Test fairness: a DIF analysis of an L2 vocabulary test

Sauli Takala and **Felianka Kaftandjieva** *University of Jyväskylä, Finland*

The purpose of this study is to analyse gender-uniform differential item functioning (DIF) in a second language (L2) vocabulary test with the tools of item response theory (the separate calibration *t*-method) and to study potential gender impact on the test performance measured by different item composites.

The results of the study show that despite the fact that there are test items with indications of DIF in favour of either females or males, the test as a whole is not gender-biased. In spite of this, it was demonstrated that some item composites are gender-biased. In view of item bank building and use, it means that some of the tests constructed on the basis of an item bank might be biased if the item bank contains items with indication for DIF. Although the results of some empirical research suggest that the requirements for items with DIF to be excluded from the final test version may on the whole be too restrictive, this study demonstrated that the traditional advice of excluding biased items gains new significance in the light of item bank building and use since doing so will prevent possible biased item composites.

I Background

Language use can hardly be seen as a fully uniform repertoire. There are, in fact, a number of variants depending on the context of use and the language user's characteristics (social and geographical origin, education, occupation, age, gender, etc.). A number of studies conducted in various contexts confirm the existence of gender-related differences in verbal ability and language use (Maccoby and Jacklin, 1974; Thorne and Henley, 1975; FUMS, 1977–79; Thorne *et al.*, 1983; Einarsson and Hultman, 1984; Mielikäinen, 1988; Nuolijärvi, 1988; Tannen, 1986; 1990). Although gender differences have been extensively studied, the research findings differ significantly from statements such as 'girls have greater verbal ability' (Maccoby and Jacklin, 1974; Elwood, 1995; Cole, 1997) through 'there are no gender differences in verbal ability' (Hyde and Lynn, 1988) to 'women obtained lower means than men on the verbal scale' (Lynn and

Address for correspondence: Sauli Takala, Professor, Centre for Applied Language Studies, University of Jyväskylä, PO Box 35, FIN-40351 Jyväskylä, Finland; email: sjtakala@cc.jyu.fi

 0265-5322(00)LT185OA

Mulhern, 1991; Lynn and Dai, 1993; Born and Lynn, 1994). The conclusion arrived at by Hyde and Lynn (1988: 62) that 'the gender difference is significantly smaller in more recent studies' was later questioned by Cole (1997: 13) who found that 'females sustained the writing advantage they had from 1960 to 1990'.

Gender studies in vocabulary report similar conflicting results. While Cole (1997) and Hyde and Lynn (1988) report small (not statistically significant) differences in favour of females, other research has found statistical differences in both first-language (L1) and second-language (L2) vocabulary knowledge in favour of males (Boyle, 1987; Lynn and Mulhern, 1991; Lynn and Dai, 1993; Born and Lynn, 1994).

A clue for the explanation of these conflicting results can be found in the comprehensive meta-analytical study conducted by Hyde and Linn (1988). Six of the 56 vocabulary studies included in the meta-analysis found a significant difference in vocabulary in favour of males, and eight reported significant differences in favour of females. Although the meta-analysis shows that there is no significant gender difference in vocabulary, there is significant heterogeneity in the effect size, which means that these studies cannot be considered as replications of each other.

These findings suggest that the differences in methodology might be one of the possible reasons for the inconsistent results of gender studies in vocabulary. Differential item functioning (DIF) is among the factors which might affect test performance in favour of one or another particular group. Although this impact is well known and recognized, a preliminary DIF analysis is not a common practice in gender studies. The fact that the Test of English as a Foreign Language (TOEFL) and First Certificate in English (FCE) do not demonstrate gender DIF (Ryan and Bachman, 1992; Wainer and Lukhele, 1997) does not mean that this is true for all measurement instruments used in gender research.

The connection between gender differences and DIF is two-way. The observed gender differences in some research might be due to the biased estimation of the observed variable but, on the other hand, actual gender differences may lead to gender DIF. Very little is known, however, about this two-way interaction between gender differences and DIF in the L2 vocabulary testing context because, firstly, it has not been widely investigated. The second reason is that the few available research studies either report that there is no gender DIF (Ryan and Bachman, 1992; Wainer and Lukhele, 1997) or the items with significant gender DIF are not discussed from the content point of view (Raju and Drasgow, 1993).

There is another important aspect of DIF analysis which deserves

more detailed consideration. Although the common practice is that the items with indications of DIF are excluded from the final test version in order to prevent test bias, a recent study shows that a test containing DIF items is not inevitably biased (Roznowski and Reith, 1999). Such a result is not surprising since DIF is a necessary but not sufficient condition for item (and test) bias.

On the other hand, if the items with existing DIF are part of an item bank, it is quite possible (at least theoretically) for some of the tests constructed on the basis of this item bank to be biased, due to an inappropriate choice of items. Since both situations are possible – to produce fair or biased tests from an item bank containing DIF items – DIF analysis should not stop at the item level but should go further to investigate how DIF items effect the total test scores based on the possible item composites (Bolt and Stout, 1996).

Taking into account the above considerations, and with a view to item bank construction and use, the purpose of this study, which is part of a bigger project aiming to build an Item Bank for English for the needs of the Finnish Foreign Language Certificate Examination, is to analyse gender DIF in an L2 vocabulary test, and to study potential gender impact on the test performance measured by different item composites.

II Method

1 Instrument

The English vocabulary test that the present study is based on was a part of the test battery used in the Finnish Foreign Language Certificate Examination, Spring 1996, Intermediate Level. It is an official, national high-stakes foreign-language examination based on a bill passed by Parliament. It is available twice a year and covers three bands of proficiency levels (basic, intermediate and advanced). The certificate that test-takers receive is officially recognized in public administration and it is also increasingly used in the private sector.

The test battery consisted of a total of 120 items, and measured all 'four skills' (i.e., listening, speaking, reading and writing) plus grammar and vocabulary, using both selected and constructed response techniques.

The English Vocabulary Test (T_T) analysed here contained 40 multiple-choice items. The tested words had been randomly sampled from a medium-sized dictionary. They were presented in an alphabetical order, and for each English word, four Finnish equivalents were given.

2 Sample

The English examination, Intermediate Level of Spring 1996 was taken by 475 examinees (182 males and 293 females) – a voluntary response sample of adults, who took the examination probably for a number of reasons (personal interest and professional advancement). For the purposes of the analysis, the total sample was divided into sub-samples, as shown in Table 1.

3 Procedure

According to the generally agreed definition, DIF is said to occur when the probability of answering the item correctly is not the same for individuals who are on the same ability level but belong to different groups. Although there are a number of widely applied methods which are not based on item response theory (IRT), this probabilistic definition of DIF makes the IRT approach preferable for DIF analysis from a theoretical point of view (Lord, 1980; Shepard, 1981; Hambleton and Swaminathan, 1985; Crocker and Algina, 1986; Cole and Moss, 1993).

The Rasch model with its simplicity and mathematical elegance, combined with the existence of a sufficient statistic for the ability estimation and statistical tests for model-data fit, has a special place in the set of IRT models. Its main disadvantage – the assumption of equal item discrimination parameters, which prevents its broader use – is overcome in one of its modifications, that is the One Parameter Logistic Model (OPLM). In the OPLM, the discrimination indices are not assumed to be identical, but they are input as known constants. In this way OPLM combines the attractive mathematical properties of the Rasch model with the flexibility of the two-parameter model (Verhelst *et al.*, 1995).

The probabilistic definition of DIF in the IRT context means that DIF occurs when the item characteristic functions for the different sub-groups are not identical. Since item characteristic functions are fully determined by the item parameters, the DIF analysis can be done

Table 1 Breakdown of the sample of those taking the English examination, Intermediate Level of Spring 1996 by category

	Size	Selection	Females	Males
Sub-sample 1	237	Random	136	101
Sub-sample 2	238	Random	157	81
Males	182	Non-random	0	182
Females	293	Non-random	293	0

either by comparison of item parameters (Lord, 1980) or by direct comparison of item characteristic curves by computing the area between them (Hambleton and Swaminathan, 1985).

In the Rasch model and its modifications there is only one parameter, item difficulty, which determines the item characteristic function. This means that the item characteristic functions will be identical for two sub-groups if and only if the item difficulty parameters are identical across the sub-groups. That is why within the framework of Rasch measurement a more common approach to DIF analysis is a comparison of item parameters rather then comparison of item characteristic curves. The separate calibration *t*-test method (Wright and Stone, 1979) is usually used for this comparison. The corresponding *t*-statistics for each item is:

$$t = \frac{(b_F - b_M)}{SQRT(SE_F^2 + SE_M^2)}$$

where b_F and b_M are item difficulty parameter estimates for females and males, based on the separate group calibrations, and SE_F and SE_M are the corresponding standard errors of estimation. The common practice for considering an item as an item with indication of DIF is when $/t/ > 1.96$; this criterion was also applied in this study, although sometimes higher critical values are suggested (Draba, 1977; Smith, 1996).

This method allows the detection of only uniform DIF, which means that there is no interaction between ability level and group membership: the probability of answering an item correctly is greater for one of the two sub-groups for all ability levels and, therefore, the item characteristic curves for both groups do not cross each other. Non-uniform DIF, on the other hand, occurs when the item discrimination parameters are different for the two sub-groups and, consequently, within the framework of the OPLM, the item will not fit the model for one or both sub-groups.

The data used in this study were analysed with the OPLM computer program (Verhelst *et al.*, 1995) and the DIF analysis included:

- model-data fit study;
- comparison of the item difficulty parameters, *b*, estimated separately for females and males;
- comparison of item empirical curves (IEC) for items with indication of DIF;
- comparison of item-fit statistics for females and males;
- comparative analysis of person parameter estimations, based on the total set of vocabulary items (T_T) and the three sub-sets of

items (T_F, T_M, T_N), constructed in such a way that the impact of DIF on the sub-test scores is either minimized or maximized:

1 Sub-Test 1 (T_F), including all vocabulary items (18) that were easier for females than for males;
2 Sub-Test 2 (T_M), including all vocabulary items (22) that were easier for males than for females;
3 Sub-test 3 (T_N), including only the vocabulary items (29) with no indications of DIF.

III Results

1 Model: data fit

The OPLM used in this study fits statistically both the total examination and the vocabulary test, both for the total sample and for the randomly chosen sub-samples as well as for the two sub-samples of females and males referred to in Table 2. Most of the statistical fit indices depend, however, on the sample size and cannot be used as the only proof of model-data fit. That is why an additional analysis of the invariance of item parameters was carried out. Its results (Figure 1) confirm the conclusion about model-data fit for all analysed tests and samples.

As can be seen in Figure 1a, the estimates of the item difficulty parameters for the vocabulary items, based on the whole English Test (all skills) and on the Vocabulary sub-set of items only (T_T), are almost identical. This confirms one of the basic model assumptions about the unidimensionality of the total test (Hambleton and Swaminathan, 1985). In addition, the invariance of item parameters – one of the basic model features of all IRT models – can be observed in

Table 2 Results of analysis of statistical fit

Test	Sample	Size	Fit(p)	Reliability (α)	
				Unweighted scores	Weighted scores*
Total test	Total sample	475	0.0620	0.961	0.955
Vocabulary test	Total sample	475	0.4049	0.807	0.820
Vocabulary test	Sub-sample 1	237	0.1394	0.832	0.841
Vocabulary test	Sub-sample 2	238	0.5523	0.774	0.795
Vocabulary test	Females	293	0.2494	0.815	0.831
Vocabulary test	Males	172	0.3813	0.785	0.793

*In all calibrations, the items are with the same fixed discrimination indices with Geometric Mean equal to 2.423. (In OPLM the discrimination parameter can vary between 1 and 15.)

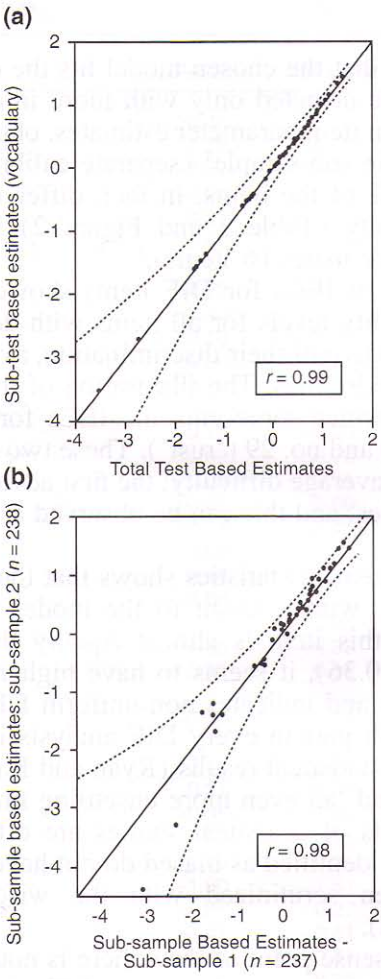

Figure 1 Model: data fit (invariance of item parameters)

Figure 1b. The estimates of the item difficulty parameters for all items (Vocabulary sub-test; T_T), based on two random sub-samples of examinees, follow the identity line ($r = 0.98$) and all are between the 95% quality control lines (Wright and Stone, 1979).

Based on this check of basic model assumptions and model features as well as the statistical fit, we can conclude that the chosen model fits the data and can be applied for supplementary analyses.

2 DIF analysis

In spite of the fact that the chosen model fits the data, DIF may play a role and it can be detected only with more in-depth analysis. The comparison between item parameter estimates, obtained separately for the female and male sub-sample[1] (separate calibration *t*-test), shows that more than 25% of the items, in fact, differ in their functioning in terms of difficulty (Table 3 and Figure 2) in favour of either females (5 items) or males (6 items).

The comparison of IECs for DIF items showed that DIF was in evidence on all ability levels for all items with demonstrated gender differences, irrespective of their discrimination, average difficulty and the direction of gender bias. The illustration of this tendency can be seen in Figure 3, which represents the IECs for two of the items: no. 34 ('turn grey') and no. 29 ('rust'). These two items differ in their discrimination and average difficulty; the first advantaged females and the second one males, and this can be observed at all levels of language proficiency.

The analysis of item-fit statistics shows that there is only one item (no. 24: 'presume') with a misfit to the model in the female sub-sample. Although this item is almost equally difficult for females (0.37) and males (0.36), it seems to have higher discrimination for females than males and indicates non-uniform DIF.

The most difficult part in every DIF analysis is the interpretation and explanation of statistical results (Ryan and Bachman, 1992; Cole and Moss, 1993) and 'an even more unsettling finding, however, has been that the results of statistical studies are often uninterpretable, that is, many items identified as biased do not have any obvious signs of bias even when scrutinized with the wisdom of hindsight' (Shepard, 1981: 96).

From a commonsense point of view there is nothing unusual in the finding that some words may be more difficult for men than for women and vice versa, but the question of why has no easy answer. Word knowledge depends on the frequency of encountering and using particular words. As Kelly (1991: 138–39) noted:

> the degree to which language users, be the language in question the native language or a foreign language, build up this knowledge depends on their experience of the word in question. The wider the variety of contexts and situations in which the item is met and the more often it is encountered, the more surely will the word be progressively mastered.

---

[1]One of the items (no. 25: reality) had to be excluded from these calibrations, because all males answered it correctly; also, very few females – only 4 out of 293 females (1.4%) – gave a wrong answer to this item.

Table 3 DIF analysis

N	Word	$b_{Females}$	b_{Males}	*t*
1	ache[F]	0.42	0.70	−2.97
2	association[M]	0.12	−0.27	2.03
3	commercial	−1.30	−1.61	0.56
4	desk	−0.02	0.12	−0.92
5	dandruff	0.97	1.01	−0.20
6	estate[M]	0.81	0.52	2.43
7	edge[M]	0.14	−0.20	2.14
8	element	−0.43	−0.59	0.58
9	foam	−0.48	−0.35	−0.63
10	form	−3.10	−2.28	−1.36
11	grease[M]	0.50	0.22	2.54
12	gutless	1.29	1.36	−0.66
13	hunting	−3.33	−4.11	0.71
14	income	−0.19	−0.20	0.08
15	immoderate	0.69	0.73	−0.32
16	jelly[F]	−0.09	0.51	−2.85
17	messenger	0.22	0.09	0.89
18	molar	1.44	1.41	0.15
19	mean	0.60	0.74	−1.83
20	numerous	0.46	0.41	0.24
21	oil-well	0.38	0.12	1.88
22	plot[F]	1.01	1.34	−2.81
23	pond	0.86	0.93	−0.59
24	presume[N]	0.37	0.36	0.06
25	reality[A]	not	enough	observations
26	roller	−0.45	−0.72	1.03
27	rookie[M]	0.11	−0.42	2.21
28	riddle	0.85	0.82	0.32
29	rust[M]	0.31	0.01	2.08
30	staff	−0.05	0.09	−0.95
31	seam	0.82	0.71	0.93
32	supply	1.10	0.97	1.11
33	sunday	−1.30	−1.61	0.56
34	turn grey[F]	−0.86	−0.27	−2.35
35	there	−0.47	−0.61	0.63
36	treasure	0.45	0.41	0.41
37	untidy	0.55	0.73	−1.49
38	villager	−1.83	−1.23	−1.66
39	a while ago	−1.71	−1.46	−0.66
40	ward[F]	1.15	1.62	−3.94

Notes: F: Items with DIF ($|t| > 1.96$) in favour of females; M: Items with DIF ($|t| > 1.96$) in favour of males; A: This item was excluded from the analysis due to the lack of variance in the male sub-sample (all males gave a correct answer); N: An item with indication for non-uniform DIF.

A number of studies, however, show that there are gender-based differences in communication patterns, writing style, interests and activities – leisure and professional (Elwood, 1995; Levine and Geldman-Caspar, 1996; Cole, 1997; Edwards, 1998; Hannah and Murachver, 1999). The observed gender differences might lead to

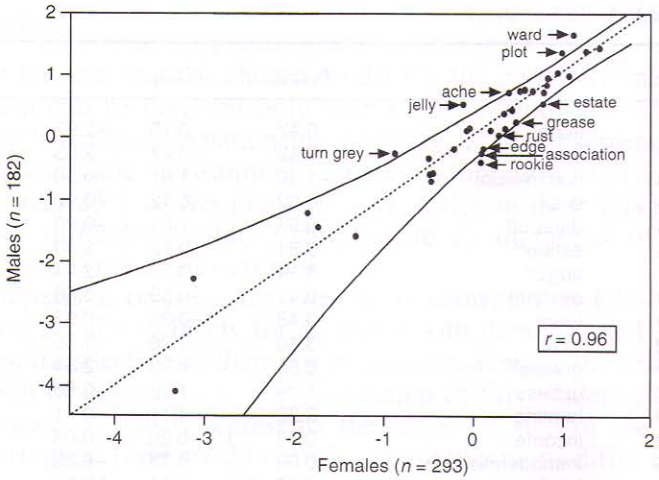

Figure 2 Differential item functioning

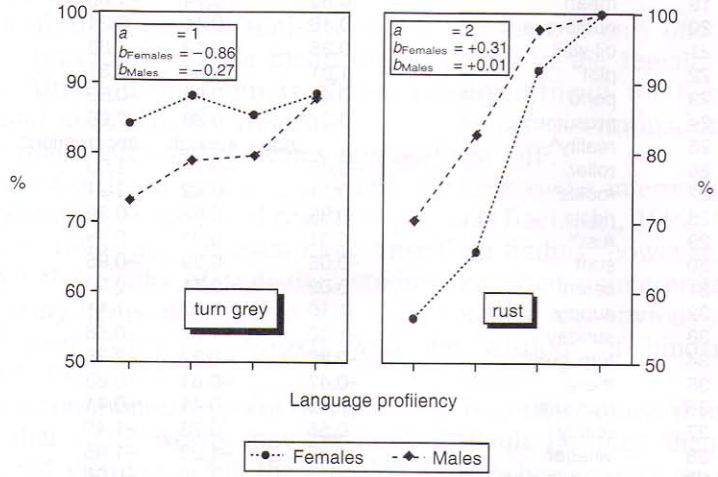

Figure 3 Item empirical curves

gender differences in the frequency of encountering and using particular words.

If males have tried to 'fix something mechanical' more frequently than females (Cole, 1997: 16–17) and are more interested 'in technological and creative aspects' of science, willing to make inventions in 'cars and other vehicles' (Levine and Geldman-Caspar, 1996) then the frequency of encountering and using words like 'grease' and 'rust' can be assumed to be higher for them than for females. (Note that the items with DIF are listed in Appendix 1 in their original format

and with an English translation of the options.) If they are also more involved in sports activities and win more awards in sports in high school than females (Cole, 1997: 16–17) and if they have to spend traditionally between 8 and 11 months of their lives in the army (as in the Finnish context) then they would have encountered more situations where the word 'rookie' is used.

On the other hand, females who more frequently than males reported that they have tried 'to figure out what is wrong with an unhealthy plant, animal' (Cole, 1997: 16–17) – and child, we would add – and who are 'more interested in the humanistic and socially relevant elements of science' (Levine and Geldman-Caspar, 1996) would have more opportunities to encounter and use the word 'ache'. Cooking also is still a more typical activity for women than for men, and usually more females are interested in reading cook books than males. Therefore, it is not too surprising if they know the meaning of the word 'jelly' better than men.

Theoretically, this phenomenon can be explained, extending Gibson's ecological approach to perception (Gibson, 1977; 1986), to cognition in general and to language learning and vocabulary knowledge in particular.

A central notion in the ecological approach is the notion of affordance and, according to Gibson (1977: 67), 'the affordance of anything is a specific combination of properties of its substance and its surfaces taken with respect to an animal' in general. What is especially important is that the affordances which the environment supplies are considered a result of the interaction between the environment and the organism. They are not the object's properties, but object–subject relations or, in other words, 'what people see – in particular, the way in which they parse their environments – is plainly in part a function of interest, desire, need, etc.' (Sanders, 1996: 6). Applying the ecological approach to language education, van Lier (1996) considers linguistic affordances provided by the interaction between the subject and the environment (physical, social, cultural, etc.) as a facilitator of language acquisition.

Human perception of the world unequivocally influences test performance and empirical studies confirm this. Familiarity with, interest in and emotional reaction to the item content were found to be among the possible determinants of gender DIF (O'Neil *et al.*, 1993; Stricker and Emmerich, 1997). Elwood (1995) also discovered some links between subject matter and gender differences in writing performance. It also makes sense to expect that the way people see and describe the world in their mother tongue will also affect their L2 vocabulary knowledge. However, the existing gender differences in frequency of encountering, use, familiarity, interest and emotional

reaction to words cannot explain the reasons for observed DIF in all 11 items. For example, the gender difference in the difficulty for words like 'ward' and 'plot', which are among those with highest indices of DIF (t), cannot be easily explained by the above-mentioned factors. Although the IRT approach ensures sample-free estimation of item parameters, and the model-data fit study confirmed the invariance of item parameter estimations, additional replication studies are needed to check whether the observed differences are genuine or artefactual.

3 Analysis of test fairness

The fact that differences were observed in the 'passive' vocabulary of adult females and males is not a surprising result in itself, but the problem is that it might produce a biased estimation of their language proficiency. On the other hand, in view of future item-bank development and use, not only should the total test be bias-free, but any item composites should also be unbiased. To analyse the impact of DIF on ability measurement, four different estimations of ability parameter for each examinee were obtained as follows:

- θ_T: ability estimation, based on the whole set of vocabulary items (T_T), consisting of 40 items;
- θ_F: ability estimation, based on sub-set 1 (T_F), consisting of the items that were easier for females; i.e., nos. 1, 4, 5, 9, 10, 12, 15, 16, 19, 22, 23, 25, 30, 34, 37, 38, 39 and 40 (18 in total);
- θ_M: ability estimation, based on sub-set 2 (T_M), consisting of the items that were easier for males; i.e., nos. 2, 3, 6, 7, 8, 11, 13, 14, 17, 18, 20, 21, 24, 26, 27, 28, 29, 31, 32, 33, 35 and 36 (22 in total);
- θ_N: ability estimation, based on sub-set 3 (T_N) consisting of the items with no significant difference in difficulty; i.e., nos. 3, 4, 5, 8, 9, 10, 12, 13, 14, 15, 17, 18, 19, 20, 21, 23, 24, 25, 26, 28, 30, 31, 32, 33, 35, 36, 37, 38 and 39 (29 in total).

As can be seen in Table 4, males had significantly better results than females on the total test (T_T). Removing items with DIF indications (T_N) slightly decreases this difference, but it is still significant. In other words, in spite of the fact that some of the items in the total test show differential item functioning, the test as a whole is not gender-biased because the observed differences in the test results remain even after excluding the items with an indication of DIF (Figure 4). A possible explanation for this might be the fact that the number of items with DIF in favour of females (5 items) is almost the same as

Table 4 Comparison between ability parameter estimations

	Females ($n = 293$)		Males ($n = 182$)		t	Significance
	Mean	SD	Mean	SD		
θ_T	0.599	0.591	0.715	0.602	−2.065	0.039
θ_N	0.593	0.603	0.705	0.592	−1.977	0.049
θ_F	0.629	0.612	0.593	0.625	+0.611	0.542
θ_M	0.544	0.600	0.791	0.587	−4.401	0.000

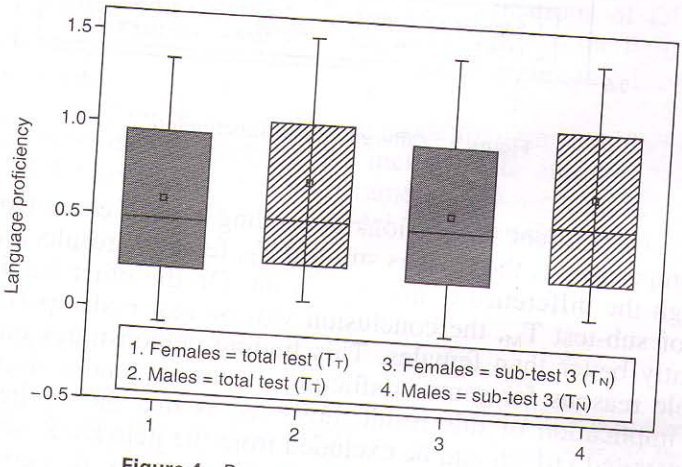

Figure 4 Person parameter estimation

the number of items that advantage males (6 items). This result supports the findings of Roznowski and Reith (1999) that a test containing differentially functioning items can still be free of bias.

Despite the fact that the total vocabulary test (T_T) is not gender biased, some item composites can still give biased estimates of latent ability. The two sub-tests T_F and T_M are a good illustration of this possibility. As can be seen in Figure 5, the sub-test T_F favours females, while T_M advantages males. Both groups have highest scores on average on the sub-test that favours them (females on T_F and males on T_M) in comparison with other tests, and for both groups the mean difference between the two sub-tests (T_F and T_M) is statistically significant at the level $p < 0.01$ ($t = 3.101$ for females and $t = -5.349$ for males).

These results could be expected because of the way these two sub-tests were constructed: T_F consisted only of the items easier for females and T_M consisted only of the items that were easier for males. The comparison between females and males, however, would lead to

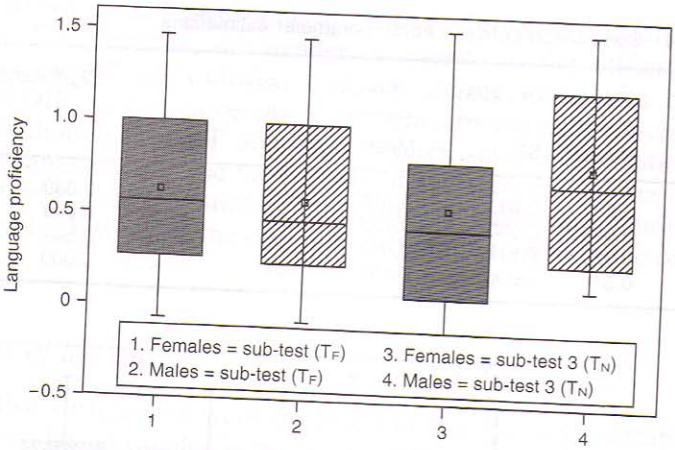

Figure 5 Differential test functioning

completely different conclusions depending on the chosen basis for this comparison. If the basis is sub-test T_F, females' results are better, although the difference is not significant. On the other hand, on the basis of sub-test T_M, the conclusion will be that males perform significantly better than females. This finding demonstrates one of the possible reasons for some conflicting results in gender studies. The main implication of this result, however, is that the 11 items with demonstrated DIF should be excluded from the item bank, since some of their composites can produce biased estimates of person parameters.

One of the basic advantages of item-response modelling is that it gives an opportunity for test-free ability estimation. This means that the person ability parameter estimates are invariant and do not depend on the choice of the items included in the test. In other words, estimating person ability parameters on the basis of different sub-tests should lead to similar results. In this study, however, splitting the total test into two parts T_F and T_M, led to inconsistent results, and the main reason for this is that these two sub-tests were constructed with a deliberate bias. Without preliminary DIF analysis, however, such biased item composites can indeed be created by chance and in this way one of the main advantages of IRT – the invariance of person parameter estimations – will be lost.

IV Implications

The results of this study show that:

- There are significant differences in the 'passive' L2 vocabulary

of females and males, which have to be taken into consideration in test construction and in the measurement of language proficiency. Otherwise, biased estimation of latent ability can occur.

- Some of the observed gender differences in vocabulary knowledge appear to be readily explainable by different sex roles and stereotypes. However, more DIF analytical studies are needed in order to identify possible determinants of gender DIF in language testing.
- Although it is theoretically possible for a test containing DIF items to be bias-free, this needs empirical verification.
- The traditional advice for items with indications of DIF to be excluded from the test gains new significance in the light of item bank building and use since it will prevent possible biased item composites.
- Any application of item-response modelling ought to be preceded by a model-data fit study, which includes DIF analysis. Statistical fit and the invariance of item parameters are not a sufficient basis for the conclusion that the model is appropriate for all examined groups.

V References

Bolt, D. and **Stout, W.** 1996: Differential item functioning: its multidimensional model and resulting SIBTEST detection procedure. *Behaviometrika* 23, 67–95.

Born, M. and **Linn, R.** 1994: Sex differences on the Dutch WISC-R: a comparison with the USA and Scotland. *Educational Psychology* 14 (2), 249–55.

Boyle, J. 1987: Sex differences in listening vocabulary. *Language Learning* 37, (2), 273–84.

Cole, N.S. 1997: *The ETS gender study: how females and males perform in educational setting*. Princeton, NJ: Educational Testing Service.

Cole, N.S. and **Moss, P.** 1993: Bias in test use. In Linn, R., editor, Educational measurement. New York: American Council on Education, Oryx Press, 201–19.

Crocker, L. and **Algina, J.** 1986: *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Draba, R. 1977: *The identification and interpretation of item bias*. MESA Memorandum 25, Chicago, IL: MESA.

Edwards, R. 1998: The effect of gender, gender role, and values on the interpretation of messages. *Journal of Language and Social Psychology* 17 (1), 52–72.

Einarsson, J. and **Hultman, T.G.**, editors, 1984: *Godmorgon pojkar och flickor: om språk och kön i skolan* [Good morning, boys and girls: on language and gender in the school]. Malmö: Liber.

Elwood, J. 1995: Undermining gender stereotypes: examination and coursework performance in the UK at 16. *Assessment in Education: Principles, Policy and Practice* 2 (3), 283–304.

FUMS 1977–79: *Könsroller i språk 1–3* [Gender roles in language], Reports 47, 61,75. Uppsala: Institution for Nordic Languages.

Gibson, J. 1977: The theory of affordances. In Shaw, R. and Bransford, J., editors: *Perceiving, acting and knowing: toward an ecological psychology*. Hillsdale, NJ: Lawrence Erlbaum, 67–82.

—— 1986: *The ecological approach to visual perception*. Hillsdale, NJ: Lawrence Erlbaum.

Hambleton, R.K. and **Swaminathan, H.** 1985: *Item Response Theory: principles and applications*. Boston, MA: Kluwer-Nijhoff.

Hannah, A. and **Murachver, T.** 1999: Gender and conversational style as predictors of conversational behaviour. *Journal of Language and Social Psychology* 18 (2), 153–75.

Hyde, J. and **Linn, M.** 1988: Gender differences in verbal ability: a meta-analysis. *Psychological Bulletin* 104 (1), 53–69.

Kelly, P. 1991: Lexical ignorance: the main obstacle to listening comprehension with advanced foreign language learners. *International Review of Applied Linguistics in Language Teaching* 29 (2), 135–50.

Levine, T. and **Geldman-Caspar, Z.** 1996: Informal science writing produced by boys and girls: writing preference and quality. *British Educational Research Journal* 22 (4), 421–40.

Lord, F. 1980: *Application of Item Response Theory to practical testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lynn, R. and **Dai, X.Y.** 1993: Sex differences on the Chinese standardization sample of the WAIS-R. *Journal of Genetic Psychology* 154 (4), 459–64.

Lynn, R. and **Mulhern, G.** 1991: A comparison of sex differences on the Scottish and American standardisation samples of the WISC-R. *Personality and Individual Differences* 12, 1179–82.

Maccoby, E.E. and **Jacklin, C.N.** 1974: *The psychology of sex differences*. Stanford, CA: Stanford University Press.

Mielikäinen, A. 1988: Naiset puhekielen säilyttäjinä ja uudistajina [Women as preservers and innovators of spoken language]. In Laitinen, L., editor, *Isosuinen nainen*. Helsinki: Yliopistopainos, 92–111.

Nuolijärvi, P. 1988: Sukupuoli kielellisen tuotoksen muovaajana [Gender as a determinant of language production]. In Laitinen, L., editor, *Isosuinen nainen*. Helsinki: Yliopistopainos, 73–91.

O'Neill, K., McPeek, W. and **Wild, C.** 1993: *Differential item functioning on the graduate management admission test*. Research Report (RR-93-35). Princeton, NJ: Educational Testing Service.

Raju, N. and **Drasgow, F.** 1993: An empirical comparison of the area methods, Lord's chi-square test, and the Mantel–Haenzel technique for assessing differential item functioning. *Educational and Psychological Measurement* 53 (2), 301–15.

Roznowski, M. and **Reith, J.** 1999: Examining the measurement quality of tests containing differentially functioning items: do biased items result

in poor measurement? *Educational and Psychological Measurement* 59 (2), 248–70.

Ryan, K. and **Bachman, L.** 1992: Differential item functioning on two tests of EFL proficiency. *Language Testing* 9 (1), 12–29.

Sanders, J.T. 1996: An ecological approach to cognitive science. *The Electronic Journal of Analytic Philosophy* 4 (4), available on the internet at: http://www.phil.indiana.edu/ejap/1996.spring/sanders.1996.spring.html [accessed April 2000].

Shepard, L. 1981: Identifying bias in test items. In Green, B., editor, *Issues in testing: coaching, disclosure, and ethnic bias*. San Francisco, CA: Jossey-Bass.

Smith, R. 1996: A comparison of the Rasch separate calibration and between-fit methods of detecting item bias. *Educational and Psychological Measurement* 56 (3), 403–19.

Sticker, L.J. and **Emmerich, W.** 1997: *Possible determinants of gender DIF on the advanced placement psychology test: familiarity, interest, and emotional reaction*. Research Report (RR-97–24). Princeton, NJ: Educational Testing Service.

Tannen, D. 1986: *That's not what I meant! How conversational style makes or breaks your relations with others*. New York: William Morrow.

—— 1990: *You just don't understand: women and men in conversation*. New York: William Morrow.

Thorne, B. and **Henley, N.**, editors, 1975: *Language and sex: difference and dominance*. Rowley, MA: Newbury House.

Thorne, B., Kramarae, C. and **Henley, N.**, editors, 1983: *Language, gender and society*. Rowley, MA: Newbury House.

van Lier, L. 1996: *Interaction in the language curriculum: awareness, autonomy and authenticity*. London: Longman.

Verheelst, N.D., Glas, C.A.W. and **Verstralen, H.H.F.M.** 1995: *One-parameter logistic model: OPLM*. Arnhem: CITO.

Wainer, H. and **Lukhele, R.** 1997: How reliable are TOEFL scores? *Educational and Psychological Measurement* 57 (5), 741–59.

Wright, D. and **Stone, M.H.** 1979: *Best test design*. Chicago, IL: MESA.

Appendix 1 Items with indications for DIF

M1.	ache			
	a. alue *area*	b. kipu *ache*	c. väite *claim*	d. tuhka *ashes*
M2.	association			
	a. kilpailu *competition*	b. yhdistys *association*	c. vieraantuminen *alienation*	d. syytös *accusation*
M6.	estate			
	a. tila *estate*	b. muuri *wall*	c. väittää *assert*	*d. estää prevent*

M7.	edge			
	a. aita *fence*	b. viha *hate*	c. kutina *itch*	d. reuna *edge*
M11.	grease			
	a. hölmö *fool*	b. rasva *grease*	c. ongelma *problem*	d. possu *piglet*
M16.	jelly			
	a. masu *tummy*	b. hyytelö *jelly*	c. hillo *jam*	d. kudos *tissue*
M22.	plot			
	a. juoni *plot*	b. aitta *shed*	c. näyte *sample*	d. solmu *knot*
M24.	presume			
	a. savustaa *smoke*	b. olettaa *presume*	c. mainostaa *advertise*	d. varmistaa *ascertain*
M27.	rookie			
	a. kukko *cock*	b. kiertotie *detour*	c. lukko *lock*	d. alokas *rookie*
M29.	rust			
	a. talonpoika *peasant*	b. sorto *oppression*	c. kuori *cover*	d. ruoste *rust*
M34.	turn grey			
	a. kalveta *turn pale*	b. homehtua *moulder*	c. harmaantua *turn grey*	d. tulla vihaiseksi *become angry*
M40.	ward			
	a. osasto *ward*	b. sarana *hinge*	c. varoitus *warning*	d. holhooja *guardian*