with another measure or criterion of the same or related construct. A high correlation between such a test and the criterion attests to the test's criterion-related validity, also known as convergent validity. In the words of Gregory (2007: 134), "convergent validity is demonstrated when a test correlates highly with other variables or tests with which it shares an overlap of constructs". Cohen and Swerdlik (2010: 197) add that "convergent evidence for validity may come not only from correlations with tests purporting to measure an identical construct but also from correlations with measures purporting to measure related constructs." The ultimate aim of criterion-related validity is to determine a test's construct validity. In other words, the higher the criterion-related validity of two tests, the more justifiable their construct validity becomes. There are times, however, when testers want to establish a test's construct validity by contrasting it with that of another test that purports to measure a different construct. This is known as establishing a test's discriminant validity. Investigating discriminant validity involves generating statistical evidence to show that a test does not correlate significantly with a measure aimed at measuring a different construct. Cohen and Swerdlik's (2010: 197) definition of discriminant validity captures this notion very well:

> A validity coefficient showing little (that is, statistically insignificant) relationship between test scores and/or other variables with which scores on the test being construct-validated should not theoretically be correlated provides **discriminant evidence** of construct validity (also known as *discriminant validity*).

Thus, the lower the correlation between scores on two tests the lower their criterion-related validity and the higher their discriminant validity.

3. Description of the Sample

The sample used for this study consisted of 108 males and 92 female first year students enrolled in various programmes offered at the Central University of Technology (CUT) in Bloemfontein, South Africa. The sample consisted of 200 participants in total. The students had successfully completed their Grade 12 exam the previous year and had subsequently gained admission to the university. Their age ranged from 18 to 21 years. The majority of them were from African languages background while the minority spoke Afrikaans and English as home languages.

4. Methodology

The AL and QL tests of the NBTs were administered for the first time at CUT in March 2012. The AL and QL components of the NBTs constitute one test administered over three hours, but the results for the two domains are reported separately as percentages and benchmarks (NBTP 2013). These tests are designed on the basis of two different constructs and are therefore meant to provide two types of information. The AL test is aimed at measuring students' "capacity to engage successfully with the reading and reasoning demands of academic study in the medium of instruction" while the QL component targets students' "ability to manage situations or solve problems in a real context that is relevant to higher education study, using basic quantitative information that may be presented verbally, graphically, in tabular or symbolic form as related to both the NSC subjects of Mathematics and Mathematics Literacy" (NBTP 2013: 8). On the one hand, the construct underpinning the AL test has been described by Cliff and Yeld (2006: 20) as being constituted by 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 text;
- separate essential from non-essential and super-ordinate from sub-ordinate information;
- understand and interpret visually encoded information, such as graphs, diagrams and flow-charts;

- understand and manipulate numerical information;
- understand the importance and authority of own voice;
- understand and encode the metaphorical, non-literal and idiomatic bases of language; and
- negotiate and analyse text genre.

The descriptive statistics of the scores for the participants in the present study on the AL test of the NBTs administered at CUT in March 2012 are captured in **Table 1** below.

Table 1: Mean and standard deviation of the scores on the AL test of the NBTs administered at CUT in 2012 (N=200)

Variable	M	SD	Max	Min
AL test	45.4	11.6	83	26

On the other hand, the construct on the basis of which the QL test was developed is outlined in the NBT Degree Standard Setting manual (2012: 18) as the test taker's ability to do the following:

- select and use a range of quantitative terms and phrases
- apply quantitave procedures in various situations
- formulate and apply formulae
- interpret tables, graphs, charts and text and integrate information from different sources
- do calculations involving multiple steps accurately
- identify trends and patterns in various situations
- apply properties of simple geometric shapes to determine measurement
- reason logically and
- interpret quantitative information presented verbally, symbolically, and graphically.

The descriptive statistics of the scores yielded by the QL test of the NBTs for the participants in the present study are captured in **Table 2** below.

Table 2: Mean and standard deviation of the scores on the QL test of the NBTs administered at CUT in 2012 (N=200)

Variable	M	SD	Max	Min
QL test	41.3	13.6	87	22

5. Results and discussion

In order to investigate the discriminant validity of the AL and QL tests of the NBTs, the Statistical Package for Social Sciences (SPSS) was used to run a correlation analysis of the participants' scores on the two tests. Correlation is a statistical procedure used to "look at two variables and evaluate the strength and direction of their relationship or association with each other" (Dörnyei, 2007: 223). The correlation coefficient is therefore an appropriate procedure for computing a test's criterion-related and discriminant types of validity. The correlation coefficient ranges from -1 to +1. A correlation coefficient of +1 means that there is 100% positive association between the variables involved while that of -1 signals a complete negative relationship between such variables (Miller et al. 2009). Mackey and 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.

A high correlation coefficient between the two tests involved in this study would therefore mean that they have high criterion-related validity and that they probably measured almost the same or highly related constructs. In contrast, a low correlation between the two would mean that they measured two largely unrelated constructs and that they possessed discriminant validity. The results of the correlational analysis of the scores

obtained by the participants on the AL and QL tests of the NBTs are summarized in **Table 3** below.

Table 3: The correlation between the scores on the AL and QL tests of the NBTs administered at CUT in March 2012 (N = 200)

Variables	Correlation Coefficient	p-value
AL test QL test	.632	0.01

As shown in **Table 3** above, the correlation coefficient of the two tests was .63 and the *p*-value was 0.01. Firstly, this means that the correlation between the two tests was high and that they therefore possessed concurrent criterion-related validity and not discriminant validity. Dörnyei (2007: 223) has pointed out 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." Secondly, the *p*-value of .01 means that the results of the study were statistically highly significant. In the words of Mackey and Gass (2005: 265) "the accepted *p*-value for research in second language studies (and in other social sciences) is .05.

A *p*-value of .05 indicates that there is only a 5% probability that the research findings are due to chance, rather than the actual relationship between or among variables." Thus, the *p*-value for the results in the present study means that the probability that the findings were a result of pure chance was 1%. Overall, this means that the correlation between the scores on the two tests by the same group of test takers was too high for any claim to be made that any of them possessed discriminant validity in relation to the other. In other words, with regard to participants in the present study, the two tests possessed convergent instead of discriminant validity. This lack of or low degree of discriminant validity by the two tests is graphically evident in their high correlation depicted in **Figure1** below.

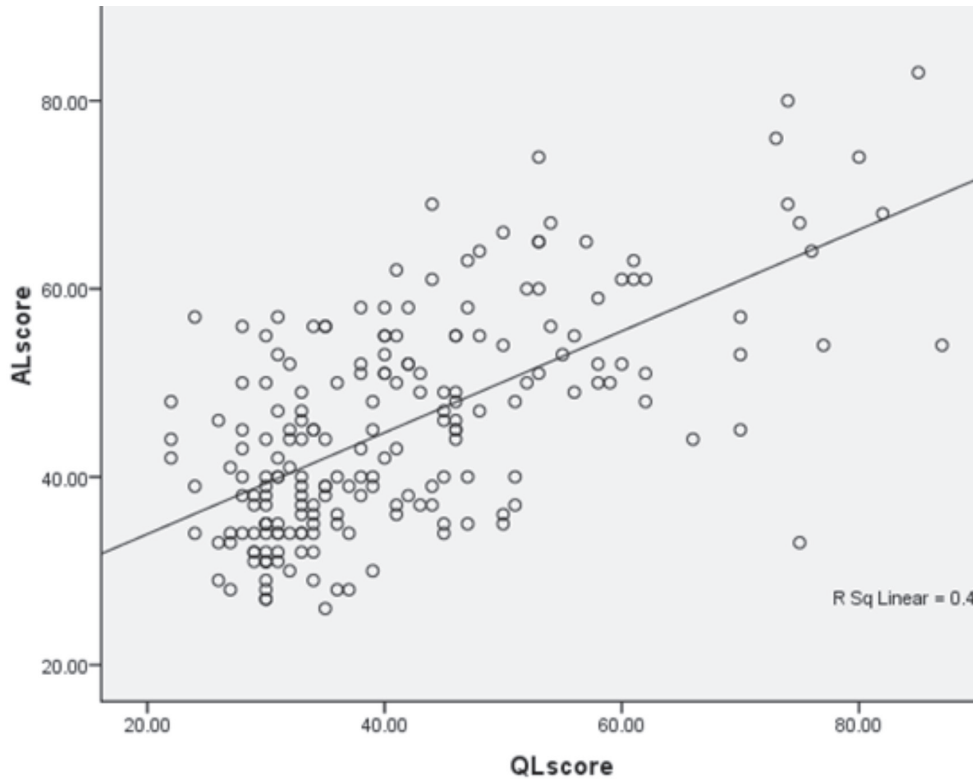

Figure 1: A graphic representation of the high degree of correlation between the scores of the AL and QL tests of the NBTs at CUT in March 2012 (N = 200).

As shown in **Figure 1** above, the regression line shows that not only was the correlation between the two tests high and statistically significant in the case of the participants in the study, it also shows that the association between the scores on the two tests was positive. This means that students who performed well in the AL test also tended to do well in the QL test and that those who performed poorly on one of the tests tended to do the same on the other. This bears testimony to the high criterion-related validity of the two tests and their low discriminant validity in relation to each other.

The results of this study show therefore that there is an overlap in some aspects of these tests which should account for their lack of discriminant validity. The first possible reason for this is that "general intelligence" is believed to relate with both Maths and language ability and could account for the high correlation in the performance on the two tests. If this is indeed what happened, it would exonerate the two tests from any possible discriminant validity related defects in their constructs. A look at the construct

of the AL test shows, however, that two skills that the test purports to measure, namely, 'understand visually encoded information, such as graphs, diagrams and flow charts' and 'understand and manipulate numerical information' should be part of the construct of the QL test instead. It is a probable result of overlaps of this kind that the developers of TALL, another test of academic literacy developed in South Africa, do not seem to distinguish between what the National Benchmark Tests Project (NBTP) categorizes as academic and quantitative literacy. This is evident in the construct underpinning the test, its content, and the fact that performance on the kind of items that measure what the NBTP would categorize as 'quantitative literacy' and 'academic literacy' is not reported separately as it is done with the NBTs. Patterson and Weideman (2013a&b) continue to disregard this distinction in their argument for the typicality of academic discourse as the basis for defining constructs of academic literacy.

To Patterson and Weideman (2013a&b), the defining feature of academic discourse is that it requires analytical and logical thinking to be processed successful. Without any doubt, these are the essential qualities of both the constructs of academic and quantitative literacy underpinning the NBTs that were presented earlier, and a possible reason for the high degree of convergence in test taker performance in the two tests. Furthermore, both these tests are presented in the multiple choice format. It would be unreasonable, however, to attribute the 60% correlation in performance on the tests to the use of a similar response format alone. Also, by its very nature, a test like the QL of the NBTs is susceptible to interference by construct irrelevant factors such as reading ability particularly when its takers are second language speakers of the language of the test (Miller *et al*. 2009).

It is important therefore that the language related demands of such a test are reviewed to minimize their role in possible construct irrelevant variance in performance (Miller *et al.* 2009). The role of language in the overlap between the tests is likely to be a factor because the same language is used in both tests. Lastly, test content can also contribute to the overlap in performance on the two tests. Whether any or all of these factors can account for the high convergence of the scores on these tests can be understood better, however, if these tests are investigated at the item level.

6. Conclusion

The aim of this paper was to investigate the extent to which the AL test of the NBTs possessed discriminant validity in relation to its QL counterpart. The two tests are developed on the basis of two different constructs and are therefore meant to give two different types of information about the test-takers. The degree of the discriminant validity of the two tests could be established by investigating the extent to which scores obtained by a group of test takers on one of the tests correlated with those they obtained on the other. A statistically significant correlation between the two would mean that they possessed convergent validity or criterion-related validity, the opposite of discriminant validity. This would mean that the two tests measured almost the same construct.

The results of the present study revealed that the two tests had a statistically significant correlation and that they therefore possessed convergent as opposed to discriminant validity. This is problematic in a context where the purpose of different tests is to advise students about choices related to university degree programmes; or in cases where the test results are sometimes used to direct students into extended or ordinary academic programmes. While the degree of the correlation is below 100%, the correlation is high enough to imply that the two tests are based on highly overlapping constructs. This defeats the purpose of using these tests for making two different types of inferences regarding the academic readiness of the students who take them.

Whatever the source of the high degree of convergence between the two tests is, it raises questions about the rationale for using two tests that are supposed to generate different information whereas they fail to stand a discriminant validity test. In the broader context of ethics and testing, test-takers could rightfully ask why it is necessary for them to take "the same" test "twice". The questions raised in this article should be considered by test-makers, test-takers and test-users in the interest of the broader elements related to the validity of the tests under discussion.

The causes for the lack of discrimination value could not be investigated fully in this article, because the researcher did not have access to the item level data. Currently, it is not easy to get access to this data. Future studies of this nature should determine the availability of the item level data first of all so that more comprehensive conclusions could be made when the study is replicated in different contexts in order to determine the generalizability of its findings. This remains a limitation of the current project and it would hopefully be pursued by researchers in future.

References

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., Mellenbergh, G. J. & Van Heerden, J. 2004. The concept of validity. *Psychological Review* 111(4): 1061 – 1071.

Bourdieu, P. & Passeron J. C. 1990. *Reproduction in education, society and culture.* Newbury Park: Sage Publications.

Cliff, A. F. & Yeld, N. 2006. Test domains and constructs: Academic literacy. In: Griesel, H. (Eds.) 2006. *Access and entry level benchmarks: The national benchmark tests project.* Pretoria: Higher Education South Africa. pp. 19-27.

Cohen, R. J. & Swerdlik, M. E. 2010. *Psychological testing and assessment.* New York: McGraw-Hill.

Dörnyei, Z. 2007. *Research methods in applied linguistics*. Oxford: Oxford University Press.

Gee, J. P. 1996. *Social linguistics and literacies: Ideology in discourse* (2nd ed.) London: Taylor and Francis.

Gregory, R. J. 2007. *Psychological testing: History, principles and applications*. New York: Pearson.

Mackey, A. & Gass, S. M. 2005. *Second language research: Methodology and design.* New York: Routledge.

McNamara, T. & Roever, C. 2006. *Language testing: The social dimension.* Language Learning Research Club, University of Michigan: Blackwell Publishing.

Messick, S. 1989. Validity. In: Linn, R. L. (Eds.) 1989. *Educational measurement.* Third edition. New York: American Council of Education/Collier Macmillan. pp. 13-103.

Miller, M. D., Linn, R. L. & Gronlund, N. E. 2009. *Measurement and assessment in teaching.* New Jersey: Pearson Education, Inc.

NBTP. 2012. Academic literacy, quantitative literacy and mathematics degree standard-setting meeting. Unpublished: University of Cape Town.

NBTP. 2013. National Benchmark Tests Project Results – National Report (1). Unpublished: University of Cape Town.

Patterson, R. & Weideman, A. 2013a. The typicality of academic discourse and its relevance for constructs of academic literacy. *Journal for Language Teaching* 47(1): 107-123.

Patterson, R. & Weideman, A. 2013b. The refinement of a construct for tests of academic literacy. *Journal for Language Teaching* 47(1): 124-151.

Stoynoff, S. & Chapelle, C. A. 2005. *Esol tests and testing.* Alexandria, Virginia: TESOL

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

Van, Dyk, T.J. 2005. Towards providing effective academic literacy intervention. *Per linguam* 21(2): 38 – 51.

Van Els, T., Bongaerts, T., Extra, G., Van Os, C. & Janssen-van Diten, A. 1984. *Applied linguistics and the learning and teaching of foreign languages.* London: Edward Arnold.

Van Wyk, A. & Yeld, N. 2013. Academic literacy and language development. In: Kandiko, C. B. & Weyers, M. (Eds.) 2013. *The global student experience: An international comparative study.* New York: Routledge. pp.62-77.

Weideman, A. 2009. Constitutive and regulative conditions for the assessment of academic literacy. *Southern African linguistics and applied language studies* 27(3): 235-251.

Weideman, A. 2012. Validation and validity beyond Messick. *Per Linguam* 28(2): 1-14.

ABOUT THE AUTHOR

Kabelo Sebolai

Central University of Technology, Free State (CUT), Private
Bag X20539, Bloemfontein, 9300, South Africa

Email address: ksebolai@cut.ac.za

Kabelo Sebolai is the coordinator of the Academic Literacy Programme at the Central University of Technology in Bloemfontein, South Africa. His research interests include academic literacy curriculum development and ESL testing.