Frans van der Slik and **Albert Weideman**

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?

ABSTRACT Responsible test design relies on close examination of a number of parameters of a test. After finding a clearly argued, rational basis (construct) for the ability being tested, then articulating this in detailed specifications for subtests and item types, and subsequently setting benchmarks for both test reliability and item productivity, there remains, after the results become available, a number of further dimensions of a test that need attention. This article examines one such dimension: that of Differential Item Functioning (DIF), asking whether there is, in the case of the test under consideration, bias towards a certain group of test-takers (testees), so that they are unfairly disadvantaged by some of the items or task types in the test. The test results across four different years (2005-2008) of a large group of first year students, the bulk of the intake at one South African university, are analysed. The fact that there are variations in DIF across the different years and across different task types (subtests) calls for specific explanations. The findings suggest that one would do well to examine test results in depth, in order to avoid conclusions that may be fashionable but inaccurate. However, the argument returns to the defensibility of the test construct, and what should legitimately be included in that, and, by extension, measured.

Keywords: test design, subtests, item types, Differential Item Functioning (DIF), bias, test results, defensibility, measurement

1. Tests must be theoretically defensible and socially accountable

Many recent studies (e.g. McNamara & Roever, 2006; Shohamy, 2001a, 2001b, 2004) have alerted test developers to the political, social and cultural prejudices that may unwittingly be built into the tests of language ability that we design. Test developers cannot today claim to be ignorant of how, historically, tests have sometimes served to reify cultural and social biases. The context of this article is therefore the broader case that must be made not only for the theoretical justification of language test design, but also for the social accountability that attaches to each responsibly designed test.

In a number of studies and analyses of the *Test of Academic Literacy Levels (TALL)* and its Afrikaans counterpart, the *Toets van Akademiese Geletterdheidsvlakke (TAG)* (Van der Slik & Weideman, 2005, 2007, 2008, 2009; Weideman, 2003, 2009; Weideman & Van der Slik, 2008), we have described how, in terms of empirical analyses and defensible theoretical assumptions and argument, language tests such as these might be rationally justified, how they might be refined, and how their developers might become socially accountable. In addition, Van der Slik (2008) has specifically examined gender bias in these tests. Similarly, the question of the equivalence of tests across different years has also been addressed (Van der Slik & Weideman, 2007). Furthermore, a groundbreaking validation study of the test (Van der Walt & Steyn, 2007), that could potentially serve as a model for others to do similar studies, has also contributed to the detailed analyses and information currently available about these tests, that are now administered to some 31000 students annually on the various campuses of the four universities that have been using them over the past six years. This collaboration among the University of Pretoria, North-West University, the University of the Free State and Stellenbosch University has recently been formalised into a partnership, the Inter-Institutional Centre for Language Development and Assessment (ICELDA).

We have undertaken the analyses referred to above for two main reasons. First, the tests affect the lives of ever larger numbers of students: their results are for the most part used to channel first year students into appropriate academic literacy interventions, but in the case of one institution even higher stakes decisions relating to access are in part based on them. Second, we believe that complacency remains the number one enemy of responsible test development. Responsible test design relies on close adherence to a number of design principles (Weideman, 2009), and an equally close examination of a number of parameters of a test.

The first step in responsible test design remains finding a clearly argued, rational basis (construct) for the ability being tested (cf. Van Dyk & Weideman, 2004a), then articulating this in detailed specifications for subtests and item types, and finally setting benchmarks for both test reliability and item productivity (Van Dyk & Weideman, 2004b) to guide the piloting process. After this, there remain, once the tests have been administered and the results become available, a number of further dimensions that need attention. How should the results be interpreted? Since no test is perfect, how can those who have potentially been negatively affected by the level of inconsistency of the measurement be given a second chance? What advice should be given to the administrators who wish to use the test results to enable them to treat fairly and with care everyone whose ability has been measured?

This article examines one such further dimension: that of Differential Item Functioning (DIF), asking whether there is, in the case of the test under consideration, bias towards a certain group of test-takers (testees), so that they are unfairly disadvantaged in comparison with others by some of the items or task types in a test. It is only one of a number of measures to ensure the fairness of test results, but a highly specific and important one, that we thus wish to draw attention to here.

2. Research questions

In this article we will therefore address several research questions: 1) To what extent does *TALL* display Differential Item Functioning (DIF)? 2) If DIF is identified, can it be linked to the specific content of the item? 3) To what extent do English first language students and students who have one of the African languages as a first language perform differently on the tests and their constituting subtests and items? 4) If the analyses reveal significant heterogeneity, and if the topic or content does not explain the observed differences, is there another, more plausible explanation? 5) What lessons can be learnt from the outcomes?

3. Method

3.1 Population and context

In January and February of every year between 2005 and 2008, the academic literacy of virtually all new undergraduate students of the University of Pretoria, the Potchefstroom and Vaal Triangle campuses of North-West University, and Stellenbosch University was tested through the administration of the *Test of Academic Literacy Levels (TALL/TAG)*. At two of these institutions, students are allowed to sit for either the English (*TALL*) or Afrikaans test (*TAG*), and so have the freedom to choose whichever language they feel more comfortable with in the academic environment.

Students are also invited to provide background information on their gender and their first (or home) language. A preliminary check of these data reveals that, besides students who have Afrikaans and English as their first language, one of the participating institutions attracts substantially more students whose first language is an African language, than do the others. Since virtually all of these students take the English version of the tests, we will restrict ourselves to the data of those who wrote *TALL* at that institution.

3.2 Description of the sample

Information made available by the student administration of the institution whose data are being used for the analysis shows that the main first year students' African languages in 2004 were Xitsonga (2%), IsiZulu (3%), Sesotho (3%), Setswana (5%), and Sesotho Sa Leboa (7%). These proportions would have remained more or less steady, and the Sotho-speaking group (at 15% of total enrolment in 2004) would have remained the majority, in subsequent years. In total, 15,192 students participated (cf. Van der Slik, 2008 for a detailed description). Students whose first language is unknown or who have another language than those mentioned before (Portuguese or French, for example) were excluded from the analyses. This applies to 1,428 students, so the results of 13,764 candidates are available for further analyses. Though everyone in the total sample wrote *TALL*, the English version of the academic literacy test, by choice, the sample is therefore made up not only of those who have English (n = 7430) as a first language, but also of a group who have an African language (n = 4400) or Afrikaans (n = 1934) as a first language.

3.3 TALL *and its design*

The 2005 and 2006 versions of *TALL* consist of 120 scoring marks, distributed over seven subtests or sections (described in Van Dyk & Weideman 2004a; 2004b, Weideman 2006), six of which are in multiple-choice format:

Section 1: Scrambled text (ST)
Section 2: Interpreting graphs and visual information (GVI)
Section 3: Understanding texts (UT)
Section 4: Academic vocabulary (AV)
Section 5: Text type (TT)
Section 6: Text editing (TE), later renamed to Grammar & text relations (GTR), which will be the label used subsequently in this article
Section 7: Writing (handwritten; marked and scored only for certain borderline cases)

The 2007 and 2008 versions of *TALL* and *TAG* each consist of 100 scoring points, distributed over the first six subtests or sections mentioned above, all of which are in multiple choice format. Section 7 (20 scoring marks) was omitted from 2007 on; borderline cases, who are identified by statistical means, are allowed to take another test, the results of which are used to decide if the students indeed have risk associated with their level of academic literacy, or have adequate levels of academic literacy.

Students have a limited time (60 minutes) to complete the test, and they earn a maximum of 100 marks (some items count 2 or 3 marks instead of 1). The time restriction to complete the test has been deliberately chosen by the test designers. It is considered to be one way of distinguishing between academically literate and academically less literate students. In addition, it has to be emphasized that for the same reason the most difficult part of the test has also purposely been placed at the end of the test. We return to a discussion of this design decision below, as well as in the conclusion.

3.4 Analyses

Two kinds of analyses were performed. T-tests were used to check if students who have English, Afrikaans and African languages as their first languages performed differently on the total tests (a Bonferroni adjustment was made for the number of comparisons). In addition, DIF analyses were performed by means of the Mantel-Haenszel statistic in the TiaPlus package (CITO, 2005). The Mantel-Haenszel DIF statistic is calculated by first partitioning students into different subsamples by language. Those whose first language is one of the African languages thus form the first group, while the language of the second group is Afrikaans, and that of the third English. Then, four different subgroups of approximately the same size according to their levels in mean ability scores on the entire test are made up for each language subsample. Finally, the ratio of the odds of success of the different language groups is calculated and the averages of these ratios across each score level are determined. DIF values in the 0-1 interval imply that the item is more difficult for the group in the first subsample. DIF values around

1 imply that the test item has approximately equal difficulty across subsamples or language groups, and DIF values greater than unity mean that the item is more difficult for the group in the second subsample. Z-scores are used to check significance. We would like to emphasize, however, that due to the large number of testees used as a total sample, even small differences may turn out to be significant.

4. Results

4.1 Discussion

Table 1 shows the mean scores broken down by first language for *TALL*, while Table 2 shows the T-values for the differences between the three first languages involved.

It is quite obvious that in the years 2005-2008 students whose first language is one of the African languages performed significantly worse than those whose first language is English or Afrikaans. In most cases, the Afrikaans speaking students did not differ significantly from the students whose first language was English. Interestingly, however, the trend is in the direction of Afrikaans speaking students outperforming the English speaking students, and in one case – the 2007 test – this difference is significant. One possible explanation for this perhaps counter-intuitive result might be that Afrikaans speaking students are for the most part from formerly privileged socioeconomic backgrounds, while the English speaking students come from more heterogeneous socioeconomic backgrounds, but we have no real evidence that this is so. Another, more plausible explanation may be that a good number of those Afrikaans first language students who voluntarily chose to write the test in English might have felt more comfortable with English as an academic language because they finished secondary school in English, an enduring trend among middle class pupils (not only among Afrikaans, but also African language speakers). A third, related reason might be that those Afrikaans first language students who chose to write the test in English might have been pressurised when they enrolled by their faculty administration to write the test in English. At this institution, it is well known that the faculty that attracts the top students requires them to write the English test. So the Afrikaans students who wrote the English test may very well be among the top students of the intake of a particular year. The observed trend of English speaking students gradually being outperformed by Afrikaans speaking students nevertheless has to be kept in mind, because it is in line with observations yet to be made.

Also, a general descending trend in average scores can be observed over the years. However, this does not necessarily mean that the average academic literacy has decreased over the years, since the trend might also be explained by increasing difficulty levels of the test (cf. Van der Slik & Weideman, 2007). Future research should resolve this relevant issue more definitely, yet some intriguing findings, seemingly pointing to a gradual decrease in English academic literacy over the years 2005-2008, will be presented in the following tables.

4.2 Differential Item Functioning

By means of TiaPlus (CITO 2005), DIF analyses were performed with the Mantel-Haenszel statistic for the results of *TALL* in 2005, 2006, 2007, and 2008. However, since the mean scores of Afrikaans and English speaking students differ so little, we have decided to present only the Z-scores associated with the DIF statistics of the English and the other African language speaking students (see Table 3).

Table 1: Mean scores on TALL *of first year students who have an African language, English, or Afrikaans as their first language*

Study	African Languages (1)			English (2)			Afrikaans (3)		
	n	*Mean*	*SD*	*n*	*Mean*	*SD*	*n*	*Mean*	*SD*
2005	1050	56.41	19.10	1416	81.41	11.73	473	81.20	11.63
2006	1198	49.35	17.97	1479	74.21	14.32	517	74.48	13.99
2007	1054	44.07	17.76	2127	68.55	17.19	415	71.37	14.88
2008	1098	46.50	16.93	2408	69.03	17.84	529	70.60	15.56

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

Table 3: Z-scores of associated Mantel-Haenszel DIF statistics for English and African language speaking students

Items[1]	2005	2006	2007	2008
ST 2	–2.62**	--	--	--
AV 1	--	–2.75**	--	--
AV 2	--	–2.83**	--	--
AV 4	--	--	–2.66**	--
AV 6	--	–3.15**	--	--
AV 8	--	–3.38***	--	--
TT 1	--	--	–2.73**	--
TT 3	--	–3.83***	--	--
UT 3	--	–2.62**	--	--
UT 6	–5.75***	–2.64**	--	--
UT 7	–5.43***	--	--	--
UT 8	–3.86***	--	--	--
UT 9	–4.45***	--	--	--
UT 10	–3.02**	--	--	--
UT 12	–3.56***	–3.93***	--	--
UT 14	–2.69**	--	--	--
GTR 1	--	–3.08**	--	--
GTR 2	–3.63***	--	--	--
GTR 3	–2.68**	–3.44***	--	--

GTR 4	–4.17***	–3.51***	--	--
GTR 5	--	–3.64***	--	--
GTR 6	--	–4.36***	--	--
GTR 8	–4.15***	–4.09***	--	--
GTR 9	–4.97***	–4.24***	--	--
GTR 10	–4.17***	–3.25**	--	--
GTR 11	–4.20***	--	--	--
GTR 12	–4.70***	–4.54***	--	--
GTR 13	–4.75***	–3.37***	--	--
GTR 14	–5.28***	–3.61***	--	--
GTR 15	–4.67***	–5.10***	–3.04**	--
GTR 16	n/a	--	–2.66**	--
GTR 17	n/a	–4.06***	--	--
GTR 18	n/a	–4.04***	--	--

Note: ** p < .01; *** p <.001; negative values imply students whose first language is English outperform students whose first language is an African language; n/a: not applicable; --: not significant

[1]: Item number and acronym of the subtest (see above, the section on *TALL* and its design)

Some quite interesting conclusions can be drawn from the results presented in Table 3. First, if DIF occurs, it flags that English speaking students perform better than students whose first language is an African language when total test performance is considered. Second, for some yet unexplained reason, DIF disappears rather abruptly as years go by. In Table 3, it should be noted, we have presented the Z-scores associated with those items in which DIF has indeed been indicated. These Z-scores are measures of the level of significance of the DIF occurring. In fact no significant DIF occurred in 2008, whereas in 2007 only four items displayed DIF. We return to this peculiar outcome at the end of this section when we discuss Figure 2. Third, it seems that what used to be called the Text editing (TE) subtest, and what is now entitled the Grammar & text relations (GTR) subtest, is particularly susceptible to DIF, at least in 2005 and 2006. Remember, however, that the candidates had only a limited amount of time to complete the test and that this subtest traditionally is the final one to complete. Perhaps less academically literate students were unable to complete the test, or due to time pressure made more guesses at the end than the more academically literate students. So, the occurrence of DIF might not be the result of item bias of some sort, discriminating against students whose first language is not English, but may be the result of purposely introduced time restrictions.

There are at least two additional observations that support such an interpretation. When one takes a closer look at the DIF statistics for the 2005 test, it seems that in addition to the Grammar & text relations (GTR) subtest, the Understanding texts (UT) subtest is rather susceptible to DIF, while in 2006 this applies to the Academic vocabulary (AV) subtest.

This is peculiar. Why would a test exhibiting DIF in 2005 not do so in 2006, and the other way around? There appears to be no other reason for this than that in 2005 the UT subtest preceded the GTR subtest, while in 2006 the AV subtest preceded the GTR subtest almost immediately. So an obvious explanation seems to be that in both cases these subtests were situated at or close to the end of the test and that DIF may have been occasioned by time restrictions, and not by the (presumably biased) content of the items. Note in this regard, as well, that the final items of the subtest are more susceptible to DIF than the first ones. So, again time restrictions to complete the test seem to be largely if not solely responsible for the occurrence of DIF.

A final observation that supports such a conclusion is the following. When one examines the test scores of the students, it is immediately apparent that many of them were unable to complete the test, since they left many questions unanswered at the end. The outcomes presented so far, however, have been calculated in such a way that missing answers have been interpreted as incorrect answers. When one repeats the analyses presented here, but now excludes these missing answers, it results in a picture that is rather different from the one presented above. It shows instead that *all* DIF statistics display a reduction in strength. Moreover, when missing answers are excluded rather than counted as incorrect answers it transpires that in 2005, 14 instead of 19 items displayed DIF, in 2006 7 instead of 22, and in 2007, none instead of 4. Figure 1 illustrates why DIF decreases when missing answers are excluded from the analyses.

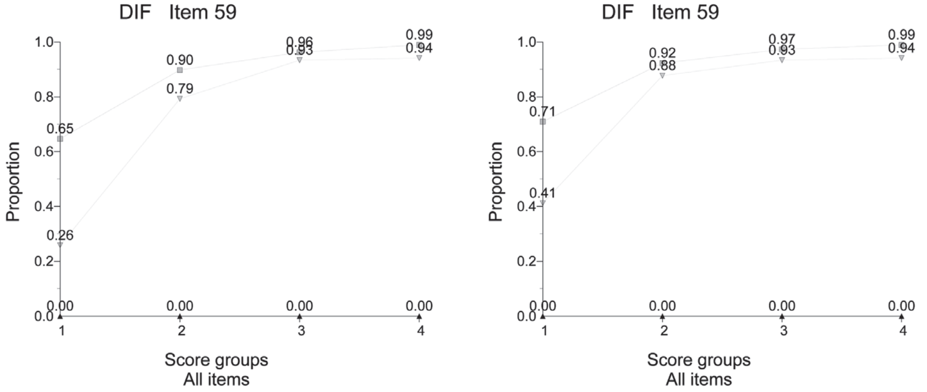

Figure 1: Grammar & text relations (GTR) item number 14 (Question 59, TALL 2005) if missing answers are counted as incorrect (left panel) and if they are excluded from the analyses (right panel).

Figure 1 can be read as follows. The TIAPLUS package has divided the testees into four score groups. Score group 1 (1) contains the 25% lowest scoring testees on all the 60 items of *TALL*. Score group 4 (4) consists of those 25% who scored highest, while score group 2 and 3 fall in between. The top line (whose scores are generally higher than those of the bottom line) represents those with English as first language. The bottom line represents those whose first language is an African language, and whose scores are generally lower. It seems quite obvious from comparing the left and the right panel in Figure 1 that a substantial number of the incorrect answers in the lowest scoring group is due, however, to some of them having given no answer at all. In case these missing answers are included and counted as incorrect, only 26% of the lowest scoring group answered this item correctly (see Figure 1, left panel). If, however, the missing answers are excluded (Figure 1, right panel), 41% of the lowest scoring group answered this item correctly, and accordingly the Z-score associated with the Mantel-Haenszel DIF statistic is reduced from Z = –5.28 (see Table 3) to Z = –3.61.

Although we have presented here only the outcomes for GTR item 14 in 2005, the general picture that emerges from all items displaying DIF in 2005, 2006, and 2007 is in accordance with Figure 1. That is: DIF decreases when missing answers are excluded from the analyses. Several observations could be made when inspecting Figure 1, but we will restrict ourselves to – in our view – the most salient one: DIF seems to occur almost exclusively because the 25% of those who have an African language as a first language who achieved the lowest score (score group 1), performed much lower than the 25% lowest achieving English speaking candidates. However, the remaining scoring groups of those with an African language as first language, on the one hand, and the English first language group, on the other, are almost similar to each other regarding their average scores. What does this mean? In our view it signifies that the items are doing exactly what they were supposed to do, i.e. making a distinction between less and more academically literate students. The primary reason why the lowest achieving African first language students score lower than their lowest achieving English first language counterparts seems to be that they more than others either gave no answer at all or were guessing to a larger degree. The validity of this conclusion is strengthened by the observation that not content but rank order in the test is associated with the occurrence of DIF.

But why are there decreases in the number of items displaying DIF over the years and why do we not observe DIF at all in 2008? Not answering the final items of the test was even more prevalent in 2008 than it was in 2005, 2006, and 2007. A good 29% of the candidates gave no answer for the final items of the test in 2008, while only 14% did not do so in 2005. We think that Figure 2 gives at least a provisional answer. In Figure 2 we present the pre-final item of the Grammar & text relations subtest, in the same way as we did in Figure 1, i.e. with and without the missing answers included.

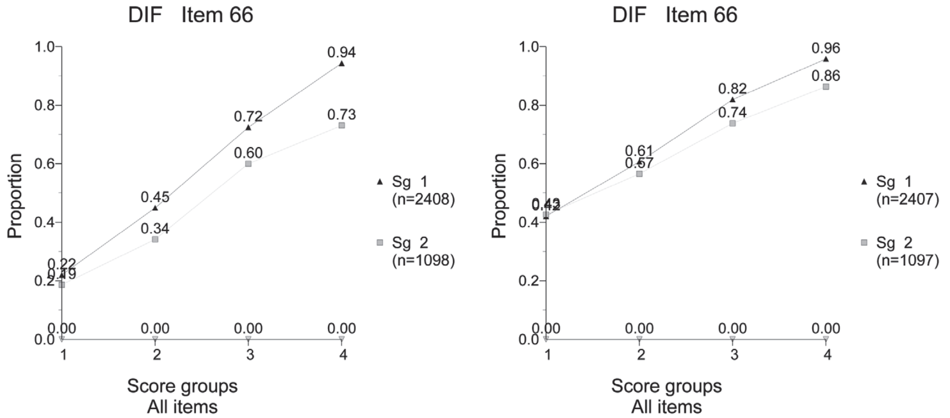

Figure 2: Grammar & text relations number 17 (2008) if missing answers are counted as incorrect (left panel) and if they are excluded from the analyses (right panel).

The top line again indicates those with English as first language, and the bottom line those with an African language as first language. It can be added that the picture that emerges from Figure 2 is representative of the remaining items of *TALL* 2008. When Figure 2 is compared to Figure 1, at least three observations can be made. First, the various scoring groups from either the English or African language group behave in much the same way, which, of course, is to be expected since no significant DIF was found. Second, the lower the achievement of a scoring group is, the more it refrained from answering. The lowest scoring groups in Figure 2, for example, answered GTR item 17 correctly in around 20% of the cases, when missing answers were included in the calculation (Figure 2, left panel). However, their proportion correct answers doubled to .43 when missing answers were excluded from the calculations. Third, and perhaps most importantly, it seems that in 2008 a substantially larger number of students whose first language was recorded as English are indicated as being less academically literate, compared to the numbers in 2005. This provides at least a partial explanation why DIF occurs less and less frequently as compared to the outcomes displayed in Figure 1. The finding that the proportion of missing answers has more than doubled since 2005, finally points into the same direction.

5. Conclusion

Rather than signalling item bias, i.e. pointing to undeservedly discriminating against the lowest achieving students who have an African language against the lowest achieving students whose first language is English, DIF flags that *TALL* is doing validly what it is supposed to do: distinguishing less academically literate students from more literate ones. The primary reason for the occurrence of DIF is not the biased content of the 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 also less able to complete than those who can more competently and fluently handle the demands of cognitive processing and language associated with tertiary education. So the question that is raised in the title of this article can be answered with an emphatic "no".

Should a time limit be imposed on such tests? There are cases of other South African tests of academic literacy that we know of where there is virtually no time limit on the completion of the test, and where testees are given anywhere between 2:5 and 3 hours to complete a test that is usually slightly longer in terms of number of items than *TALL*, which has a time limit of an hour. The declared purpose of such tests, however, has always been to determine whether those from disadvantaged backgrounds have the potential to deal with tasks that are challenging from the point of view of the language to be employed academically by them, something that TALL does not set out to do, since "potential" is such a theoretically problematic construct. One may argue that such a lack of time limit might be appropriate for that purpose (though, given the analysis here, we are now more sceptical about that assumption, which has never been tested empirically), but that, in the case of *TALL* a limitation on time is entirely appropriate, and is theoretically defensible. One's fluency in and ability to handle academic discourse has to be part of one's level of academic literacy.

We should also note that the time limitations set in the case of *TALL* were not randomly determined, but the outcome of careful piloting. Several adjustments were made, and are in fact still being made, during the piloting process of each of the tests in question. The limitations, in other words, are justifiably related to what is being measured by this test of academic literacy.

Hunter and Schmidt (2000:151; cf. too Koch & Dornbrack, 2008) present a strong case as to why professionally developed tests such as *TALL* may be free from item bias. Despite allegations from certain quarters that cognitive ability and educational achievement tests are predictively biased, especially against cultural groups such as minorities, research has consistently shown that there is no bias at the level of total test scores. They (2000:153) find the hypothesis that individual items in such tests are biased to be inconsistent with the findings of large testing organisations. Rather, findings of biased items can be attributed to statistical or mathematical errors resulting from statistical procedures that are not founded on substantive theory. Kline (2004: 559) emphasises that measurement bias should be made at the test level, since this is the level at which decisions about individual persons are made. She advises using caution before simply removing or revising items that show DIF, a view not shared by De Beer (2004:53) who advocates the identification and elimination of DIF for the purposes of improved test construction.

In their study on bias in cross-cultural assessment, Van de Vijver and Poortinga (1997:30) maintain that bias is not considered to be an inherent property of an assessment instrument, but that inferences drawn from scores can become biased. They (1997:33) caution that valid intergroup differences in average scores can be confounded with bias, particularly where the cultural distance between two groups is large.

We should thus be wary of using DIF analyses to jump to a fashionable conclusion. As Lumley and Brown (2005:838) point out, item bias or DIF exists only where the differences in test scores are unrelated to the test construct, the ability being tested. In the case of *TALL* the results indicate that English speaking students perform better in terms of total test performance than do students whose first language is an African language. The difference in performance can be attributed in part to the time restrictions imposed on testees and the rank order of test types. A further contributing factor may be related to the fact that the English first language group was representative of heterogeneous socioeconomic backgrounds, including persons from formerly privileged socioeconomic backgrounds. We are of the opinion that the differences noted in the analyses of TALL are related entirely to what is being measured and that the test design is justifiable and fair both from a theoretical perspective, as well as in respect of its social accountability.

Acknowledgements

We are indebted to Colleen du Plessis, our research assistant, who unearthed a number of further studies that shed light on a topic that, we believe, is often misunderstood.

REFERENCES

CITO. 2005. *TiaPlus, Classical Test and Item Analysis©*. Arnhem: Cito M. & R. Department.

De Beer, M. 2004. Use of Differential Item Functioning (DIF) analysis for bias analysis in test construction. *SA Journal of Industrial Psychology*, 30(4): 52-58.

Geldenhuys, J. 2007. Test efficiency and utility: Longer and shorter tests. *Ensovoort*, 11(2): 71-82.

Hunter, J.E. & Schmidt, F.L. 2000. Racial and gender bias in ability and achievement tests. *Psychology, Public Policy, and Law*, 6(1): 151-158.

Kline, T.J.B. 2004. Gender and language differences on the Test of Workplace Essential Skills: Using overall mean scores and item-level Differential Item Functioning analyses. *Educational and Psychological Measurement*, 64(3): 549-559.

Koch, E. & Dornbrack, J. 2008. The use of language criteria for admission to higher education in South Africa: issues of bias and fairness investigated. *Southern African Linguistics and Applied Language Studies*, 26(3): 333-350.

Lumley, T. & Brown, A. 2005. Research methods in language testing. In: Hinkel, E. (ed.). *Handbook of research in second language teaching and learning*. Mahwah, New Jersey: Lawrence Erlbaum Associates: 833-855.

McNamara T. & Roever C. 2006. *Language testing: The social dimension*. Oxford: Blackwell.

Shohamy E. 2001a. Fairness in language testing. In: Elder, C., Brown, A, Grove, E., Hill, K., Iwashita, N., Lumley, T., McNamara, T. & O'Loughlin, K. (eds.). *Experimenting with uncertainty: Essays in honour of Alan Davies*. Cambridge: Cambridge University Press: 15-19.

Shohamy E. 2001b. *The power of tests: A critical perspective on the uses of language tests*. Harlow: Pearson Education.

Shohamy E. 2004. Assessment in multicultural societies: Applying democratic principles and practices to language testing. In: Norton, B. & Toohey, K. (eds.). *Critical pedagogies and language learning*. Cambridge: Cambridge University Press: 72-92.

Van der Slik, F. 2008. Gender bias and gender differences in two tests of academic literacy. *Southern African Linguistics and Applied Language Studies special issue: Assessing and developing academic literacy* [Geldenhuys, J. (ed).] 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. 2007. Testing academic literacy over time: Is the academic literacy of first year students deteriorating? *Ensovoort*, 11(2): 126-137.

Van der Slik, F. & Weideman, A. 2008. Measures of improvement in academic literacy. *Southern African Linguistics and Applied Language Studies*, 26(3): 363-378.

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* special issue: Assessing and developing academic literacy [Geldenhuys, J. (ed.)] 27(3): 253-263.

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

Van de Vijver, F.J.R. & Poortinga, Y.H. 1997. Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, 13(1): 29-37.

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

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

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

Weideman, A. 2006. Assessing academic literacy in a task-based approach. *Language matters*, 37(1): 81-101.

Weideman, A. 2009. Constitutive and regulative conditions for the assessment of academic literacy. *Southern African Linguistics and Applied Language Studies* special issue: Assessing and developing academic literacy [Geldenhuys, J. (ed.)] 27(3): 235-251.

Weideman, A. & Van der Slik, F. 2008. The stability of test design: Measuring difference in performance across several administrations of a test of academic literacy. *Acta academica*, 40(1): 161-182.

ABOUT THE AUTHORS

Frans van der Slik
Research Associate in the Department of English, UFS
Radboud University
Nijmegen
Netherlands

Albert Weideman
Department of English
University of the Free State
Bloemfontein
9300
Email: albert.weideman@ufs.ac.za