Kabelo Sebolai
Central University of Technology

Disparate Impact, Justice and fairness: A case study of the Test of Academic Literacy Levels

Abstract

The teaching of academic literacy has become critically important at South African universities in the post-apartheid period. The reason for this is that universities that were previously exclusively accessible to white students are currently within reach of non-white students. Most of these students, however, graduated from public schools where they received poor education in English, a medium of instruction at most universities. A result of this has been that the students struggle to handle the demands of university education in the language. This contributes to their failure to complete their studies in scheduled time and to drop out. As a way to deal with this challenge, Higher Education South Africa (HESA) has introduced a National Benchmark Test of academic literacy to assess the reading, writing and thinking abilities of these students to ensure their proper placement at universities. The Inter-institutional Centre for Language Development and Assessment (ICELDA), a partnership of the Universities of Pretoria, Stellenbosch, North-West and Free State, has also developed a test of academic literacy known as the Test of Academic Literacy Levels (TALL) for the same purpose. This paper was a case study of the impact, justice and fairness of this test. The findings were that the test possessed an acceptable degree of justice and fairness and that it aimed for a positive impact on the test-taker.

Keywords: impact, justice, fairness, academic literacy, test.

Journal for Language Teaching, Volume 48, Number 1, pp. 89 - 107. 2014.
ISSN 0259-9570. http://dx.doi.org/10.4314/jlt.v48i1.5

1. Introduction

In the period starting from 1948 and ending in 1993, the South African society bore witness to the introduction and institutionalization of apartheid laws that promoted the segregation of South Africans on the basis of race. A product of these laws was the Bantu Education Act of 1953. In accordance with the broader apartheid policies, Bantu Education introduced segregated and discriminatory education for Coloureds, Indians and Blacks; access to education was free and compulsory for whites while 'non-whites' had to pay for their education (Mdepa & Tshiwula, 2012: 20). Not only did Bantu Education have racial segregation and discrimination as its goals, its ultimate aim was to ensure that good quality education was accessible to whites only and that poor quality education was delivered to non-whites to prepare them for subservient roles in the South African society.

In contrast, the new constitution (Act 108 of 1996) of a democratic South Africa introduced rights of citizenship and equality for all (Mdepa & Tshiwula, 2012: 21). For example, Section 29 of this constitution guaranteed the right to all levels of education for all races while Section 29.2 (c) "refers to the need to redress the results of past discriminatory laws and practices that institutionalized difference" (Mdepa & Tshiwula, 2012: 21). A result of these laws was an influx of non-white students to historically white universities:

> Higher education institutions opened their doors to all race groups. As noted by (Badat, 2010: 7), total student enrolment increased from 473, 000 in 1993 to 799, 388 in 2008. In 1993, 40% of all students were African (191, 000 students), and 52% were black; by 2008, African enrolment had risen to 64.4% (514, 370 students) and black enrolment stood at 75% of overall enrolment (Mdepa & Tshiwula, 2012: 22).

Historically white universities, however, use English as one of the languages of learning and teaching, an additional language to most non-white students and one in which they received poor education at high school. Consequently, the students struggle to succeed in dealing with the reading, writing and thinking demands of university education in the language, and this factor contributes to their failure to complete their studies in time or even to graduate at all. In the face of this challenge, Higher Education South Africa (HESA) has introduced a National Benchmark Test (NBT) of academic literacy to measure the reading, writing and thinking ability of first year university students to ensure that they are properly placed within universities and that their curriculum needs can be appropriately met. Similarly, the Inter-institutional Centre for Language Development and Assessment (ICELDA), a partnership of the Universities of Pretoria, Stellenbosch, North-West and Free State, has designed a test of academic literacy known as the Test of Academic Literacy Levels (TALL) for the same reason (Le, Du Plessis & Weideman, 2011). The scores from this test are used to decide whether the students need to enroll for academic literacy interventions. In other words, cut-scores or benchmarks are determined and used either to exempt the students from or to enroll them in compulsory academic literacy programmes. Given the relatively high stakes purpose for which the test is used, it is important that its social impact, justice and fairness are investigated.

2. Fundamental concepts in language testing

Language testing is a subfield within the broader field of applied linguistics. Applied linguistics has been defined by Weideman (2006: 72) "as a discipline that devises solutions to language problems". In view of this, it is necessary that we "develop a theory of applied linguistics which shows that the constitutive and regulative conditions exist for doing applied linguistics designs" (Weideman 2007: 30). In applied linguistics, such designs include language tests, language courses and language policies. In language testing, Weideman's (2009: 1) constitutive conditions of test design include validity, reliability, and a test's theoretical defensibility while the regulative conditions include among others, accessibility, transparency, accountability, impact, justice and fairness. Weideman's (2009: 2012) separation of the constitutive from the regulative conditions of test design is an indication on his part that test appraisal does not only involve a consideration of the empirical (constitutive requirements) but that it also requires paying attention to the social (regulative) dimensions of testing. Indeed, in the words of Rambiritch (2012b: 112), the constitutive conditions of test design namely, validity and reliability,

> do not function in isolation but in harmony or accordance with other factors, qualities or modes such as the lingual, the social, economic, aesthetic, juridical and ethical dimension of reality, and the way that these are reflected in concepts and ideas such as, respectively, the technical interpretability of the scores/outcomes of the test, the implementation of the test, its technical utility, alignments with needs of students and administrators, transparency, accountability and fairness.

This is no less relevant to the field of language testing in particular 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). It is for this reason that McNamara and Roever (2006: 2) have further argued that "a psychometrically good test is not necessarily a socially good test". This is an important observation "because a core concern here is the social responsibility that test developers have, not just to the test taker but to everyone affected by the test – supervisors, parents, test administrators and society at large" (Rambiritch, 2012b: 109). Part of this responsibility involves ensuring that testing is carried out justly and fairly and that it has a positive impact on all those who holds a stake in it.

2.1 Impact, justice and fairness in language testing

In language testing, the term 'impact' refers to the consequences of the decisions taken on the basis of test performance. These consequences can be either positive or negative. The word 'justice' relates to the degree to which such consequences are in the interests of the test-taker. Test impact and justice are two of the regulative conditions of test design that relate to the notion of test 'fairness'. Test 'fairness' refers to the degree to which "a test is used in an impartial, just, and equitable way" (Cohen & Swerdlik, 2010: 203). In this sense, a test that is fair should have a positive impact on those involved and should therefore possess a degree of justice and vice versa. Thus, while they are important as regulative conditions of testing in their own right, test impact and justice are intertwined with test fairness. Kunnan (2004: 27) has argued that “the concept of test fairness is arguably the most critical in test evaluation” because language test scores are often used for making a number of critical decisions such as the selection and placement of students, and as a tool to assess learning progress and diagnose learning difficulties (Bachman & Palmer 1996: 96-97). Language test scores are also used by some universities for student certification, language programme evaluation, and teacher professional development (Bachman & Palmer 1996: 96-97). Kunnan (2004: 31) observes, however, that the importance of test fairness, especially in the current ‘use-oriented’ testing milieu, has resulted in the practice of focusing test validation on the empirical properties of a test at the expense of the social dimensions of testing, especially test impact, justice and fairness. This restricted focus on reliability and validity in test appraisal has prompted Kunnan to generate a framework that may be used to evaluate test fairness and by extension, impact and justice. This framework is captured in Table 1 below.

Table 1: Kunnan’s framework of test fairness

Main quality	Main focus
1. Validity	
Content	Representativeness of items, tasks, topics
Representativeness/coverage	
Construct or Theory based validity	Representativeness of construct/ underlying trait
Criterion-related validity	Test score comparison with external criteria
Reliability	Stability, alternate form, inter-rater and internal consistency

Main quality	Main focus
2. Absence of bias	
Offensive content or language	Stereotypes of population groups
Unfair penalization	Content bias based on test takers background
Disparate impact and standard setting	DIF in terms of test performance; criterion setting and selected decisions
3. Access	
Educational	Opportunity to learn
Financial	Comparable affordability
Geographical	Optimum location and distance
Personal	Accommodations for test takers with disabilities
Equipment and condition	Appropriate familiarity
4. Administration	
Physical setting	Optimum physical setting
Uniformity and security	Uniformity and security
5. Social consequences	
Washback	Desirable effects on instruction
Remedies	Re-scoring, re-evaluation, legal remedies

(Kunnan 2004: 46)

As can be seen from **Table 1** above, Kunnan's framework consists of five main qualities. These are validity, absence of bias, access, administration and social consequences. In the words of Kunnan (2004: 37) himself, the framework looks at fairness from the point of view "of the whole system of a testing practice, not just the test itself". Thus, except that Kunnan groups the empirical and social requirements of test design under the umbrella concept of ‘fairness’, his framework reflects his agreement with Weideman’s (2009; 2012) view that both the constitutive and regulative conditions of test design merit disparate attention in test design and appraisal.

Arguably, a lot has already been written about the empirical or, as Weideman (2009; 2012) calls them, constitutive qualities of test design that also feature in Kunnan’s Test Fairness Framework, namely validity and reliability. The words of Rambiritch (2012a: 87) attest to this:

> Many testing professionals hold the view that testing research has always focused on issues of fairness through the concepts of validity and reliability, indicating … a concern only with empirical evidence and, preferably, that kind of empirical evidence that can be stated in numbers.

To some extent, studies on the properties of TALL have not been immune to this. For this reason, the four regulative qualities of fairness featuring in Kunnan’s (2004) framework namely, absence of bias, access, administration and social consequences are considered in this study to evaluate the impact, justice and fairness of this test. More specifically, research evidence for the absence of bias in the test, the use of remedies for dealing with the social consequences of the test, efforts to establish test-takers’ familiarity with the test, and the degree to which the administration of the test is standardized are investigated to establish the degree to which TALL is just and fair and the nature of its impact on those who take it. In the rest of the article, the concepts of justness and beneficence will be explored in relation to TALL.

3. TALL and justice

While some scholars (e.g. Cohen and Swerdlik 2010: 199) have argued that validity and test bias should be treated as separate issues in test design and appraisal, ensuring that a test is free from bias is, by logic, closely associated with the validity of a test and the appropriateness of the interpretation and use of scores obtained on such a test. It is difficult, if not impossible, therefore, to talk about test impact, justice and fairness, the ultimate culminations of test bias, without linking them especially to construct validity. This does not mean, however, that all these concepts are the same. Indeed, Kunnan’s framework on test fairness separates validity from the unempirical requirements of test design to underline this point. Similarly, Weideman (2009; 2012) has classified validity as a constitutive component of testing on the one hand, and impact, justice and fairness as regulative requirements of test development on the other.

From the point of view of psychological measurement, test bias “is a factor inherent in a test that systematically prevents accurate and impartial measurement” (Cohen & Swerdlik, 2010: 199). Test bias is therefore a consistent and systematic failure by a test to provide a reliable and justifiable measurement of an ability a test was designed to measure, as a result of some factor that is a function of the background of the test-takers involved and that is unrelated to the construct underpinning the test. In other words, a test is biased in favour of test-takers of a common background, such as males, if it discriminates against another group of test-takers such as females. Such a test would be male-oriented in some way and would make it impossible for its user to make meaningful inferences about both the males and female students involved. The impact, justice and fairness of such a test would be questioned, since its construct would be giving a measure unrelated to what it can validly test. In the words of Jensen (1980: 444), the essence of a test that is fair and just and whose impact is positive because of the absence of bias in its content is that

> … any person showing the same ability as measured by the whole test should have the same probability of passing any given item that measures that ability, regardless of the person's race, social class, sex, or any other background characteristics. In other words, the same proportion of persons from each group should pass any given item of the test, provided that the persons all earned the same total score on the test.

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

For the purpose of determining, from the point of view of test bias, the impact, justice and fairness of TALL further, Van der Slik and Weideman (2010) conducted a study to investigate if the test would function differently for students at the Universities of Pretoria, Stellenbosch and North-West from three first language backgrounds, namely, African languages, English and Afrikaans. T-tests were used and Differential Item Functioning (DIF) analyses were carried out by means of the Mantel-Haenszel statistic to determine this. The outcomes of the T-tests and DIF analyses of performance by the three groups of students are shown in Table 2 below:

Table 2: T-values of differences between mean scores on TALL of first year students who have an African language, English, or Afrikaans as their first language

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

[1]: with Bonferroni adjustment

(Van der Slik & Weideman 2010: 111)

The T-tests and DIF statistics as indicated in **Table 2** above show that there were negligible differences in performance between the three different native language groups that took this test. The overall finding, however, was that the DIF could be accounted for by the less proficient test takers' lack of ability to complete all the test tasks and that the DIF was evidently not related to the content of the test items. Van der Slik and Weideman (2010: 115) explain this finding as follows:

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

To this end, the two studies show that TALL has been consistent with the first of the two principle of Frankena's system from which Kunnan's framework of test fairness and by extension, impact and justice derives (Kunnan, 2004: 33):

Principle 1: *The principle of justice:* A test ought to be fair to all test takers; there is a presumption of treating every person with equal respect.

Sub-principle 1: A test has to have comparable construct validity in terms of its test score interpretation for all test takers.

Sub-principle 2: A test ought not to be biased against any test-taker groups, in particular by assessing construct-irrelevant matters.

4. TALL and beneficence

As pointed out earlier, in most cases, testing is not carried out for measuring a trait, knowledge or ability for the mere sake of it. It is, in the words of Gregory (2007: 127), commonly carried out "in the service of decision making". Gregory (2007: 127) gives a few of the general decisions often taken on the basis of test scores:

> The personnel manager wishes to know whom to hire; the admissions officer must choose whom to admit; the parole board desires to know which felons are good risks for early release; and the psychiatrist needs to determine which patients require hospitalization.

It is evident from all the reasons for testing that Gregory advances above that using testing for decision making is particularly relevant to predicting future behavior. Since no test is 100% perfect, however, it is possible, for example, that a test can predict that examinees who actually fail will pass and that those who actually pass will fail (Gregory, 2007: 128; Cohen & Swerdlik, 2011: 189-190). These cases are known as false positives and false negatives respectively (Gregory, 2007: 128; Cohen & Swerdlik, 2010: 189-190). In the words of Whiston (2013: 68), "A false positive occurs when the instrument predicts that individuals have 'it' (the criterion) when in fact they do not" while "a false negative occurs when the instrument predicts that the test takers do not have it when in fact they do" (Whiston, 2013: 68). Put differently,

> A **false positive** classification error is when we classify a test taker into the higher, or mastery group, when his ability is actually at the level of the lower, non-mastery group. A **false negative** classification error, on the other hand, is when we classify a test taker into the lower group, when his ability is actually at the level of the higher group (Bachman, 2004: 198-199).

Together, false positives and false negatives are known as *misses* because in both cases a test makes inaccurate predictions (Gregory, 2007: 128). The opposite of *misses* are called hits. These are the cases in which a test correctly predicts future test-taker performance or behavior. A way to deal with *misses* in order to ensure justice, fairness and positive impact in testing involves making use of what is, in the language of psychometrics, known as Decision Theory (Whiston, 2013; Erford, 2013). Applying this theory to the design and development of tests involves using established procedures to determine the accuracy of the decisions taken on the basis of test scores. Decision Theory enables the test developer to determine the frequency with which tests accurately classify the test-takers and how often they classify such test-takers inaccurately (Whiston, 2013; Erford, 2013). In other words, through Decision Theory procedures, the test developer is able to identify what are known as false positives and false negatives. In essence, false positives and false negatives are misclassifications that are inherent to testing because, as pointed out earlier, no test is 100% flawless. Gregory (2007: 128) explains this situation thus:

> False positives and false negatives are unavoidable in the real world use of selection tests. The only way to eliminate such selection errors would be to develop a perfect test, an instrument which has a validity coefficient of +1.00, signifying a perfect correlation with the criterion measure. A perfect test is theoretically possible, but none has been observed on this planet. Nevertheless, it is still important to develop selection tests with very high predictive validity, so as to minimize errors.

Given the cost that might result from a classification test's inaccurate prediction of a test- taker's level of ability, knowledge or trait of interest to the test user and the need for testers to ensure positive test impact, justice and fairness in testing, it is necessary that test designers and developers find ways to handle misclassifications (Van der Slik & Weideman, 2005).

The analysis of the scores from TALL has involved the use of the TiaPlus software package to identify these misclassifications. The software has enabled the developers of the test to use two types of scenarios that are derived from Cronbach's alpha and Greatest Lower Bound (GLB) statistics to identify false positives and negatives (Weideman, 2011). These scenarios are the correlation between TALL and a hypothetical parallel test as well as the correlation between observed and 'true' scores (Van der Slik & Weideman, 2005; Van der Slik & Weideman, 2009; Weideman, 2011). Based on the results of this analysis, false negatives in particular are given a second chance to demonstrate their academic literacy levels, and parameters are set for determining the size of such false negatives (Van der Slik & Weideman, 2009). This is the extent to which the developers and users of TALL attempt to ensure that the test has a positive impact and that it is just and ultimately fair to all those who take it. In the words of Cohen and Swerdlik (2010: 203), test fairness is, as indicated earlier, "the extent to which a test is used in an impartial, just, and equitable way". This means that a test whose results are unfairly used especially for those who are misclassified as not having the relevant criterion falls short of meeting the regulative test design criteria of impact, justice and fairness (cf. Kunnan, 2000; Weideman, 2009). The statistics of the potential misclassifications of the test-takers of TALL at the Universities of Pretoria, Stellenbosch and North-West from 2005 to 2008 are shown in Table 3 below:

Table 3: Potential misclassifications on the English version of the academic literacy test (Percentage of this tests population). [In italics the corresponding interval (in terms of standard deviations) around the cut-off points.]

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

(Van der Slik & Weideman, 2009: 258)

As can be seen from the last row in **Table 3** above, in TALL, false negatives have generally been found to "occur more or less within the expected range of scoring points around the cut-off point, i.e. around 0.25 standard deviations around the cut-off point" (Van der Slik & Weideman, 2009: 258). This is evidence that the test had had a fairly positive impact and that it had been reasonably just and fair because the extent to which it had been misclassifying the test-takers had been minimal. As pointed out earlier, the extent to which the test has met these regulative conditions of test quality, has further been enhanced by giving those who are potentially misclassified a second chance to take the test.

Thus, this study shows that TALL has been consistent with the second of the two principle of Frankena's system from which Kunnan's framework of test fairness and by implication, impact and justice derives (Kunnan, 2004: 34):

Principle 2: *The principle of beneficence:* A test ought to bring about good in society; that is, it should not be harmful or detrimental to society.

Sub-principle 2: A test ought to promote good in society by providing test score information and social impacts that are beneficial to society.

Sub-principle 2: A test ought not to inflict harm by providing test-score information or social impacts that are inaccurate or misleading.

5. The experiences of TALL test-takers concerning impact, justice and fairness

In the main, tests are designed, developed and administered to measure the test-taker's mastery of the ability that the test user is interested in. To use the words of Davies (1990: 17), tests are "intended above all to clarify the difference in the matter under test, in what is being tested (proficiency, aptitude, achievement) among the candidates". In language testing, however, studies have generated evidence to show that variance in test scores is also affected by the different processes, experiences and strategies that test takers engage in when taking a test as well the degree of the test taker's access to or familiarity with a test (Bachman, 2004). The role played by these processes and experiences should therefore be considered when a test's construct validity, impact, justice and fairness are under scrutiny (Messick, 1989). Bachman (2004: 276) raises questions that point to the relevance of these experiences to the validity of a test's construct and by extension, its impact, justice and fairness:

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

Measurement researchers have addressed this concern by asking test-takers to give a report of their own experiences of taking a test (Van der Walt & Steyn, 2007; 2008). Such a report can be generated by the test-taker while in the process of responding to test tasks in what is known as "think aloud" protocols (Bachman, 2004: 276; Van der Walt & Steyn, 2008). Alternatively, the report can be compiled after the test is taken in what is called a retrospective verbal report (Bachman, 2004: 276). Records of these verbal reports are known as verbal protocols and can subsequently be qualitatively and quantitatively analyzed by the test developer in what is known as verbal protocol analysis (Bachman, 2004: 276). Both these reports are a way to enable the test developer to determine especially the extent to which a test is accessible or familiar to those who take it.

In a bid to establish the construct validity, impact, justice and fairness of TAG from the angle of test-taker experience, Van der Walt and Steyn (2007) distributed questionnaires to extract feedback from a group of 754 test-takers at the Potchefstroom campus of the North-West University regarding their familiarity with the tasks used in the test. The feedback the two researchers received was that the test was not adequately transparent and that its developers had to make some effort to make the test and its format more familiar to test-takers. Secondly, using the same questionnaire, Van der Walt and Steyn (2007) elicited information from the test-takers regarding their perception of the conduciveness of the conditions under which the test was administered. The general perception of the test-takers was that such circumstances were not ideal and that this could negatively impact the validity of the test's scores and the degree of its justice and fairness as well as the nature of its impact. Thirdly, the researchers wanted to establish the test-takers' perception of whether the test seemed relevant to their studies. Only 45% of the respondents felt that the test had relevance to their studies. Finally, Van der Walt and Steyn (2007) aimed at finding out through the questionnaire whether the test-takers were clear about what was required of them by the test tasks. Only 68% of the respondents indicated that they were confident about how they were expected to respond to most tasks.

Asked if they could finish taking the test in the allotted time, only 14 percent indicated they had been able to do so. The feedback from the takers of the test showed that some aspects related to its impact, justice and fairness needed attention. Firstly, one must point out, however, that the fact that the developers of TALL have made the test available for external scrutiny on their perception of various aspects of the test, especially its familiarity to them, is aimed at enough justice, fairness for the test-taker and a positive impact. Secondly, some room should be allowed for the shortcomings of the test as revealed by this study (Van der Walt & Steyn, 2007) because everybody is in agreement that no test is 100% perfect.

Like Van der Walt and Steyn (2007) did in their study, it is important that testers obtain information about test-taker perceptions of a test because language tests have often been so unfairly used that this has attracted a degree of criticism on the impact, justice and fairness of language testing. The language testing literature is full of examples of this unfair use of tests. The earliest example of this is according to McNamara (2004) and McNamara and Roever (2006) a one item language test in the Bible where, in a situation of war, people were asked to pronounce the word 'Shibboleth' to determine if they were Ephraimites or Gileadites. The 42 000 Ephraimites who could not pronounce the word were put to death by the Gileadites. Also, McNamara (2004: 774) gives a recent example of how language tests can sometimes be misused. According to him, in the 1990s, the German communities from former Eastern bloc countries who tried to immigrate into a united Germany were administered a German language test in the form of an interview. The presence of any evidence of non-standard forms of German in the applicant's speech, was interpreted to mean that they were not proper Germans and were, as a result, denied access into the country.

Against the background of this unfair use of tests, through two imaginary characters engaged in an imaginary conversation, Fulcher and Davison (2008) make the point that human beings are deprived of happiness by institutions in society and that the most evil of such institutions is testing. To this end, one of these characters argues that,

> Testing is the method by which the powerful remain in power and decide what knowledge is to be valued. The test takers are mere objects that have no choice but to comply with the demands of the powerful. The purpose is to establish domination through endless testing, thereby placing value on what is cherished by the powerful, thus maintaining society's status quo (Fulcher & Davidson, 2008: 408).

Conversely, Fulcher and Davidson (2008: 412) argue that "… tests, used correctly, have the power to grant access to opportunities and goods that were previously unavailable to the ordinary people." From the empirical evidence generated by studies related to the impact, justice and fairness of TALL to date, it is evident that the test's aim is to promote "access to opportunities and goods that were previously unavailable to the ordinary people (Fulcher & Davidson, 2008: 412". TALL is a test of academic literacy used to measure the levels of academic literacy among first year university students. In the words of Rambiritch (2012a: 30), "it is used to determine whether the student is equipped with the knowledge, language ability and skills needed to deal with the kind of language she or he will encounter specifically at university level."

At the University of Pretoria, for example, students who obtain low scores from the test stand the risk of failing to succeed at their studies and are therefore required to enroll in an academic literacy intervention programme offered by the university to help them develop the academic language abilities they need for success. TALL is, in this sense, aimed at having a positive effect on the academic lives of the students involved as well as being just and fair to them. Low levels of proficiency in academic literacy among students are, according to Weideman (2003: 56), risky "(a) for students, who fail to complete their courses in time; (b) for parents (who have to foot the bill for additional years of study); (c) for themselves (universities) in the loss of subsidy; and for the education system as a whole".

6. Considering the administration of TALL in the service of impact, justice and fairness

In educational and psychological assessment, the phrase "test administration" is a term used to refer to the process of giving a test (Kaplan & Saccuzzo, 1997: 12). Gregory (2007: 17) has argued that commonly, a misconception exists among psychologists and educators that test administration is a simple and straightforward procedure that can be carried out by anyone. In other words, some educators and psychologists commonly believe that test administration merely involves "passing out forms and pencils, reading instructions, keeping time, and collecting the materials" (Gregory, 2007: 17). Test administration is, however, a factor in the reliability and validity of a test and therefore impacts the meaningfulness, consequences, justice and fairness of how test scores are interpreted and ultimately used. It is important therefore that it is carried out with care and that those who participate in it are adequately trained and familiar with all the procedures involved. In the words of Gregory (2007: 17), "… careless administration … can impair group test results, causing bias for the entire group or affecting only certain individuals." Gregory (2007) mentions two aspects of test administration that can impact the consequences, justice and fairness of a test such as TALL. These are the failure by those who administer a test to keep to the time allocated for the test, differences regarding the physical condition under which a test is administered and noise (Gregory, 2007). The manual for TALL states, however, that it is a standardized test which presupposes that,

> certain standard criteria … are maintained at a constant level from one test to the next. The criteria dictate standard procedures for conducting the test… There is a link between the reliability and the validity of a test and its standardization. If the standard procedures are not complied with, the reliability of the test is influenced, resulting in possible discrimination against certain students (Van Dyk, 2006: 1).

While this does not guarantee that the users of the test will adhere to its standardized procedures of administration, it is just and fair to the test taker that the manual for this test contains information aimed at promoting a standardized, just and fair administration of the test and that the test should therefore impact those who take it positively. The manual outlines detailed pre-, while-, and post administration procedures that must be adhered to for this purpose. In the words of Gregory (2007: 19), in the kind of group testing for which tests like TALL are used, "deviations from the instructions are simply unacceptable."

7. Conclusion

Language testing is, undeniably, a critical aspect of first and additional language teaching and learning programmes. Its sometimes negative impact on those involved notwithstanding, it remains the common means through which teachers get to know how much progress their students are making or if they have ultimately achieved the objectives of a language course. While ample research studies on the empirical qualities of testing, namely, validity and reliability are commonly carried out on language tests, it rarely happens that the impact, justice and fairness of such tests are adequately considered as aspects of test design in their own right. Conventionally, test impact, justice and fairness are almost always implicitly dealt with through investigations of validity and reliability. While it cannot be denied that the latter test properties have a direct bearing on the former and that that the latter possibly cannot be considered without touching on the former, it is important to acknowledge that test impact, justice and fairness are not the same as validity and reliability and that they therefore cannot be replaced by them.

They need to be seen as principles of test design on their own. Using Kunnan's (2004) framework of test fairness in collaboration with the one advanced by Weideman (2009; 2012) on the constitutive and regulative requirements of test design to evaluate the impact, justice and fairness of TALL, this study reveals that, from the point of view of test bias, access, administration, and social consequences, TALL is a reasonably just and fair test of academic literacy, the use of whose scores for decision making should not have a negative impact on those involved.

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.

Davies, A. 1990. *Principles of language testing.* Cambridge: Basil Blackwell. pp. X-X.

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

Erford, B. T. 2013. *Assessment for counselors.* USA: Brooks/Cole Cengage Learning.

Fulcher, G. & Davidson, F. 2008. Tests in life and learning: A deathly dialogue. *Educational philosophy and theory* 40(3): 407 – 417.

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

Jensen, A. R. 1980. *Bias in mental testing.* New York: Free Press.

Kaplan, R. M. & Saccuzo, D. P. 1997. *Psychological testing: Principles, applications, and issues.* Boston: Cole Publishing Company

Kunnan, A. J. 2000. Fairness and justice for all. In: Kunnan, A. J. (Eds.) 2000. *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium*, Orlando, Florida. Cambridge: University of Cambridge Local Examinations Syndicate. pp. 1-14.

Kunnan, A. 2004. Test fairness. In: Milanovic, M. & Weir, C. (Eds.) 2004. *Studies in language testing*. Cambridge: Cambridge University Press. pp. 27 – 45.

Le, P.L., Du Plessis, 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.

McNamara, T. 2004. Language testing. In: Davies A. & Elder C. (Eds.) 2004. *The handbook of applied linguistics.* Malden: Blackwell Publishing. pp. 763-783.

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

Mdepa W. & Tshiwula L. 2012. Widening participation and lifelong learning. *Special Issue*, 13: 19 -33.

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

Rambiritch A. 2012a. Transparency, accessibility and accountability as regulative conditions for a postgraduate test of academic literacy. Unpublished Ph.D Thesis. University of the Free State.

Rambiritch A. 2012b. Challenging Messick: Proposing a theoretical framework for understanding fundamental concepts in language testing. *Journal for Language* Teaching 46(2): 108-127.

Van der Slik., F. 2008. Gender bias and gender differences in tests of academic literacy. *Southern African Linguistics and Applied Language Studies* 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 der Walt, J. L. & Steyn, H. S. (Jnr.) 2007. Pragmatic validation of a test of academic literacy at tertiary level. *Ensovoort* 11(2): 138-153.

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

Van Dyk, T. 2006. Test of Academic Literacy Levels: Standard procedures for test administration: Unpublished Manual: Pretoria.

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

Weideman, A. 2006. Transparency and accountability in applied linguistics. *Southern African linguistics and applied language studies* 24(1): 71-86

Weideman, A. 2007. A responsible agenda for applied linguistics: Confessions of a philosopher. *Per Linguam* 23 (2): 29-53.

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

Weideman, A. 2011. Academic literacy tests: Design, development, piloting and refinement. *Journal for Language Teaching* 45 (2): 100-113.

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

Whiston, S. C. 2013. *Principles and applications of assessment in counseling.* United States: Brooks/Cole Cengage Learning.

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 testing.