Kabelo Sebolai
Language Centre: Stellenbosch University

Revisiting the meaning of validity for language testing: The case of two tests of English language ability

Abstract

Validity is probably the most crucial of all concepts that govern all kinds of measurement. This is more so the case in educational and psychological testing where high stakes decisions often need to be taken about individuals and institutions. From the time it saw the light of day, however, the concept of validity has been a source of inconclusive contestation. Two schools of thought have arisen from this debate. In the main, the first of these, also known as the traditional view, regards validity as a property of a test while the second, also known as the unitary view, locates it in the way test scores are interpreted and used. This creates a challenge for test developers with regard to exactly what the object of test validation should be. The aim of this article was to determine which of these views is defensible particularly for language testing. Using two studies focusing on the validity of two tests of language ability as the basis, the article demonstrates that the unitary view of validity is problematic for these tests as it leaves them susceptible to the possibility of being used for what they are not designed for.

Keywords: validity, language ability, language testing, meaning, traditional view, unitary view

***Journal for Language Teaching*, Volume 52, Number 1, pp. 152 – 168. 2018.**
ISSN 0259-9570. https://dx.doi.org/10.4314/jlt.v52i1.8

1. Introduction

Since about a decade before the advent of its new government in 1994, South Africa witnessed unprecedented growth in the body of students gaining admission to its universities. Positive as it is, this development has unfortunately been accompanied by concerns that the majority of these students have proven to be under-prepared for higher education. This challenge is often attributed to the country's political history of apartheid and the concomitant socio-economic inequalities that created an imbalance with regard to academic preparation for university education among different races. Observations have also been made at the same time, however, that academic under-preparedness is a phenomenon that cuts across high school leavers regardless of political and socio-economic background. These opposing views notwithstanding, the fact remains that South African universities all agree that they are faced with a first year student body that is not adequately prepared for higher education and for whom extra support has had to be provided to see the students through the academic challenges that are typical of university education. The need to provide this support has in turn, necessitated testing for academic readiness so that levels of academic preparedness are determined prior to the start of academic instruction. This was the reason for the introduction and current widespread use of tests by these universities of academic readiness such as those developed by the National Benchmark Tests Project (NBTP) and those owned by the Inter-institutional Centre for Language Development and Assessment (ICELDA). The belief in the potential of these instructional interventions to help improve throughput rates is evident in the fact that a sizeable portion of the Teaching and Learning Grants provided by the Department of Higher Education and Training to universities is invested in such interventions. This places an obviously huge responsibility on those who develop tests of academic readiness to ensure that such tests possess a degree of validity, the ability to measure what they purport to measure.

For about a century and a half to date, the concept of validity has been viewed as the essence of quality assurance in all educational and psychological measurement. This stems from the fact that tests are administered for making decisions about those who take them and that what is measured by tests should logically be aligned with the particular decision aimed at. This alignment is critical because it unites testing with its purpose and gives ultimate meaning to the term validity. It is this understanding that has provided the basis for many to concede that of all qualities that accord a test its identity as a measurement tool, validity is the first. In language testing in particular, the prime status of validity is evident, for example, in observations such as the one Rambiritch (2012: 62) makes that "one would be forgiven for assuming that all questions find their answers in the concept of validity, for it is the concept of validity that seems to dominate the literature on language testing". This notwithstanding, from the time it came into being, a consensus meaning of this term has been elusive and continues to be contested to date. This will become evident in the history of the evolution of this concept which is briefly dealt with below.

2. The concept of test validity

Controversy over the meaning of validity begins with the decision by the North American National Association Directors of Educational Research to work towards achieving consensus on the terminology and procedures to be used in psychological and educational measurement (Newton & Shaw, 2014). From the literature (See Sireci, 2009; Newton & Shaw, 2014), it appears that this decision was a result of the proliferation and widespread use of various kinds of tests to take high stakes decisions about individuals and institutions from around the middle of the 19th century. In the words of Newton and Shaw (2014: 17) , "by the end of the 19th century, belief in the potential of structured assessment was high, and results from written examinations were used for all sorts of different purposes, from selecting individuals for jobs in the Civil Service to holding schools to account for the quality of their education." The subjectivity involved in some of these assessments would logically lead some to question their assumed accuracy and the need for their quality to be determined (Newton & Shaw, 2014). This need for quality assurance came to be expressed in the term validity, which came to be defined as the degree to which a test measures what it purports to measure (Newton & Shaw, 2014).

Towards the end of the 19th century, the procedure for establishing test validity came to centre around producing evidence for performance on a test and some specified external outcome. In the words of Sireci (2009: 21), the "earliest definitions of validity were largely pragmatic, defining validity in terms of correlation of test scores with some criterion". In other words, in the view of the proponents of this approach, "the correlation coefficient – or the 'validity coefficient' as it came to be known – provided definitive evidence of validity, even when the content of a test appeared to be quiet far removed from the attribute that was supposedly being measured" (Newton & Shaw, 2014: 18). This empirical approach of correlation to validity received further impetus from the subsequent emergence of factor analysis, which came to be particularly valuable in validating traits or constructs that were believed to manifest in test performance (Sireci, 2009).

The early 20th century came to witness the evolution of validity and validation beyond test criterion relationships and factor analysis (See Sireci, 2009; Newton & Shaw, 2014). As Sireci (2009: 23) observes, while psychometricians of this time were fascinated by contemporary empirical approaches to validity, "others were thoroughly dissatisfied with the notion that validity referred merely to 'what the test measured' and that tests could be validated solely through correlational and factor analytic studies." The issue of logically analysing the content of a test as a form of evidence for validity also started to attract attention (Sireci, 2009; Newton & Shaw, 2014). Consistent with this, arguments came to be made that educational achievement tests in particular needed, in addition to being subjected to test criterion validation, to exhibit evidence of validity in its content (Sireci, 2009; Newton & Shaw, 2014). The arguments went, in other words, that "it ought to be obvious, from logical analysis of test content alone, whether it measured what it was supposed to measure (Newton & Shaw, 2014:19)."

The next phase in the evolution of the meaning of the term validity begins in 1952 with the work done by a committee of the American Psychological Association (APA) that was tasked with the responsibility to develop standards for psychological testing (Sireci, 2009; Newton & Shaw, 2014). This committee produced a document that was published in the same year and which expanded on the work already done on the meaning of validity by categorizing it into four types namely, content, predictive, status and congruent (Sireci, 2009; Newton & Shaw, 2014). The second edition of this document was published two years later in which content and predictive validity were left untouched while status and congruent validity were changed to concurrent and construct validity respectively (Sireci, 2009; Newton & Shaw, 2014). It was not until the publication of the third edition of this document in 1966, however, that these four categories were reduced to only three types namely, content, construct and criterion-related types (Sireci, 2009; Newton & Shaw, 2014). This perspective is what is now known as the traditional view of validity.

The last phase of the evolution of validity takes its lead from the introduction of construct validity in the third publication of the standards document referred to earlier (Sireci, 2009; Newton & Shaw, 2014). A further elaboration of this concept by Cronbach and Meehl (1955), two members of the APA committee on standards referred to above, particularly paved the way for the emergence of this phase. Cronbach and Meehl (1955: 282) argued, among others, that

> Construct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not 'operationally defined.' The problem faced by the investigator is 'What constructs account for variance in test performance?'

While Cronbach and Meehl (1955) indicate at the beginning of their paper on this elaboration that they had no intention to elevate construct validity over its counterparts, Newton and Shaw (2014) argue, however, that this was not necessarily the case. To this end, the latter have pointed out that

> even as early as their [Cronbach and Meehl] landmark paper, there were indications that at the very least, construct validity might be considered first among equals. For example, they had stated explicitly that construct validation was important at times for every sort of test, including both aptitude tests and achievement tests. (Newton & Shaw, 2014: 21)

Cronbach and Meehl's (1955) views about construct validity gained more prominence in the years between 1974 and 1999, a period which Newton and Shaw (2014) have rightly labelled the 'Messick years' (Newton & Shaw, 2014). As this label suggests, Messick was key in taking Cronbach and Meehl's (1955) thinking on construct validity to a level that brought "the majority of measurement professionals of his generation around to the viewpoint that all validity ought to be understood as construct validity" (Newton & Shaw, 2014). What Messick (1980; 1989) essentially achieved was to unify other types of validity under construct validity, a development which has earned his thinking about validity the adjective 'unified' or 'unitary'.

Messick's conviction about construct validity being the essence of validity led him to argue that validity is a function of how test scores are interpreted and used and that it was not an attribute of the test involved. It seems logical that somebody whose conviction is that a construct is separate from the test used to measure it would locate validity in how test developers and users understood the meaning of test scores and not in the instrument used to generate such scores. Messick (1989: 13) defines validity exactly in the following words:

> An integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment.

A careful reading of the literature reveals that Messick's view of validity has been and remains the dominant definition of this concept especially in North America. Kane (2006: 17) has, for example, defined validation as "… the process of evaluating the plausibility of proposed interpretations and uses…" and validity "as the extent to which the evidence supports or refutes proposed interpretations and uses". The latest edition of the standards document which was referred to several times earlier in this article captures the same view of validity:

> Validity refers to the degree to which evidence and theory support the interpretation of test scores entailed by proposed use of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests … Validation logically begins with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use. (AERA, APA, NCME 1999: 9)

These definitions attest to the dominance of the view of validity as a property of the interpretation of test scores from the time that Messick (1980) explicitly proposed it for the first time to date.

But this has not been the end of the validity story. Some have, in recent years, reverted to the pre-Messick view wherein validity was viewed as a property of a test and not of the scores that such a test generates. The proponents of this view "define validity in terms of a causal relationship between the attribute being measured and performance on the test tasks" (Sireci, 2009). For example, Borsboom et al. (2004: 1), a group of scholars that are strongly for the definition of validity as a property of a test have argued that

> A test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure.

Similarly, except that their view of validity heavily inclines towards content representativity in a test, Lissitz and Samuelsen (2007: 441) appear to take the same position as

Borsboom et al. (2004) on validity, which is that a test must first be valid for it to produce valid scores:

> … the test is a combination of tasks and these tasks are the operational definition that is captured … by the name and description of the domain to be measured by the test.

The view of validity as the degree to which a test measures what it purports to measure has also been evident in the language testing literature in South Africa in recent years. Weideman (2009: 2012) has argued, for example, that the objective ability of a test to produce objective scores cannot be confused with one's ability to interpret those results appropriately. In his view, no amount of appropriateness or accuracy in how the results are interpreted can make the test yielding those results valid. In his own words (Weideman 2012: 4), removing validity as a characteristic of a test "runs the risk of downplaying the quality of the instrument. No amount of interpretation can improve the measurement results (score) obtained from an inadequate instrument that gives a faulty and untrustworthy reading". In fact, Weideman (2009: 241) views Messick's (1980; 1989) concept of validity and its location in test scores as an obscured way of attaching validity to the test itself:

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

The current lack of consensus on the meaning of validity means that there is little clarity on how test validation should proceed and that as a result, tests that are not valid might be used to take high stakes decisions about those involved. This is particularly a challenge in higher education in South Africa where tests of academic readiness are used to take access decisions. This is the research problem that necessitated this article.

The aim of the article is to contribute towards the necessary clarity on the meaning of validity with reference to the two opposite views of validity i.e. validity is a property of test scores and validity as a quality of a test, that were dealt with above especially with regard to tests of language ability. The specific position that the article takes is that the first of these views i.e. validity as a function of test score interpretation and use is itself a threat to the validity of language tests in particular and opens them up for misuse. In pursuance of this point, the article considers two validation studies of two tests of language ability and draws on the insights such studies provide for the meaning of language test validity. These tests are the Test of Academic Literacy Levels (TALL) and Proficiency Test in English Second Language Advanced Level (PTESLAL).

3. Methodology

This article is a case study of the meaning of validity for language tests. It analyses the results of two predictive validity studies on two tests with the view to determine which of the contesting views of validity is defensible and which is not for the two tests and those of language ability in general. This approach was adopted for the possibility it offered to yield insights about the essence of validity for tests of language ability. As, Babbie (2013: 338) points out, "case study researchers may seek only an idiographic understanding of the particular case under examination, or … case studies can form the basis for the development of more general nomothetic theories". The latter is the kind pursued in this article. This is the kind that Dornyei (2007: 152) calls an instrumental case study, one which is "intended to provide insight into a wider issue while the actual case is of secondary interest; it facilitates our understanding of something else." In the case of this article, the actual case, the validity of the two tests involved, is of secondary interest while the overall theory of validity for language tests is the primary focus.

4. The predictive validity study on TALL

The Test of Academic Literacy Levels was mooted against the background of reportedly low levels of academic literacy among high school leavers entering universities for the first time. The test has been and continues to be used by some South African universities to assess these levels for the purpose of student placement. The literature on what TALL measures and the theories of language ability that inform its construct are adequately covered elsewhere (See Weideman, 2003; Van Dyk & Weideman, 2004; Weideman, 2011) and will therefore not be dealt with in this article. It suffices to say that the test aims to assess the kind of language ability that is typically required to handle the demands of academic education in the language of teaching and learning.

Adequate literature (see Van der Slik & Weideman, 2005; Van der Slik, 2008; Van der Slik and Weideman, 2009; Van der Slik and Weideman, 2010; Weideman, 2009; Le, Du Plessis & Weideman, 2011; Van der Walt & Steyn, 2007; 2008; Van Dyk, 2015) is publicly available on most of the psychometric properties of this test and its accompanying social dimensions. As pointed out in the introduction to this article, however, it is studies on the predictive validity of this test that are particularly relevant to this article. Given the space restriction that the nature of this article imposes, only one of such studies will be considered here. This is the study carried out in 2014 by Sebolai (2016) on the predictive validity of TALL for a total number of 604 first year students at a South African university of technology. The students were enrolled in different programmes that are offered within the four faculties at this university. The outcome variable in the study was the students' end of year average performance. The linear regression methodology was used to determine and measure the predictive ability of the test through the use of the SAS statistical package. The results of this analysis are captured in **Table 1** below.

Table 1: The results of a linear regression of scores on TALL as a predictors of end of year average performance (n=604)

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

5. The predictive validity study of PTESLAL

PTESLAL is a test of English proficiency developed by the Human Sciences Research Council (HSRC) and used by the South African university of technology referred to earlier to determine the linguistic readiness of first time entrants to its academic programmes. A study similar to the one dealt with above was carried out with this test as a predictor of end of first year academic performance of 303 first years in 2012. In this case too, the participants were enrolled in different programmes within the four faculties and the SAS statistical package was used to regress the outcome variable on the predictor. The results of this analysis are summarized in **Table 2** below.

Table 2: The results of a linear regression of performance on PTESLAL as a predictor of end of year average performance (n=303)

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

6. Discussion

As pointed out earlier in this article, TALL is, according to its owners, a test of academic literacy aimed at measuring test taker ability to handle, with some degree of success, the language demands of academic education at the first year level of study. Any validity claim for this test to that effect will need to be based on a reasonable amount of evidence of a predictive relationship with first year academic performance in a broad sense. This is so because TALL is essentially a criterion-referenced test whose criterion of reference is academic performance. The key interpretation of performance on this test should therefore be that test takers who perform satisfactorily on it are unlikely to struggle with the language demands of academic education and that those who do not are likely to be hindered by language in their effort to study. Given its entirely criterion basis, the reference point for its validity should be that criterion itself. Choosing any other way to validate a test of this kind ahead of criterion validation would not leave the test unquestioned as it might mean that the test owner is not confident about its validity. Kane (2009: 44) rightly argues that in cases such this one where extrapolating the validity of a test of TALL's kind merely from its content, for example, is inadequate, "and the decisions to be made have moderate to high stakes, the relationship between the test scores and more direct measures of the performance of interest can be evaluated in a criterion validity study …" Kane (2009: 44) points out, in addition, "that the need for criterion-related evidence depends on the intended use of the test scores to estimate or predict some non-test performance." In the case of a test of academic preparedness such as TALL, this need is fundamental.

The key results of the predictive validity study of TALL are represented by the *t* and *p* values in the last two columns of **Table 1** above. The former is a measure of the predictive value of this test while the latter is an indication of whether the former is statistically significant. In applied linguistics research, the acceptable *p* value for statistical significance is 0.05 and below (See Mackey & Gass, 2005). Any *p* value above 0.05 means that the *t* value was merely a result of chance. So, the *p* value result of the predictive validity study of TALL presented in **Table 1** means that before the *t* statistic was not a matter of chance and that this test therefore possessed a degree of predictive relationship with the participants' average performance at the end of their first year of study. This justifies the way that performance on this test is interpreted by its developers: a measure of levels of readiness for university education with regard to language. In other words, just as the developers of the test claim it does, performance on this test evidently possessed a predictive relationship with the participants' end of first year average performance. This means that the way that these results are interpreted is appropriate and adequate, as Messick would say. For this to be the case, however, two conditions must first be satisfied. The first is that the test used must be underpinned by a valid construct. A valid construct for a test such as TALL is one for which evidence of its relationship with its criterion of reference is established. The second condition is that the types of tasks used to elicit levels of this construct are designed in a way that ensures that they are efficient in doing so. This is what was referred to as content validity earlier in this article. These conditions reside in the test itself and not necessarily in how its scores are interpreted. The bottom line, however, is that for tests of academic language ability such as TALL at least, it is construct and content validity that determine the predictive validity of such tests. The criterion referenced nature of these tests means that the appropriate way to interpret their results is that they relate positively with the criterion of interest and that this is not achievable if there are any flaws in the construct and content of these tests. For this reason, talking about the validity of these kind of tests only as a property of their scores does not make sense. On the contrary, not only does defining validity as a quality of a language test such as TALL make sense, it also protects it from possible misinterpretation and resultant misuse by those who might take advantage of the freedom that Messick's concept of validity affords them to interpret tests scores in a manner that they deem appropriate for them even though this interpretation might not align with the purport of the test involved. As I demonstrate with regard to PTESLAL below, tests of language ability such as TALL are by their very nature susceptible to misuse if their validity is not firmly located in such tests themselves and is left instead to how scores from these tests are interpreted and used.

It is the results of the validity study on PTESLAL as depicted in **Table 2** above that are particularly revealing on the weakness of the view of validity as a property of score interpretation and use in language testing. As can be seen from this table, the *p* value for the *t* statistic in this case was higher than 0.05, meaning that although it was positive, the *t* statistic was a result of mere chance. This means that performance on this test did not predict the outcome variable as assumed by its user. As pointed out earlier in this article, this test is used by the institution to judge students' academic readiness with regard to language. What this means is that from the point of view of this institution, PTESLAL is a test of the kind of language ability that is required for success at university study and which is now commonly known as academic literacy. Performance on this test is, in other words, interpreted to mean that students are academically literate enough to succeed academically if they perform to the satisfaction of the test user. When one considers what the owner of the test developed it for, however, one realizes immediately that the way that performance on this test is interpreted is inappropriate and in the view of Messick (1980; 1989), invalid. PTESLAL is, however, a test of English proficiency whose purport is different from that of a test of academic literacy such as TALL. The test came into being "in response to the perceived needs of education departments and various sectors of South African society" to measure the test takers' "level of general language development" (HSRC 1991: 15). The HSRC (1991: 15) defines the purpose of a proficiency test such as PTESLAL as follows:

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

It is clear from this that this test can therefore not be used to assess language readiness the way that a test of academic language ability like TALL can or should. It is this possibility for a test user to interpret the result of a test inappropriately that must have stimulated Messick to locate test validity in the way scores are interpreted. In cases where this happens, it makes sense for one to say that score interpretation is invalid. But does this also mean that the test yielding such scores is invalid? Not in the case of PTESLAL in this article. As can be seen in the HSRC's own description of the purpose of this test above, it is a test of general language proficiency that has been validated for that purpose only and not necessarily for assessing academic language readiness. In light of this, it makes sense for one to attribute validity to a test designed for a particular purpose and not necessarily to how its results are interpreted. The results also show that locating validity in how test scores are interpreted opens a test up for being used for purpose for which it was not developed. This is clearly the case with PTESLAL at the university that uses it for access. Test results can, in other words, easily be interpreted incorrectly if a test that is used to generate such results is left out in the definition of validity. This is particularly likely in language testing, where different language abilities need to be defined distinctly keeping in mind the specific context in which particular language resources are pertinent. The fact that context determines the kind of language ability required to accomplish a particular communicative task makes it all the more important that in language testing at least, definitions of validity are removed from score interpretation and are attached to a test whose particular purpose has a clear identity. This point is captured quiet clearly by Weideman (2017: 209) when he says,

> … we must not … overstretch the limitations of a language test by employing it for purposes outside of the technical range that it was designed to measure. A test designed for one purpose, say assessing academic literacy, cannot measure proficiency in lingually negotiating a successful business transaction, for example. Each artefact has technical boundaries and limits …

This means, in other words, that a language test should be accorded validity for the particular purpose for which it was developed. As implied by Weideman's (2017) observation above and shown by the validity study on PTESLAL dealt with earlier in this article, leaving the meaning of validity to how scores from that test are interpreted creates room for a possible misinterpretation of such scores and consequent misuse of the test itself.

The argument pursued in this article draws broader contextual support from the now well-known distinction that Cummins (2009) makes between the kind of language ability that is assessed in TALL as opposed to the kind needed to communicate in social settings as measured by PTESLAL, for example. Cummins refers to the former as Cognitive Academic Language Proficiency (CALP) and calls the latter kind Basic Interpersonal Communicative Skills (BICS). From the results of the present study, it is evident that academic literacy as defined for and measured in TALL is a different kind of ability from general language proficiency as measured in PTESLAL. Patterson and Weideman (2013) have observed that the kind of language ability that TALL purports to measure is distinct in that it requires analytical and logical thinking and that this is not necessarily a defining features of what Cummins describes as BICS. Patterson and Weideman (2013: 111) make this point thus:

> It is evident that the typicality of academic discourse is stamped and guided by specific dimension of experience – namely, the analytical. While each academic field is circumscribed by one or more modes of reality, … academic discourse as a whole is qualified by the analytical (or logical mode) … In other words, work within every academic discipline … is guided and led by the logical dimension of experience which involves analysis as its defining kernel.

These views underscore the importance of the need to distinguish between different kinds of language ability and by extension, the purpose for which a language test is valid. As demonstrated by the results of the two validity studies dealt with in this article, appropriate score interpretation depends on what a test is validly able to do. A test must, for example, first be valid as a test of academic language ability for its results to be interpreted as such. Otherwise, the results might be interpreted for a purpose that is unrelated to a valid test thereby invalidating it for its purported focus. No matter how one looks at it, validity is in the first instance a function of the degree to which a test measures what it was developed to measure. If it is unable to do this, its results cannot be interpreted otherwise. Similarly, if it is able to do it, its results can still not be interpreted otherwise. I say "in the first instance", since there are further ways in which validity can be disclosed, but in its original, undisclosed meaning it refers to the power of a test to measure according to its purpose. That is material for a further discussion, however, and can, for reasons of space, not be dealt with here.

From the argument above, it is clear that the purpose for which a language test is developed is contingent upon the context of language use involved. This shows that concerns about the validity of a test's purpose becomes a serious issue from the time that test itself is conceptualized and throughout the whole process of its development. It would not make sense in other words, for validity to be a property of test scores if it is in fact a crucial consideration in test development long before such scores come into being. More precisely, in language testing in particular, validity becomes an issue at the very earliest stage of test development when the particular language ability, also known as the construct, of interest to the test developer needs to be defined. The way that McNamara (2000: 13) explains this process of construct definition clearly shows how its intended purpose and ultimate validity are crucial determinants of what a test will do even before it is administered for score generation. In his view "defining the test construct involves being clear about what knowledge language consists of, and how that knowledge is deployed in actual performance (language use). Understanding what view the test takes of language use in the criterion is necessary for determining the link between test and criterion in … testing." This, McNamara (2000: 13) further argues, "is not just an academic matter. It has important implications, because according to what view the test takes, the 'look' of the test will be different, reporting of scores will change, and test performance will be interpreted differently". Clearly, the way that a test's construct is defined and how the test itself is designed is the epicentre of everything about such a test's validity. In fact, it makes sense to argue that test scores or how they are interpreted can only serve as the basis for determining whether the test itself does what it is intended to do. In others words, once developed and administered, a test's scores are the primary set of data on the basis of which its purported validity can be confirmed or disconfirmed. Bachman and Palmer (1996: 10) capture this point as follows:

> If we want to use the scores from a language test to make inferences about individuals' language ability, and possibly to make various types of decisions, we must be able to demonstrate how performance on that language test is related to language use in specific situations other than the language test itself.

The point that Bachman and Palmer's (1996: 10) make here about one being "able to demonstrate how performance on that language test is related to language use in specific situations other than the language test itself" underlines the role of test scores as data for investigating whether a test is valid for the particular purpose that necessitated its origin. This is particularly true for the construct and criterion-related types of validity for which empirical procedures are available to allow a quantitative analytic validation of a test on the basis of its scores. The argument that scores are the basis for investigating a test's validity makes complete sense when one acknowledges that tests can exist without scores while scores cannot exist independent of a test. The relationship between tests and scores is, in other words, not of a chicken and egg kind: Tests give rise to scores but scores do not give rise to tests. Similarly, valid or invalid tests produce valid or invalid results, not the other way round.

8. Conclusion

The aim of this article was to engage with the current debate on the meaning of test validity. The basis of this debate is the lack of consensus over whether test validity is a function of the way test results are interpreted and used, or a quality of the test that produces those results. These are the two polarised views on the meaning of the term validity, also known as the unitary and traditional views, respectively. The article argues that the unitary view, the one in which validity is regarded as a property of test score interpretation and use, is not defensible for language testing. It presents two validity studies of two language tests, one that was developed to measure general language proficiency and another one developed to measure academic language ability, to demonstrate how this view of validity can lead to the misuse of language tests. More specifically, a article shows how a test of general language proficiency has been used to take invalid decisions about academic language readiness by a university as a probable result of lack of understanding by its user that these two types of language ability are different and that decisions related to a specific language ability cannot be taken on the basis of performance on any language test.

In this article, support for this point resides in the fact that a criterion related validity study of two tests showed that general language proficiency is not as good a predictor of academic performance as a test designed specifically for assessing academic language ability is.

The article's broader context for this argument is that the current approach to language teaching and testing begins with a definition of the particular language ability to be taught and tested as dictated by the relevant context. It would be unwise therefore for anyone to offer a language course or use a language test in a context that is not in sync with the specific purpose of these artefacts.

The main argument of this article is that locating validity in the way test scores are interpreted and used gives freedom to language test users to use tests as they see fit and not necessarily for the purpose they were developed. This is not to say that the view of validity as a function of score interpretation and use was definitely the basis in the case of the wrong reason for which the language proficiency test dealt with in this study has evidently been used. The point of this article, however, is that the more the meaning of validity continues to be understood to reside in the way scores are interpreted and used, the more likely it will happen that test users will use language tests developed for a specific purpose for assessing any language ability they are interested in and which may have very little to do with the very purpose of the tests themselves. This, as argued throughout this article, promotes language test misuse and invalidity.

References

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. 1999. *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Babbie, E. 2013. *The practice of social research*. Wadsworth: Cengage Learning.

Bachman, L.F., & Palmer, A. S. 1996. *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press.

Boorsboom, D., Mellenbergh, G. J., & Van Heerden, J. 2004. The concept of validity. *Psychological Review*, 111, 1061-1071.

Cummins, J. 2009. Fundamental psychological and sociological principles underlying educational success for linguistic minority students. In Skutnab-Kangas, T., Phillipson, R., Mohanty, A. K. & Panda, M. (Eds.), *Social justice through multilingual education*. Bristol: Multilingual Matters, pp. 19-35.

Cronbach, L. J., & Meehl, P. E. 1955. Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

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

Human Sciences Research Council. 1991. *Manual for proficiency test English second language advanced level*. Pretoria: Human Sciences Research Council.

Kane, M. T. 2006. Content-related validity evidence in test development. In Downing, S. M. & Haladyna, T. M. (Eds.) *Handbook of test development*. New York: Routledge. 131-153.

Kane, M. T. 2009. Validating the interpretations and uses of test scores. In Lissitz, R. W. (Ed). *The concept of validity: Revisions, new directions, and applications*. Charlotte, NC: Information Age Publishing, INC. pp. 39-64.

Le, P.L., du Plesssis, C. & Weideman, A. 2011. Test and context: The use of the Test of Academic Literacy Levels (TALL) at a tertiary institution in Vietnam. *Journal for Language Teaching*, 45 (2): 115-131.

Lissitz, R. W., & Samuelsen, K. 2007. A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36, 437-448.

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

McNamara, T. 2000. *Language testing*. Oxford: Oxford University Press.

Messick, S. 1980. Test validity and the ethics of assessment. *American Psychologist*, 35, 1012 -1027.

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

Newton, P. E. & Shaw, S. D. 2014. *Validity in educational and psychological assessment*. Los Aangeles: SAGE.

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

Rambiritch, A. 2012. Transparency, accessibility and accountability as regulative conditions for a postgraduate test of academic literacy. PhD thesis. Bloemfontein: University of the Free State. URI: http://hdl.handle.net/11660/1571.

Sebolai, K. 2016. The incremental validity of three tests of academic literacy in the context of a South African university of technology. PhD thesis. Bloemfontein: University of the Free State. Available: http://hdl.handle.net/11660/5408.

Sireci S. G. 2009. Packing and unpacking sources of validity evidence: History repeats itself again. In Lissitz, R. W. (Ed). *The concept of validity: revisions, new directions, and applications*. Charlotte, NC: Information Age Publishing, INC. pp. 19-37.

Van der Slik., F. 2008. Gender bias and gender differences in tests of academic literacy. *Southern African Linguistics and Applied Language Studies Special* *issue*: Assessing and developing academic literacy (Ed. Geldenhuys, J.), 27 (3): 277-290.

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

Van der Slik, F., & Weideman, A. 2009. Revisiting test stability: Further evidence relating to the measurement of difference in performance on a test of academic literacy. *Southern African Linguistics and Applied Language Studies*, 27 (3): 253-263.

Van der Slik, F., & Weideman, A. 2010. Examining bias in a test of academic literacy: Does the Test of Academic Literacy Levels (TALL) treat students from English and African language backgrounds differently? *Journal for Language Teaching*, 44 (2): 106-118.

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

Van Dyk, T. 2015. Tried and tested. *Tijdchrift voor Taalbeheersing*, 37(2): 159-186.

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

Van der Walt, J. L, & Steyn, F. 2008. The validation of language tests. *Stellenbosch Papers in Linguistics*, 38, 191-204.

Weideman, A. 2003. Assessing and developing academic literacy. *Per Linguam*, 19 (1&2): 55-65.

Weideman, A. 2009. Constitutive and regulative conditions for the assessment of academic literacy. *Southern African Linguistics and Applied Language Studies,* 27: 1-26.

Weideman, A. 2011. Academic literacy tests: Design, development, piloting and refinement. *Journal for Language*

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

Weideman, A. 2017. *Responsible design in applied linguistics: theory and practice*. Cham: Springer International Publishing. [Online]. DOI 10.1007/978-3-319-41731-8.

ABOUT THE AUTHOR

Kabelo Sebolai

Language Centre: Stellenbosch University

Email: ksebolai@sun.ac.za

Kabelo Sebolai is the deputy director in the Language and Communication Development section of the Language Centre at Stellenbosch University. His professional background is Teaching English to Speakers of Other Languages (TESOL). His research interest revolves around academic literacy teaching and assessment.

Tydskrif vir Taalonderrig - Journal for Language Teaching
- Ijenali yokuFundisa iLimi - IJenali yokuFundisa iiLwimi -
Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta Polelo
- Buka ya Thuto ya Puo - Jenale ya Thuto ya Dipuo - Ijenali
Yekufundzisa Lulwimi - Jena?a ya u Gudisa Nyambo
- Jenala yo Dyondzisa Ririmi - Tydskrif vir Taalonderrig -
Journal for Language Teaching - Ijenali yokuFundisa iLimi
- IJenali yokuFundisa iiLwimi - Ibhuku Lokufundisa Ulimi
- Tšenale ya tša Go ruta Polelo - Buka ya Thuto ya Puo -
Jenale ya Thuto ya Dipuo - Ijenali Yekufundzisa Lulwimi
- Jena?a ya u Gudisa Nyambo - Jenala yo Dyondzisa
Ririmi - Tydskrif vir Taalonderrig - Journal for Language
Teaching - Ijenali yokuFundisa iLimi - IJenali yokuFundisa
iiLwimi - Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta
Polelo - Buka ya Thuto ya Puo - Jenale ya Thuto ya Dipuo -
Ijenali Yekufundzisa Lulwimi - Jena?a ya u Gudisa Nyambo
- Jenala yo Dyondzisa Ririmi
- Tydskrif vir Taalonderrig
- Journal for Language
Teaching - Ijenali
yokuFundisa iLimi -
IJenali yokuFundisa
iiLwimi - Ibhuku
Lokufundisa Ulimi
- Tšenale ya tša
Go ruta Polelo -
Buka ya Thuto
ya Puo - Jenale
ya Thuto ya Dipuo
- Ijenali Yekufundzisa
Lulwimi - Jena?a ya u
Gudisa Nyambo - Jenala yo
Dyondzisa Ririmi - Tydskrif vir Taalonderrig
- Journal for Language Teaching - Ijenali
yokuFundisa iLimi - IJenali yokuFundisa iiLwimi -
Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta
Polelo - Buka ya Thuto ya Puo - Jenale ya Thuto ya
Dipuo - Ijenali Yekufundzisa Lulwimi - Jena?a ya
u Gudisa Nyambo - Jenala yo Dyondzisa Ririmi
- Tydskrif vir Taalonderrig - Journal for Language
Teaching - Ijenali yokuFundisa iLimi - IJenali
yokuFundisa iiLwimi - Ibhuku Lokufundisa Ulimi -
Tšenale ya tša Go ruta Polelo - Buka ya Thuto ya Puo -
Jenale ya Thuto ya Dipuo - Ijenali Yekufundzisa Lulwimi
- Jena?a ya u Gudisa Nyambo - Jenala yo Dyondzisa
Ririmi - Tydskrif vir Taalonderrig - Journal for Language
Teaching - Ijenali yokuFundisa iLimi - IJenali yokuFundisa
iiLwimi - Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta
Polelo - Buka ya Thuto ya Puo - Jenale ya Thuto ya Dipuo -
Ijenali Yekufundzisa Lulwimi - Jena?a ya u Gudisa Nyambo
- Jenala yo Dyondzisa Ririmi - Tydskrif vir Taalonderrig -
Journal for Language Teaching - Ijenali yokuFundisa iLimi
- IJenali yokuFundisa iiLwimi - Ibhuku Lokufundisa Ulimi
- Tšenale ya tša Go ruta Polelo - Buka ya Thuto ya Puo -
Jenale ya Thuto ya Dipuo - Ijenali Yekufundzisa Lulwimi
- Jena?a ya u Gudisa Nyambo - Jenala yo Dyondzisa
Ririmi - - Tydskrif vir Taalonderrig - Journal for Language
Teaching - Ijenali yokuFundisa iLimi - IJenali yokuFundisa
iiLwimi - Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta