Southern African Linguistics and Applied Language Studies 2009, 27(3): 277–290
Printed in South Africa — All rights reserved

SOUTHERN AFRICAN LINGUISTICS AND APPLIED LANGUAGE STUDIES
ISSN 1607–3614 EISSN 1727–9461
DOI:

Gender bias and gender differences in two South African tests of academic literacy

Frans van der Slik
Department of English, University of the Free State, PO Box 339, Bloemfontein 9300, South Africa;
Department of Linguistics, Radboud University, PO Box 9103, 6500 HD Nijmegen, The Netherlands
e-mail: f.v.d.slik@let.ru.nl

Abstract: How much empirical evidence is there of gender bias in the Test of Academic Literacy Levels (TALL) and its Afrikaans counterpart, Toets van Akademiese Geletterdheidsvlakke (TAG)? This paper reports on meta-analyses of 24 academic tests, containing data of more than 60 000 first-year South African students. TALL appears to be gender-neutral, while for TAG some evidence was found that men slightly outperform women. This outcome can be attributed almost entirely to male students outperforming female students on the subtest requiring the interpretation of graphs and visual information. Women outperform men slightly on the subtest dealing with text types. These outcomes are presented in light of discussions on gender differences versus gender bias.

Introduction

Freedom from systematic error remains a central concern in achievement or ability testing. Without it, a test and the decisions taken on the basis of the results it yields cannot be considered fair:

> Fair decisions *are those that are equally appropriate, regardless of individual test takers' group membership. We need to ask whether the decision procedures and criteria are applied uniformly to all groups of test takers.* (Bachman & Palmer, 1996: 32f, emphases in the original)

Systematic error occurs when testing conditions or rating procedures are not equal for all candidates or when groups of candidates are treated differently. A test is said to be gender biased if, for example, women are placed at a disadvantage against men, or vice versa.

The current paper examines the question of possible gender bias in assessing academic literacy over a period of time at North West University (NWU), the University of Pretoria (UP) and Stellenbosch University (SU). It belongs to a series of investigations that have been done to determine how the Test of Academic Literacy Levels (TALL) and its Afrikaans counterpart, Toets van Akademiese Geletterdheidsvlakke (TAG), meet various criteria for language tests in general, and academic literacy tests in particular (Weideman, 2003a; 2006a; Van Dyk & Weideman, 2004a; 2004b; 2006b; Van der Slik & Weideman, 2005; 2007; 2008; Van der Walt & Steyn, 2007; Weideman & Van der Slik, 2008). The tests are designed and used only for placement purposes, i.e. to determine what level of academic literacy support is required (if at all) *after* the student has gained access, and are therefore not high-stakes, but medium- to low-stakes tests. For students at NWU and UP, the effects of the results of TALL and TAG are limited to submitting to a compulsory academic literacy course (Weideman, 2003b), while limited action is taken for underachieving students from SU. The compulsory courses are intended to assist students in eliminating one of the factors most closely associated with lack of academic success and performance (Van Rensburg & Weideman, 2002). From 2008 onwards, the purpose of the test at the SU has changed rather drastically. Instead of being used as a low-stakes test with no social or political consequences, it has been upgraded to a high-stakes test, since it is now part of SU's admission test battery.

The question of gender bias in testing is awkward, and occasionally even a rather inconvenient one. Suppose it is found that women on average obtain higher scores on a test of academic literacy than men. Does this necessarily imply that these are artificial outcomes of a gender-

biased test? Men are perhaps no less academically literate than women, but the test may have put them at a disadvantage. Or could it be that women are generally speaking indeed more academically literate than men, and does the test merely reveal that difference in a valid way? The female/male ratio at universities in, for example, the United States (Coley, 2001), the Netherlands (Statline, 2008) and South Africa (current paper) seem indeed to point to the latter. The issue can even be pushed a bit further. Suppose no gender differences are found regarding academic literacy. Does this, then, imply that the test is gender-neutral? Even this is not necessarily the case: the test may mask real existing differences by discriminating against women who on average may be academically more literate than men (cf. Shealy & Stout, 1993; Stout *et al.*, 2003). Clearly, more elaborate methods have to be used to formulate at least provisional answers to these questions. Fortunately, there is a wealth of literature regarding both gender bias as well as gender-based differences. A brief review is presented below. A cautionary note seems to be appropriate, however, since a distinction between gender bias and gender-based differences is difficult to make (Goldstein, 1993).

Review of the literature

Differential Item Functioning

One way of testing the occurrence of test bias is to perform Differential Item Functioning (DIF) analyses (Holland & Wainer, 1993; McNamara & Roever, 2006). Items exhibit DIF when testees with different background characteristics (such as gender, or cultural, social or linguistic) differ in their probability of answering these items correctly, after controlling for ability (Camilli & Shepard, 1994), or, formulated more accurately, overall test performance (Abbott, 2006). DIF analyses reveal if candidates that are matched on overall test performance nevertheless score differently because they belong to different groups. In that case, inappropriate interpretations of the test outcomes might occur, which result in discrimination against a specific group. Gierl *et al.* (1999: 15) have emphasised, however, that the occurrence of DIF does not necessarily point to item or test bias: 'If … the performance difference can be attributed to actual knowledge and experience differences the test is designed to measure, then the outcome can be interpreted as item impact.' For this reason, items exhibiting DIF are carefully scrutinised by test developers or applied linguists, and eventually eliminated from the test if it is concluded that the item in question does not tap the construct it is intended to measure. An example from Dorans and Kulick (1983) illustrates how gender bias might operate. Students were presented the following analogical reasoning item to complete in the Scholastic Assessment Test (SAT):

Decoy: Duck:

A) net: butterfly B) web: spider C) lure: fish D) lasso: rope E) detour: shortcut

Dorans and Kulick found this item to be more difficult for female candidates than for male candidates when overall test performance was controlled for, and they attributed this to gender differences in background knowledge regarding hunting and fishing – two traditionally male-oriented recreational activities (Dorans & Kulick, 1983). If this is indeed the case, then the item measures more than it is intended to measure and in that case it is advisable to remove it from the test battery. Unfortunately, it is not always as clear why individual items exhibit DIF as in the Decoy:Duck example. In fact, in more than half of the cases, the items with large DIF are found to be uninterpretable (see Gierl *et al.,* 1999). This poses a huge challenge to test developers, because eliminating all items that show DIF without knowing the exact cause of its occurrence, as some authors (Van de Vijver & Leung, 1997; De Beer, 2004; Hambleton, 2005) advocate, may adversely affect the validity of the construct being measured for the reasons Gierl *et al.* (1999: 15) have stated. Moreover, tests are seldom one-dimensional in the sense that they capture just one attribute and that fact alone may result in the occurrence of DIF. Routinely eliminating all items that demonstrate DIF may thus mask overall differences between the sexes regarding attributes that are considered of central importance with respect to the construct being measured (Goldstein, 1993).

Gender bias

Because the explanation of the occurrence of DIF for individual items may prove to be less successful than hoped for, researchers have searched for patterns across bundles of items rather than individual items. The rationale is that individual items may exhibit rather weak but systematic DIF which will stay unnoticed in a single item approach, but, taken together as a bundle, DIF is amplified (Nandakumar, 1993). As such, Differential Bundle Functioning (DFB) analysis is regarded as a natural extension of DIF analysis (Boughton *et al.*, 2001). As noted before, the occurrence of DFB or DIF does not necessarily flag item bias. As Shealy and Stout (1993) have noted, test items are intended to measure a primary (ability or achievement) dimension. Items that produce DIF are said to measure more than this primary dimension, and are assumed to measure at least one additional dimension. These dimensions are referred to as secondary dimensions. They can be characterised as reflecting impact if the constituting items are included intentionally, because, in that case, they measure part of the construct of the test. These constituting items reflect item bias, however, if the secondary dimension slipped in unintentionally (Gierl *et al.,* 2001). Using the procedures developed by Stout and his colleagues (2003), several authors have found evidence for gender bias.

Gierl *et al.* (1999), for example, evaluated the 1997 and 1998 Canadian Grade 6 Science Achievement tests and found content topic differences between girls and boys. Boys performed better on understanding the concepts and processes of science when air and aerodynamic topics were involved; while girls did better when observation and inference topics were involved. Boughton *et al.* (2001) also found gender differences across content areas for the 1991 and 1992 Alberta Social Studies Grade 12 diploma examinations. The exams consisted of 70 multiple-choice items, which tested understanding, terminology, recall of concepts, and synthesis and analysis. Source materials on a variety of topics such as maps, graphs, and political and economics texts were provided. Female candidates performed better on economic theory items, while the bundles of items on history and control tactics were favoured more by male candidates. Because the male-oriented topics were overrepresented compared to the female-oriented topics, female candidates scored consistently lower on the Social Studies exams than male students. The validity of these outcomes may be questioned, however, since women demonstrate 'a 0.1 to 5.2 percentage point advantage, in school awarded marks' (Boughton *et al.,* 2001: 6). These results partly provide support for the theory developed by Walter and Young (1997) that males perform better on this test because the constituting items are more masculine in character or consist of items that predominantly present a male perspective. It should, therefore, not come as a great surprise that males outperformed females. The relationship between item content and performance has been found repeatedly (see Buck *et al.*, 1997). Bügel and Buunk (1996) have found that in tests of foreign language text comprehension, men scored significantly higher on 'male' topics, (i.e. topics dealing with the economy, politics, crime, sports and technology); while women scored significantly higher on 'female' topics (topics concerning, for example, human relations, female professions, self-care, household and art). These authors provide evidence that gender differences could partly be explained by differences in reading habits and prior knowledge.

The research outcomes reviewed above indicate that the content of a test in terms of the topics that are treated may affect the scores of female and male candidates differently. It seems advisable, therefore, to include, if possible, only items on gender-neutral topics in a test of academic literacy, or at least to match male-oriented topics with feminine topics (Maccoby & Jacklin, 1974).

Such an approach is no guarantee, however, that the test outcomes of male and female testees do not differ. Other factors than topic content may also affect test scores of men and women differently. Walstad and Robson (1997), for example, found that the response format of the questions may also have an impact on test results. In line with previous research (see Willingham & Cole, 1997; Boughton *et al.*, 2001), they reported that female candidates perform worse on multiple-choice tests. Lumsden and Scott (1987) found that women do better on essay tests. Such findings may point to differences in cognitive functioning between men and women. In such a scenario, a test that displays no differences between men and women may mask existing gender differences because the effect of an overrepresentation of female-oriented topics is neutralised by the response

format of the test. Other researchers have, therefore, dug somewhat deeper and have searched for gender-based differences – i.e. differences that cannot be attributed to unfair treatment by a test *per se*, but to differences in cognitive functioning.

Gender differences

On the premise that tests may exhibit gender differences in various degrees, depending on a variety of causes (such as gender bias) that may interact with each other in complex ways, researchers have tried to neutralise these variations by adopting a meta-analytic approach. In meta-analysis the results of studies that address the same research hypothesis are combined. The main purpose is to find an average effect size over these studies and to control for individual study characteristics. The outcome can be considered as a more powerful estimation of the true population effect size than the outcome of a single study. Macoby and Jacklin (1974: 26), for example, performed a meta-analysis and identified 85 studies that reported females outperforming males on verbal ability. Their conclusion was that girls 'do better on tests of grammar, spelling, and word fluency'.

Maccoby and Jacklin's study was severely criticised by Hyde and Linn (1988) for its lack of methodological sophistication. Maccoby and Jacklin used a simple vote counting procedure, i.e. counting significant results in the predicted direction against remaining results. This method can lead to false conclusions if the reviewed studies lack statistical power. Hyde and Linn looked at 58 vocabulary studies and used more elaborate meta-analytical techniques. They did not find a significant gender effect, but they did find a sizeable heterogeneity in effect sizes, implying that the studies can hardly be seen as replications of each other. Such a variation in effect sizes might at least partially be the result of a variation in topic content – a possibility that was elaborated upon in the previous section. In general, women fared slightly better in reading, writing, speaking and general verbal ability than men; but, according to Hyde and Linn, these differences were too small to account for, and they concluded that the textbooks on gender differences have to be rewritten.

This bold conclusion has proved to be premature, however. Cole (1997) and Willingham and Cole (1997) have concluded in an even larger study with more than four million students that women have retained their language advantage over the past 30 years. Women fare much better in writing and language use (grammatical conventions, expression, spelling, etc.), while small effect sizes are detected for reading and vocabulary reasoning. These outcomes have been found to be largely independent of socio-cultural background. Lin and Wu (2003), for example, found in a report on Chinese candidates who did the English Proficiency Test, that the bundle of listening comprehension items favoured females; while the bundles of grammar, vocabulary and cloze items favoured male candidates slightly. Pae (2004), using a huge sample of Korean English Foreign Language (EFL) learners, found that items expressing Mood/Impression/Tone tended to be easier for females; whereas items classified as Logical Inference were more likely to favour males, regardless of item content. Bügel and Buunk (1996) found evidence that women generally fare better than men on reading abilities in narrative text types (cf. Bielinski & Davison, 1998).

It has to be emphasised that gender differences have also been found in a variety of ability tests other than language-use tests. Men have been found to score higher on mathematical reasoning tests, while women fare better on computational tests. In addition, men outperform women on spatial tests – particularly those of mental rotation (Hamilton, 1988; Willingham & Cole, 1997; Kimura, 1999). Such findings might help to explain why language-test items that simultaneously involve visual representation and reading comprehension might be easier for men, because such tasks are cognitively rather complex – an ability in which men have been found to outperform women (Engelhard, 1990).

Gender bias versus gender differences

The reasons for the occurrence of gender differences may be manifold. Though few researchers today point exclusively to a single cause, some authors are primarily interested in biological or genetic differences, while others give prevalence to socio-cultural differences (for reviews see Maccoby, 1966; Caplan *et al.,* 1997; Kimura, 1999). Kimura (1999), for example, offers biological or genetic explanations and provides ample evidence that differences between male and female

cognitive functioning can be explained (though not exclusively) by different hormonal configurations. The production of male sex hormones from early childhood in boys is assumed to be critical in this respect. As a result, a masculinisation of behaviour and cognition occurs, leading to a variety of differences between males and females in, for example, motor skills, spatial abilities, mathematical aptitude, perception, and verbal abilities. Yet, other researchers dismiss genetic or biological explanations as primary explanations for differences between the sexes and point to the socio-cultural environment and the way labour and child care are organised as primary causes for gender differences in various abilities such as mathematics, science and language (Caplan *et al*., 1997). Other authors take a less extreme stand in this nature-nurture debate (Maccoby, 1966; Halpern *et al.,* 2007) and emphasise that genetic, biological or hormonal causes interact in a complex way with causes that stem from the socio-cultural environment people live in.

Research questions

In this paper, I will address several research questions: (1) To what extent do the tests for academic literacy levels display DIF? (2) Can DIF be linked to the specific content of the item? (3) To what extent do male and female candidates perform differently on the tests and their constituting subtests? (4) If the analyses reveal significant heterogeneity in test impact, can topic content explain the observed differences?

Method

Population and context

In January and February of 2005[1], 2006, 2007 and 2008, the academic literacy of virtually all new undergraduate students of the University of Pretoria, the Potchefstroom and Vanderbijlpark campuses of North-West University, and the University of Stellenbosch were tested through the administration of the Test of Academic Literacy Levels (TALL/TAG). At the University of Pretoria and North-West University, students are allowed to sit for either the English (TALL) or Afrikaans (TAG) test, and so have the freedom of choosing whichever language they feel more comfortable with in the academic environment. At the University of Stellenbosch, however, students have to take both tests. At this university, the English test was administered one day after the Afrikaans test. As stated in the Introduction, Stellenbosch University has decided to use the academic literacy tests, from 2008 onwards, as part of an admissions test battery, and the 2007 tests were used for that purpose in the period from June to September 2008. In total, 64 357 students participated (but they do not necessarily represent *different* students – see the first note to Table 1); 34 604 took the Afrikaans test, while the remaining 29 753 students participated in the English version. See Table 1 for a detailed description.

The TALL and TAG tests and their design

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

- Section 1: Scrambled text (ST)
- Section 2: Understanding graphs and visual information (GVI)
- Section 3: Understanding texts (UT)
- Section 4: Academic vocabulary (AV)
- Section 5: Text types (TT)
- Section 6: Text editing (TE)
- 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, all of which are in multiple-choice format. Section 7 was omitted from 2007 onwards; borderline cases, who are identified by statistical means, are allowed to take another test, the results of which are used to decide if the student has passed or failed the test, and has risk in terms of academic language.

Table 1: Number of first-year students at the University of Pretoria (UP), Stellenbosch University (SU) and North-West University (NWU) taking the TALL and TAG tests (2005–2008)

TALL	UP	SU*	NWU	Total
2005	3 310	1 729	134	5 173
2006	3 652	3 710	143	7 505
2007	3 905	4 165	140	8 210
2008**	4 325	4 282	258	8 865
Total	15 192	13 886	675	29 753
TAG	UP	SU*	NWU	Total
2005	2 701	1 702	2 521	6 924
2006	2 547	3 703	2 650	8 900
2007	2 582	4 160	2 707	9 449
2008**	2 333	4 287	2 711	9 331
Total	10 163	13 852	10 589	34 604

* All Stellenbosch students took the TALL the day after they took the TAG

** In 2008, the Stellenbosch students took the 2007 TALL and TAG tests as part of an admission test

Students have 60 minutes to complete the test, and they earn a maximum of 100 marks (some items count for 2 or 3 instead of 1 mark). In total, the 2005, 2006, 2007 and 2008 tests contain 512 items.

It is important to note that there is no selection bias – a methodological flaw that hinders many meta-analytic studies – in the current study, since all tests and test scores were available to the author. In addition, so far none of the tests has been screened on gender bias, and finally, all samples are sufficiently large to reveal differences, if any, and have been drawn from the same population.

Analyses

Three types of analyses were performed. These included *t*-tests to check if female and male students performed differently on the total tests. In addition, DIF analyses were performed by means of the Mantel-Haenszel statistics in the TiaPlus package (CITO, 2005). The Mantel-Haenszel DIF statistic is calculated by first partitioning male and female candidates into subgroups with same total scores on the test. Then the ratio of the odds of success of the females over the odds of success of the males 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 in the first subgroup; DIF values around 1 imply that the test item has approximately equal difficulty across subgroups; and DIF values greater than unity mean that the item is more difficult for the second subgroup. *Z*-scores are used to check significance. Finally, the StatsDirect package (StatsDirect, 2007) was used in order to perform meta-analyses on both the total tests and the six constituting subtests. The Random effects *pooled d+* statistic (DerSimonian & Laird, 1986) was used to test if tests across the years 2005, 2006, 2007 and 2008 and across universities reveal a significant difference between female and male candidates. In addition, the heterogeneity statistic I^2 was used, which measures the proportion of total variation in study estimates that is accounted for by heterogeneity rather than by chance (StatsDirect, 2007). Because the Chi-square test has a low sensitivity in detecting heterogeneity, an α-value of less than 0.10 was used to determine significant heterogeneity. An I^2 value of more than 40% indicates significant heterogeneity.

Results

Description of the sample

Table 2 shows the outcomes at the scale level for TAG for female and male students separately. As can be seen, there is a general trend of lower mean scores for female students as compared to their male counterparts for the three universities included in this study over the years 2005 through 2008. In fact, half of the data show a significant difference, while the other half, though non-significant, are

Table 2: Mean scores of female and male students taking the TAG test

Study	Females			Males			
	n	Mean	SD	*n*	Mean	SD	*t*-value
2005UP	1 450	69.12	13.65	1 205	71.57	13.26	−4.68***
2006UP	1 341	60.02	14.68	1 139	60.64	15.08	−1.03
2007UP	1 383	55.89	15.59	1 134	57.69	15.14	−2.93**
2008UP	1 227	54.53	14.28	1 073	57.32	14.12	−4.71***
2005NWU	1 384	63.15	14.86	1 032	63.41	14.95	−0.42
2006NWU	1 485	54.17	15.25	1 027	54.22	15.18	−0.08
2007NWU	1 570	50.72	15.24	1 066	52.01	15.83	−2.08*
2008NWU	1 558	47.46	14.74	1 145	50.31	15.07	−4.91***
2005SU	970	62.35	19.90	708	64.84	18.58	−2.63**
2006SU	1 898	54.07	18.24	1 747	53.11	17.70	1.62
2007SU	2 210	51.81	18.79	1 879	52.24	18.76	−0.73
2008SU#	2 187	50.88	19.01	2 100	50.07	19.31	1.38

Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Test is identical to the 2007 test

in the same direction with two exceptions. Interestingly, however, the female/male ratio exceeds unity without exception. In addition, there is no indication of greater test score variation among male students.

Table 3 shows the mean scores broken down by gender for TALL. The general trend observed for TAG was not replicated very closely for TALL. Male students performed significantly better than female students only at Stellenbosch University in 2005, 2006 and 2008. The latter outcome is somewhat problematic since the 2008 tests at Stellenbosch University are identical to the 2007 tests. Perhaps the differences in student populations can account for these outcomes. Remember that the 2007 students at Stellenbosch University took the test when they were already admitted to the university, while the 2008 students took the test when they had not yet been admitted. Though not significantly, the trend of males outperforming females seems to be reversed for North-West University and for the University of Pretoria in 2005 and 2008. Very few students at North-West University opted for the English version of the test, so it seems that these students' characteristics differed from those at the University of Pretoria and Stellenbosch University. No gender-related trend in test score variability could be observed for TALL either. Finally, the general trend was again that more females than males decided to study at the tertiary level.

Differential Item Functioning

By means of TiaPlus (CITO, 2005), DIF analyses were performed with the Mantel-Haenszel statistic over the 2005, 2006, 2007 and 2008 academic literacy tests, TAG and TALL. Since the tests were the same for the three universities involved (except in 2008), it was expected that identical items would display DIF for the three universities. For the most part, this turned out to be true, as can be seen in Tables 4 and 5.

The general picture that emerges from Tables 4 and 5 has two primary aspects. First, if DIF occurs, it flags that, almost without exception, male students perform better than female first-year students when total test performance is accounted for. Second, and in accordance with outcomes of previous research, it seems that the subtest on understanding graphs and visual information (GVI) is particularly susceptible to DIF. This holds for the Afrikaans version of the test for the year 2006 and 2007 in particular; but the 2007 English version of the academic literacy levels test also contains a substantial number of DIF-exhibiting GVI items. It was already noted that, in 2008, the 2007 academic literacy tests were administered to prospective Stellenbosch University students as part of an admission test. Virtually the same items that were susceptible to DIF in 2007 also flag DIF in 2008.

One possible explanation for this may be that the content or theme of the items rather than some unmeasured sample characteristic is responsible for the occurrence of DIF. But is it? The

Table 3: Mean scores of female and male students on TALL

Study	Females			Males			
	n	Mean	SD	*n*	Mean	SD	*t*-value
2005UP	1 798	72.12	18.43	1 337	71.66	20.21	0.65
2006UP	2 077	64.00	19.43	1 460	65.25	20.47	−1.83
2007UP	2 175	60.98	20.08	1 653	61.40	21.28	−0.62
2008UP	2 446	62.97	19.28	1 808	62.48	21.02	0.77
2005NWU	55	63.69	20.70	62	60.02	22.12	0.93
2006NWU	85	59.13	17.46	57	52.84	20.09	1.93
2007NWU	73	53.64	20.71	66	46.95	22.12	1.83
2008NWU	133	56.70	21.11	120	52.21	18.82	1.79
2005SU	969	75.78	14.72	685	78.92	13.62	−4.47***
2006SU	1 901	67.06	16.57	1 752	69.95	16.34	−5.31***
2007SU	2 212	64.91	16.78	1 881	65.67	16.44	−1.46
2008SU[#]	2 183	65.46	16.58	2 099	66.98	16.85	−2.97**

Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

[#] Test is identical to the 2007 test

2007 GVI subtest of TAG is about global daily oil consumption – a technical topic that is perhaps more appealing to male students than it is to female students. Yet, in that case one would expect the 2006 GVI subtest of TALL to exhibit DIF as well, since the topic of this subtest is global oil production. Nothing of this sort can be observed, however. This subtest does not result in DIF. On the other hand, the 2006 GVI subtest of TAG is on patterns of spending in the household; a topic that may (stereotypically) have been expected to have greater appeal to female than male students. Again, the outcomes of the DIF analyses contradict such an expectation; controlling for overall performance, this subtest was easier for male than for female students. The 2007 GVI subtest of TALL deals with population growth of metropolitan areas across the globe. Though this topic appears to be gender neutral, the DIF analyses again contradict such a hypothesis. Male students outperform female students. And finally, the TALL 2005 GVI subtest and the TAG 2008 GVI subtests are both on pension provision in certain countries consuming ever larger amounts of money (measured as proportion of GDP). With the preceding outcomes in mind, one would perhaps expect male students to outperform female students. That turns out to be a wrong guess again: no DIF is observed. The general conclusion that can be derived from the preceding outcomes is thus that content or theme does not appear to make any difference. No lessons can in such a case be learned from this line of argument, since the occurrence of DIF appears to be entirely unpredictable. The only conclusion that can be derived is that, if DIF occurs on the GVI subtests, it generally favours men over women.

To push the argument a bit further, I have performed more detailed analyses on the subtest level. If the GVI-subtests are scrutinised separately, just **two** GVI-items display DIF. And they did this for TAG 2005 and for University of Pretoria students only. GVI-item 22 favours male students, while GVI-item 10 (not detected previously as displaying DIF, by the way) favours **female** students. The occurrence of DIF may thus be attributed entirely to chance. Hence, these more detailed analyses demonstrate rather nicely that in the present study the occurrence of DIF depends almost entirely on the rest of the test and, thus, appear to reflect gender differences rather than gender bias.

Meta-analyses[2]

StatsDirect (2007) was used to perform meta-analyses on the TAG and TALL data. These analyses were conducted on the scores of the total tests and also on the scores of the constituting subtests. As the 2008 tests of Stellenbosch students were identical to the 2007 tests of academic literacy, the 2008 test scores for Stellenbosch University were not included. A possible threat to the validity of the outcomes of a meta-analysis is the incompatibility of the test scores over the years. Fortunately, this threat appears to be rather minor. Though the length of the entire test and of its constituting

Table 4: *Z*-scores of associated Mantel-Haenszel DIF statistics for TAG

TAG 2005	UP (*n* = 2 655)	SU (*n* =1 1678)	NWU (*n* = 2 416)
11 AV#	−3.80***	−0.85	−3.04**
18 GVI	−3.53***	−2.83**	−3.91***
22 GVI	−4.28***	−2.41	−4.17***
33 UT	−3.98***	−2.79**	−5.49***
59 TE	−3.26***	−1.50	−2.25
TAG 2006	(*n* = 2 480)	(*n* = 3 645)	(*n* = 2 512)
6 GVI	−2.97**	−3.73***	−2.23
8 GVI	−3.99***	−3.69***	−2.49
9 GVI	−3.50***	−3.76***	−2.93**
34 AV	−2.80**	−2.31	−2.75**
38 AV	−2.43	−0.64	−2.74**
TAG 2007	(*n* = 2 517)	(*n* = 4 089)	(*n* = 2 636)
7 GVI	−2.96***	−2.46	−3.29***
8 GVI	−4.30***	−2.45	−3.88***
9 GVI	−3.35***	−3.17**	−3.92***
10 GVI	−4.08***	−4.89***	−3.51***
11 GVI	−5.64***	−5.07***	−4.82***
12 GVI	−4.00***	−3.38***	−3.66***
42 AV	2.12	2.22	2.97**
48 AV	2.19	3.14**	1.28
TAG 2008	(*n* = 2 300)	(*n* = 4 287)	(*n* = 2 703)
11 GVI	−2.97**	n.a.	−1.63
27 UT	−6.46***	n.a.	−6.23***
36 UT	2.11	n.a.	2.76**
44 AV	−2.19	n.a.	−3.18**
7 GVI (2007)##	n.a.	−2.86**	n.a.
8 GVI (2007)	n.a.	−5.21***	n.a.
9 GVI (2007)	n.a.	−3.45***	n.a.
10 GVI (2007)	n.a.	−4.39***	n.a.
11 GVI (2007)	n.a.	−5.14***	n.a.
12 GVI (2007)	n.a.	−3.72***	n.a.
48 TT (2007)	n.a.	2.62**	n.a.

Note: ** $p < 0.01$; *** $p < 0.001$; negative values imply men outperforming women; n.a. = not applicable

Item number and acronym of the subtest (see p.281)

At SU the 2007 version of TAG was used in 2008

subtests varies slightly over the years, the weights of the items may also vary; but they do so in such a way that the relative weight of the subtest within the entire test remains more or less constant. In addition, total test scores have remained constant over the years (100 points). In short, the scores on the literacy tests could be compared in a straightforward way (cf. the discussion on test equivalence in Van der Slik & Weideman, 2007; Weideman & Van der Slik, 2008).

The package provides a pooled effect size measure *d*, and its associated 95% confidence interval. It also provides an inconsistency measure I^2 which can be used to test the uniformity of the effect sizes. It can thus be used as a replication measure. A large I^2 indicates that the effect sizes *d* vary considerably across tests, indicating that tests are poor replications of each other.

Regarding TAG (see Table 6), a significant though weak effect size was found which indicated that male students, on average, slightly outperformed female students ($d+ = -0.08$ (95% CI = −0.10 to −0.05). It was also found that the effect size variation was substantial: $I^2 = 80.2\%$ ($p < 0.001$). Table 7 shows that only a very weak, though significant, difference between average male student and female student performance could be detected across the TALL tests (Pooled $d+ = -0.05$) (95% CI = −0.05 to −0.02). However, again, considerable effect size variation was found between the eleven tests for academic literacy ($I^2 = 81.1\%$; $p < 0.001$).

Table 5: *Z*-scores of associated Mantel-Haenszel DIF statistics for TALL

TALL 2005	UP (*n* = 3 135)	SU (*n* = 1 654)	NWU (*n* = 117)
22 TT[#]	2.87**	1.83	0.86
29 UT	−2.87**	−0.78	−0.60
50 TE	−4.04***	−3.60***	−0.42
TALL 2006	(*n* = 3 537)	(*n* = 3 653)	(*n* = 142)
22 UT	−2.58**	–1.85	−0.81
41 AV	−1.81	−2.89**	−0.58
TALL 2007	(*n* = 3 828)	(*n* = 4 093)	(*n* = 139)
6 GVI	−3.69***	−2.81**	−0.65
7 GVI	−3.57***	−3.32***	−0.89
8 GVI	−2.76**	−4.58***	−0.72
11 GVI	−2.78**	−1.57	−0.20
12 GVI	−3.52***	−3.91***	−0.43
25 UT	2.34	2.62**	0.15
TALL 2008	(*n* = 4 254)	(*n* = 4 282)	(*n* = 253)
9 GVI	−3.67***	n.a.	−1.01
10 GVI	−4.56***	n.a.	−1.17
20 AV	−2.67**	n.a.	−0.07
6 GVI (2007) [##]	n.a.	−4.60***	n.a.
7 GVI (2007)	n.a.	−4.44***	n.a.
8 GVI (2007)	n.a.	−4.61***	n.a.
12 GVI (2007)	n.a.	−2.93**	n.a.
25 UT (2007)	n.a.	3.15**	n.a.
37 AV (2007)	n.a.	−2.78**	n.a.
50 TE (2007)	n.a.	2.68**	n.a.

Note: ** $p < 0.01$; *** $p < 0.001$; negative values imply men outperforming women; n.a. = not applicable

[#] Item number and acronym of the subtest (see p.281)

[##] At SU the 2007 version of TALL was used in 2008

Perhaps the most important conclusion is that the tests of academic literacy levels TAG and TALL are hardly affected by gender differences. I will return to this in the final section. The effect sizes are significantly heterogeneous, however, and it therefore seems a good idea to check if an analysis of the subtests will result in more homogeneous effect sizes.

Tables 6 and 7 capture the meta-analytic statistics for the TAG and TALL subtests, respectively, and show that some subtests are largely responsible for the occurrence of heterogeneity. For TAG, the subtest on understanding graphic and visual information displays the largest amount of variation in effect sizes. Additional analyses have revealed that the effect sizes for GVI vary from $d = -0.74$ (UP 2007) to $d = -0.32$ (NWU 2006). Despite this large amount of variation, it is evident that female testees consistently perform worse on the GVI subtest than their male counterparts. This does not only apply to the Afrikaans version of the test, but also – though less pronounced – to the English version of the academic literacy test. Yet, other subtests also display large amounts of effect size variation, particularly the English versions of the Text Type and Text Editing, and the Afrikaans version of the Understanding Text subtests. The individual outcomes for the Text Type subtests (see note 2) indicate that – though all positive – the effect sizes vary substantially. Regarding the Understanding Text subtest of TAG, which also shows a large amount of effect size variation, it was found that the individual effect sizes display an inconsistent pattern from $d = -0.19$ (UP 2008) to $d = 0.09$ (NWU 2005). Male students fare slightly but consistently better on Academic Vocabulary than female students – an outcome that again cannot be explained by the content of the items.

Regarding the pooled effect sizes, it can be concluded that female first-year students performed worse than male first-year students on the GVI subtests. In accordance with previous research, new

Table 6: Meta-analysis – pooled *d*+ and I^2 statistics for TAG and its six subtests

Subtest	TAG	
	Pooled *d*+	I^2
Scrambled text (ST)	0.01	46.2%*
Understanding graphs and visual information (GVI)	−0.44***	90.3%***
Understanding texts (UT)	−0.02	84.4%**
Academic vocabulary (AV)	−0.12***	60.0%**
Text types (TT)	0.06**	65.0%**
Text editing (TE)	0.01	70.6%***
TAG (total test)	−0.08***	80.2%***

Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 7: Meta-analysis – pooled *d*+ and I^2 statistics for TALL and its six subtests

Subtest	TALL	
	Pooled *d*+	I^2
Scrambled text (ST)	−0.01	77.5%***
Understanding graphs and visual information (GVI)	−0.28***	48.4%*
Understanding texts (UT)	−0.00	73.8%***
Academic vocabulary (AV)	−0.04**	46.8%*
Text types (TT)	0.04***	86.2%***
Text editing (TE)	−0.00	84.7%***
TALL (total test)	−0.05***	81.1%***

Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

female students perform relatively better than their male counterparts on the Text Type subtests of both TAG and TALL.

Conclusion

Perhaps the most important outcome of this study is that both the English and the Afrikaans versions of the test for academic literacy levels are affected to only a minor degree by gender differences. The majority of the constituting subtests do not display significant differences between male and female testees, and the total tests also exhibit only very small differences between new female and male students. In addition, the vast majority of test items do not display significant DIF, which strengthens the conclusion that gender bias is rather weak. In fact, a mere one out of 13 items display DIF – a figure that is well below the general international finding of one out of three (cf. De Beer, 2004). There are, nevertheless, some observations that merit further attention.

First, the DIF analyses show that the subtest on understanding GVI is particularly sensitive to item bias. This conclusion is prompted by the outcomes of the Mantel-Haenszel analyses of the 2007 tests and the 2008 Stellenbosch outcomes. In future research, test developers should, therefore, be aware of the fact that some subtests may be more vulnerable to gender bias than others. On the other hand, it has to be emphasised that gender bias cannot entirely account for the occurrence of differences between female and male testees, since this gender difference is found rather consistently over the years; i.e. not just in 2007, as was demonstrated by the meta-analytic approach. The occurrence of these differences at least to some extent seems to validly represent existing differences between male and female cognitive functioning. On the other hand, it seems that the finding of superior female performance on the subtest Text type underlines exactly the same reasoning; i.e. that genuine differences between men and women exist, and in some areas women perform cognitively better than men, and vice versa.

The current research has also demonstrated the value of a meta-analytic approach. If only the 2007 data for the Afrikaans test were submitted to a DIF analysis, for example, it may well have

concluded – entirely in line with previous studies, but nevertheless erroneously – that topic or content is responsible for male students outperforming female students. The resolution of the issue of gender bias versus gender difference may indeed sometimes resemble the story of Baron von Münchhausen, who claimed to have escaped from a swamp by pulling himself (and his horse) up by his own hair. We should remind ourselves that both gender bias and gender differences can only be detected by means of tests, which makes it extremely difficult to disentangle bias from real, existing differences. The use of meta-analyses has the virtue of putting research findings into a broader perspective, thereby enabling the detection of often subtle though persistent differences. Answering the question as to whether these differences are either inherited or have a socio-cultural base has not been a goal of this study and, thus, remains an unresolved issue.

Second, the effect size for gender differences was found to be rather small. Several authors (Burnett, 1986; Johnson & Meade, 1987; Willingham & Cole, 1997) have emphasised, however, that even small effect sizes (usually detected by means of Cohen's *d*'s) can potentially have rather substantial public consequences – particularly when the variation in test scores is greater for male than for female candidates, as has repeatedly been attested in a number of studies. In that case, the female/male ratio at the upper end of the test scale may be 1:4. If a test is used for admission purposes, then female candidates might be selected to a much smaller degree than male candidates, even when the mean differences between males and females tend to be rather small. In the current study, no differences in the variation in test scores for male and female testees were found.

Third, the tests for academic literacy are in multiple-choice format. This might imply that the observation of male students performing slightly better than female students might be neutralised or even reversed if different scoring formats were employed; since women are generally found to perform less well on the answering format.

This brings me to my final comment. It has now become apparent that the outcomes of a test depend on more than the abilities of individual testees alone. This observation calls for a responsible decision on the part of test developers. They must ask themselves what kinds of tasks have to be included in a test for, say, academic literacy. Some parts will be easier for men than for women, or vice versa, and this may affect the outcomes – a fact which demands careful consideration of which cognitive tasks have to be part of the concept of academic literacy, and which degree of importance has to be ascribed to them.

In this study I have analysed the effects of gender. Given the complex cultural and linguistic situation in South Africa, it seems appropriate to extend this research to cultural groups in future research (cf. De Beer, 2004).

Notes

[1] For Stellenbosch University, the 2005 test was a trial run and not all faculties participated in TAG and TALL.

[2] An exhaustive report of the StatsDirect meta-analyses can be retrieved at: http://www.let.ru.nl/~f.v.d.slik/index_bestanden/MetaTAG_ALL.rtf and http://www.let.ru.nl/~f.v.d.slik/index_bestanden/MetaTALL_ALL.rtf.

References

Abbott ML. 2006. ESL reading strategies: differences in Arabic and Mandarin speaker test performance. *Language Learning* **56**(4): 633–670.

Bachman LF & Palmer AS. 1996. *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press.

Bielinski J & Davison ML. 1998. Gender differences by item difficulty interactions in multiple-choice mathematics Items. *American Educational Research Journal* **35**(3): 455–476.

Boughton KA, Dawber TE & Hellsten L-AM. 2001. Differential bundle functioning on social studies high school certification exams. Paper presented at the Annual Meeting of the American Educational Research Association (AERA), Seattle, 10–14 April.

Buck G, Kostin L & Morgan R. 2002. *Examining the Relationship of Content to Gender-Based*

Performance Differences in Advanced Placement Exams (College Board Research Report RR-02-25). New York: College Entrance Examination Board.

Bügel K & Buunk BP. 1996. Sex differences in foreign language text comprehension: the role of interests and prior knowledge. *The Modern Language Journal* **80**(i): 15–31.

Burnett SA. 1986. Sex-related differences in spatial ability: Are they trivial? *American Psychologist* **41**: 1012–1014.

Camilli G & Shepard L. 1994. *Methods for Identifying Biased Test Items.* Newbury Park, CA: Sage.

Caplan PJ, Crawford M, Hyde JS & Richardson JTE. (eds) 1997. *Gender Differences in Human Cognition.* New York/Oxford: Oxford University Press.

CITO. 2005. *TiaPlus, Classical Test and Item Analysis*. Arnhem: Cito Measurement and Research Department.

Cole NS. 1997. *The ETS Gender Study: How Females and Males Perform in Educational Setting.* Princeton, NJ: Educational Testing Service.

Coley W. 2001. *Differences in the Gender Gap: Comparisons across Racial/Ethnic Groups in Education and Work.* Princeton, NJ: RTSS.

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

DerSimonian R & Liard N. 1986. Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**: 177–188.

Dorans NJ & Kulick EM. 1983. *Assessing Unexpected Differential Item Performance of Female Candidates on SAT and TSWE Forms Administered in December 1977: An Application of the Standardization Approach* (ETS Technical Report RR-83-9). Princeton, NJ: ETS.

Engelhard G. 1990. Gender differences in performance on mathematics Items: evidence from the United States and Thailand. *Contemporary Educational Psychology* **15:** 13–26.

Gierl MJ, Bisanz J, Bisanz GL, Boughton KA & Khaliq N. 2001. Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice* **20**: 26–36.

Gierl M, Khaliq N & Boughton K. 1999. Gender differential item functioning in mathematics and science: prevalence and policy implications. Paper presented at the annual meeting of the Canadian Society for the Study of Education, Sherbrooke, Quebec.

Goldstein H. 1993. Assessing group differences. *Oxford Review of Education* **19**: 141–150.

Halpern DF, Benbow CP, Geary DC, Gur RC, Hyde JS & Gernsbacher MA. 2007. The science of sex differences in science and mathematics. *Psychology in the Public Interest* **8**(1): 1–51.

Hambleton RK. 2005. Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In Hambleton RK, Merenda PF & Spielberger CD (eds), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment.* Mahwah, NJ: Lawrence Erlbaum, pp 3–38.

Hamilton LS. 1998. Gender differences on high school science achievement tests: do format and content matter? *Educational Evaluation and Policy Analysis* **20**(3): 179–195.

Holland PW & Wainer H. (eds) 1993. *Differential Item Functioning*. Hillsdale, NJ: Erlbaum.

Hyde J & Linn M. 1988. Gender differences in verbal activity: a meta-analysis. *Psychological Bulletin* **104**: 53–69.

Johnson ES & Meade AC. 1987. Developmental patterns of spatial ability. *Child Development* **58**: 725–740.

Kimura D. 1999. *Sex and Cognition.* Cambridge, London: The MIT Press.

Lin J & Wu F. 2003. Differential performance by gender in foreign language testing. Poster for the 2003 annual meeting of NCME, Chicago.

Lumsden KG & Scott A. 1987. The economics student re-examined: male-female differences in comprehension. *Journal of Economic Education* **18**(3): 365–375.

Maccoby EE. (ed.) 1966. *The Development of Sex Differences.* Stanford, CA: Stanford University Press.

Maccoby EE & Jacklin CN. 1974. *The Psychology of Sex Differences*. Stanford, CA: Stanford University Press.

McNamara T & Roever C. 2006. *Language Testing: The Social Dimension*. Oxford: Blackwell.

Nandakumar R. 1993. Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement* **16**: 159–176.

Pae T-I. 2004. Gender effect on reading comprehension with Korean EFL learners. *System* **32**(2): 265–281.

Shealy R & Stout W. 1993. A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika* **58**(2): 159–194.

Statline. 2008. [Online] Available at: http://statline.cbs.nl/StatWeb/Table.asp?STB=G2,G6&LA=nl&DM=SLNL&PA=70943ned&D1=a&D3=0&D4=a&D5=a&D7=a&HDR=T,G5,G4,G1,G3 [Accessed on 27 February 2008].

StatsDirect. 2007. Version 2.6.5. [Online] Available at: http://www.statsdirect.com/help/statsdirect.htm [Accessed on 5 December 2007].

Stout W, Bolt D, Froelich AG, Habing B, Hartz S & Roussos L. 2003. *Development of a SIBTEST Bundle Methodology for Improving Test Equity with Applications for GRE Test Development.* ETS Research Report RR-03-06. Princeton, NJ: Educational Testing Service.

Van de Vijver F & Leung K. 1997. *Methods and Data Analysis for Cross-Cultural Research.* Thousand Oaks: Sage.

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 Walt JL & Steyn HS (jnr). 2007. Pragmatic validation of a test of academic literacy at tertiary level. *Ensovoort* **11**(2): 138–153.

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.

Van Rensburg C & Weideman AJ. 2002. Language proficiency: current strategies, future remedies. *SAALT Journal for Language Teaching* **36**(1&2): 152–164.

Walstad WB & Robson D. 1997. Differential Item Functioning and Male-Female Differences on Multiple-Choice Tests in Economics. *Journal of Economic Education* **28**(2): 155–171.

Walter C & Young B. 1997. Gender bias in Alberta social studies 30 examinations: cause and effect. *Canadian Social Studies* **31**: 83–89.

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

Weideman AJ. 2003b. *Academic Literacy: Prepare to Learn*. Pretoria: Van Schaik.

Weideman A. 2006a. Transparency and accountability in applied linguistics. *Southern African Linguistics and Applied Language Studies* **24**(1): 71–86.

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

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.

Willingham WW & Cole NS. 1997. *Gender and Fair Assessment.* Mahwah, NJ: Lawrence Erlbaum.