

Differential Item Functioning (Test Bias) Analysis Paradigm across Manifest and Latent Examinee Groups (On the Construct Validity of IELTS)

* Parviz Birjandi

Islamic Azad University

** Mohadeseh Amini

Islamic Azad University

Abstract

When important decisions are made based on test scores, it is critical to avoid bias, which may unfairly influence examinees' scores. Bias is the presence of some characteristics of an item that results in differential performance for individuals of the same ability but from different ethnic, age, sex, academic, cultural, or religious groups. The investigation of differential item functioning (DIF) is crucial in language proficiency tests in which test-takers with adverse backgrounds are involved, because DIF items pose a considerable threat to the validity of tests. This study mainly focuses on the detection and investigation of the possible multi-dimensional causes of DIF in IELTS, Listening and Reading sub-sections. The main aim of this project is to provide test constructors with information as detailed as possible about sources of DIF in order to help them avoid item bias in future forms of the test. The study was carried out in three steps. First, DIF items were identified using Item Response Theory (IRT) Likelihood Ratio Approach (LR) and Mantel-Haenszel statistical procedure. The second stage was an investigation of the multi-variant sources of DIF for different groups of candidates and whether it worked to their advantage or disadvantage. At the third stage, it was decided whether a possible source of DIF was relevant or irrelevant to the construct the test claimed to measure.

Keywords: Bias, Construct Validity, Differential Item Functioning (DIF), Manifest vs. Latent Groups, Mantel Haenszel Procedure

تحلیل الگویی عملکرد متفاوت سؤال بر گروه‌های آزمودنی آشکار و پنهان (در خصوص سنجش اعتبار ساختاری آزمون IELTS)

پرویز بیرجندی

استاد، دانشگاه آزاد اسلامی واحد علوم و تحقیقات

محدثه امینی

دکتری آموزش زبان انگلیسی، دانشگاه آزاد اسلامی

چکیده

هنگامی که تصمیم‌گیری‌های مهم براساس نمرات آزمون اتخاذ می‌شود به سبب شدت از هرگونه جهت‌گیری اجتناب کرد، چرا که ممکن است نمرات آزمون‌دهندگان به طرز ناعادلانه‌ای تحت تأثیر قرار گیرد. جهت‌گیری عبارت است از حضور ویژگی‌ای در سؤال که به عملکرد متفاوت افراد می‌انجامد؛ افرادی که متعلق به گروه‌های توانش یکسان بوده ولی به گروه‌های متفاوت سنی، قومی، جنسیتی، تحصیلی، فرهنگی و یا مذهبی تعلق دارند. مطالعه و بررسی عملکرد متفاوت سؤال (DIF) از اهمیت به‌سزایی در آزمون‌های توانش زبانی برخوردار است که در آن‌ها آزمون‌دهندگان با زمینه‌های متفاوت حضور دارند، چرا که سؤالاتی که عملکردی متفاوت دارند رویی آزمون را مورد تهدید قرار می‌دهند. مطالعه حاضر عمدتاً به جستجو و بررسی دلایل چندگانه محتمل عملکرد متفاوت در قسمت‌های مهارت خواندن (Reading) و شنیدن (Listening) آزمون IELTS می‌پردازد. هدف اصلی این پروژه آن است که اطلاعاتی تا حد امکان جزئی پیرامون دلایل عملکرد متفاوت در اختیار طراحان سؤال قرار دهد، تا به آن‌ها در اجتناب از جهت‌گیری در گونه‌های بعدی آزمون یاری رساند. این مطالعه در سه مرحله صورت پذیرفته است. نخست، سؤالات عملکرد متفاوت با استفاده از نظریه‌های رویکرد نسبت احتمالاتی پاسخ (IRT) و رویکرد آماری منتل - هنزل مشخص گردیدند. مرحله دوم به بررسی دلایل چندگانه عملکرد متفاوت سؤالات در گروه‌های مختلف آزمون‌دهندگان پرداخته است و نیز این‌که آیا عملکرد متفاوت سؤالات به نفع یا ضرر آن‌ها می‌انجامد. و مرحله سوم به تصمیم‌گیری پیرامون این مسأله پرداخته است که آیا دلایل احتمالی عملکرد متفاوت سؤالات با ساخت زیربنایی آزمون مرتبطند یا خیر.

کلیدواژه‌ها: سوگیری، اعتبار ساختاری، کارکرد متفاوت سؤال، گروه‌های آشکار در برابر گروه‌های پنهان، آزمون منتل - هنزل

* Ph.D in TESL from University of Colorado. Professor, Faculty of Foreign Languages, Research and Science Campus.

** Ph.D in TEFL, Faculty of Foreign Languages, Research and Science Campus.

1 - Introduction

In spite of its claims, IELTS often acts partial across reference and focal groups. As for the younger examinees, IELTS Academic Reading and Writing tasks seem cognitively demanding as evidenced in their varying degrees of success depending in part on their age and experience. Also, when candidates retake IELTS, it can be frustrating for them to see their score on one component improves and their score on another component falls down, leaving the overall result changed. There might be several reasons why this might happen. One reason may be the very nature of language learning. Language learning is a dynamic process involving both acquisition (improving ability in some aspects of language) and attrition (loss of ability in others). Between IELTS tests (a minimum period of 90 days) both of these processes might take place which can affect score profiles. Other reasons derive from the nature of tests and measurement. In addition to the candidate's language ability, differences in test content across versions and other variables such as the test taker's mood or state of health at the time can also affect their scores and contribute to unexpected variations. There might also be potential impact of gender in different sub-sections especially Reading, thus rendering the test a strongly gender differentiated event. A prior familiarity with the content as dictated by the candidate's academic background can also bring about construct irrelevant group variations. The existence of all such potential sources of bias (sex, age, academic background, familiarity with the content, etc.) and their unfair impact on test-takers' overall scores has motivated this research.

The following research questions will be specifically investigated in this study:

- Q1. Is there a risk of gender bias in the studied versions of IELTS?
- Q2. Does the structure of the test (item format and type) unfairly affect the (Sciences vs. Humanities)?
- Q3. Do the contents of the IELTS modules unfairly affect the performance of different focal groups?

The null hypotheses below would next be formed based on the questions above:

- H1: There is no risk of gender bias in the studied versions of IELTS.
- H2: The structure of the test (item format and type) does not unfairly affect the performance of the members of different focal groups, e.g., academic majors (Sciences vs. Humanities).
- H3: The contents of the IELTS modules do not unfairly affect the performance of different focal groups.

In addition to this introduction, this paper consists of six other sections. In section 2, a background of the study would be presented including a discussion on test bias and the existing methods used to investigate differential item functioning. Section 3 provides a review of the previous studies. Section 4, methodology, offers discussions on subjects, instrumentation and the type of

analyses. Section 5 deals with substantial and statistical analyses of the results. The next section discusses the results of the research, followed by a final section on concluding remarks and recommendations for further research.

2 - Background

English as second or foreign language proficiency tests are used mainly to measure the English language ability of test-takers whose native language is not English. The effect of test-takers' diverse characteristics on their performance in these tests has been one of the primary concerns among language testers and researchers. It is vital to investigate whether a test includes potential sources of bias against some particular groups of test-takers. That is, it is essential that a test be fair towards all applicants, and not be biased against a segment of the applicant population.

Bias results in systematic errors that distort the inferences made (Angoff 1993). In many cases, test items are biased due to the fact that they contain sources of difficulty that are irrelevant or extraneous to the construct being measured, and these extraneous or irrelevant factors affect test performance. Perhaps at such times, the item is tapping a secondary factor or factors over-and-above the one of interest. If systematic patterns of difference in test performance are observed across different language groups, the source of the disparity needs to be investigated as a potential threat to the validity of the score interpretations.

In other words, bias occurs when tests yield scores or promote score interpretations that result in different meanings for members of different groups e.g., race, ethnicity, language, culture, gender, disability, or socio-economic status (Camilli and Shepard 1994). Bias is often attributed to construct-irrelevant dimensions that differentially affect the test scores for different groups of examinees. Group differences can also be attributed to item *impact*. *Impact* occurs when construct-relevant dimensions differentially affect the test scores for different groups of examinees. In this case, the item would be a relevant measure of the target construct and the difference between the groups reflects a true difference on that construct. Differential item functioning (DIF) studies are designed to identify and interpret these construct-related dimensions using a combination of statistical and substantive analyses. The statistical analysis involves administering the test, matching members of the base (reference) and one or more comparison (focal) groups on the measure of ability derived from that test, and applying statistical procedures to identify group differences on test items. A focal group is commonly a subpopulation of interest to the researcher, and the reference group serves as the standard for comparison. An item exhibits DIF when examinees from the reference and focal groups differ, on average, in their probabilities of answering that item correctly, after controlling for ability. The substantive analysis builds on the statistical analysis because DIF items are often scrutinized by expert reviewers (e.g., test developers or content

specialists) who attempt to identify construct-related dimensions that produce group differences. A DIF item is considered biased when reviewers identify some dimensions, deemed to be irrelevant to the construct measured by the test, that place one group of examinees at a disadvantage. Conversely, a DIF item displays impact when the dimension that differentiates the groups is judged to be relevant to the construct measured by the test.

A variety of methods have been developed for detecting DIF (e.g., the Mantel-Haenszel procedure, the Standardization procedure, logistic regression, logistic discriminant function analysis, Lord's chi-square, Raju's area measures, the likelihood ratio test, etc.). And considerable progress has been made in the development and refinement of such methods (Clauser & Mazor 1998; Millsap & Everson 1993), but the development and refinement of substantive methods designed to aid with the interpretation of these items have lagged far behind (Bond 1993; Camilli & Shepard 1994; Englehard, Hansche & Rutledge 1990).

3 - Previous Studies

One of the earliest DIF investigations, as applied to language tests, came from Chen and Henning's (1985) study, which examined DIF on the English as a Second Language Placement Examination (ESLPE) for examinees with different language backgrounds (i.e., Chinese and Spanish). For DIF detection, they used Transformed Item Difficulty (TID) or Delta method developed by Angoff (1993). The basic idea of the TID method is to compare the relative ordering of item difficulty indices across two groups, and items that are outliers in terms of item difficulty are flagged for bias. Item level data from 111 examinees (i.e., 77 Chinese and 34 Spanish) were utilized for the estimation of Rasch item difficulty parameter. However, the sample size was too small for the difficulty parameter to be reliably calibrated. Furthermore, the TID method is not based on conditioning on ability (Camilli and Shepard 1994), which may raise questions about the results of the study.

Ryan and Bachman (1992) employed a more advanced technique for the detection of items that function differentially across Indo-European (IE) and Non-Indo-European (NIE) language groups on the First Certificate of English (FCE) and the Test of English as a Foreign Language (TOEFL). For the analysis, they used the Mantel-Haenszel procedure, which reports an averaged weighted odds-ratio difference (i.e., MH) between the focal and reference group across an entire score level (Dorans and Holland 1993; Holland and Thayer 1988). Results of the study identified a total of 65 TOEFL items as showing DIF, with 32 items easier for the IE group and another 33 items easier for the NIE group. Similarly, the FCE had a total of 25 DIF items with about an equal number of items favoring each language group. It must be noted, however, that the Mantel-Haenszel method is not sensitive to non-uniform DIF (e.g., Thissen *et al.* 1988; 1993), and hence is not recommended for DIF studies focusing on probability differences in item difficulty as well as item discrimination.

DIF has also been examined for tests such as the Scholastic Aptitude Test (SAT; e.g., Lawrence *et al.* 1988; Lawrence and Curley 1989; Carlton and Harris 1992), the Graduate Record Exam (GRE; e.g., Scheuneman and Gerritz 1990), the Graduate Management Admission Test (e.g., O'Neill *et al.* 1993), and the National Teacher Exam (e.g., McPeck and Wild 1992). For instance, Lawrence *et al.* (1998) investigated gender DIF on the verbal sections of the SAT using the Standardization approach, which reports a standardized averaged difference (i.e., DSTD) in proportion correct with the reference and focal groups weighted by the standardization group (e.g., Dorans and Holland 1993). The studies have shown that females tend to perform less well on items with technical reading passages than a matched group of male examinees. It has also been found that for sentence completion, gender DIF appears to be associated with science vs. non-science item content, and non-technical science items are easier for females.

Much of the research regarding the effects of language background on second language test performance has been concerned with whether ESL/EFL language proficiency and placement tests measure the same constructs for different language groups (Brown 1999). Only a few studies have examined how examinees from different language groups perform differently on dichotomously scored proficiency or placement tests at the item level (see Chen and Henning 1985; Ryan and Bachman 1992).

Although the studies mentioned here shed significant light on DIF, they are not without limitations. First, most studies focused on the detection of uniform DIF, and non-uniform DIF, which results from probability differences in item discrimination, received relatively little attention. Furthermore, most studies, especially those based on the Mantel-Haenszel and the standardization procedure, used total test scores as a matching criterion to make comparisons with comparable examinees. However, in most cases, items comprising the total test scores were not purified before DIF detection. This may threaten the trustworthiness of the results because the presence of initial biased items influences the accurate estimation of ability as measured by total test scores, and accordingly, the contaminated measure of ability may distort DIF detection.

4 - Method

4-1- Participants

The present study utilized dichotomous-converted scored item level data from the Reading (General and Academic) and Listening modules of the 2003 and 2005 International English Language Testing System (IELTS) of the Iranian candidates. A total of 580 examinees – of different genders, educational/professional backgrounds, and age-groups – took part in this research. Out of this number, a sample of 430 examinees (240 Sciences and 190 Humanities), with 230 male and 200 female candidates across academic and general groups

were selected for the analysis. About 47% of the examinee population was female, and the age of the examinees ranged from 16 to 54.

4 - 2 - Instrumentation

The instruments used in this study consisted of the Listening and Reading comprehension modules of the 2003 and 2005 Academic and General IELTS tests of English language proficiency.

In the Listening modules of the tests – lasting 30 minutes each – there were 40 questions (in each test) in four sections. The first two sections were concerned with social needs and the two final with situations related more closely to educational or training contexts. The candidates' Listening proficiency was assessed via a variety of question types including: multiple choice items, short-answer questions, sentence completion, notes/ summary/ diagram/ flow-chart/table completion, classification and so on.

The Reading modules –lasting 60 minutes each – included three reading passages (40 questions) with a total of 2,000 to 2,750 words. In the academic versions, texts were taken from magazines, books and journals with a specialist taste in either Sciences (e.g., Medicine, Bio-technology, Chemistry, etc.) or Humanities (e.g., Law, English Language Teaching, etc.).

4 - 3 - Analysis

Using the DIF framework, substantive and statistical analyses were conducted to identify and interpret construct-related dimensions that can produce biased differences on the Listening and Reading comprehension constructs. The Listening comprehension scripts and Reading comprehension passages were compared across different manifest (gender, academic background, etc) and latent groups.

As per suggestions from Thissen *et al.* (1988; 1993), a subtest of anchor items to serve as the matching criterion, which are free from DIF, was initially identified based on the Mantel-Haenszel procedure. The item level data based on 430 General and Academic IELTS candidates were subjected to the IRT 3PL model that presents the probability that a randomly selected examinee with an ability of θ (i.e.,) answers an item correctly, using item difficulty (b parameter), item discrimination (a parameter), and pseudo-guessing (c parameter) (Hambleton *et al.* 1991).

The fit of the sample data to a selected IRT model was assessed using the DIM test statistic (Stout 1987), which evaluates an overall goodness-of-fit between data and a selected IRT model via tetrachoric factor analysis. The results of the DIM test showed that a modified 3 parameter logistic (3PL) IRT model fitted both the Listening Comprehension ($p= 0.2958$) and the Reading Comprehension ($p= 0.5205$) subscales, thus suggesting that the subscales were essentially unidimensional.

After screening the DIF items through IRT and the Mantel-Haenszel tests, the multi-variant sources of such bias were determined. At last, it was decided whether such differences were statistically significant to endanger the tests' construct validity thus rendering them biased.

5 - Results

After the two versions of the IELTS test (2003 & 05) were administered, a total of 430 General and Academic candidates were selected for the present study. Based on their purified (total test score minus the DIF items based on a preliminary DIF-screening study) total test scores, the candidates were first classified into manifest reference and focal groups (males vs. females and Sciences vs. Humanities).

Using the Mantel-Haenszel statistical procedure and the IRT Likelihood Ratio approach, DIF items performing differentially across different reference and focal groups were flagged as nuisance items. Such items went through later scrutiny to shed light on the causes of such differential functioning (bias). The results of this section were further validated by expert judgments and comments from the student-think-aloud experiment. Finally, to check for how much the old-existing manifest classifications mapped onto latent groupings, the candidates were divided into groups based on the similarity in their response patterns. A significant correlation was observed here.

5 - 1 - Item Response Theory

The Listening and Reading Comprehension item level data of 430 General and Academic IELTS candidates was subjected to the IRT 3PL model that presents the probability that a randomly selected examinee with an ability of theta (i.e., θ) answers an item correctly, using item difficulty (b parameter), item discrimination (a parameter), and pseudo-guessing (c parameter) (Hambleton *et al.* 1991). The mathematical expression of the 3 PL IRT model is as follows:

$$P(X=1/\theta) = C + \frac{1-C}{1 + e^{-Da(\theta-b)}}$$

where x is an item response, θ is the estimated ability, a is item discrimination, b is the item difficulty, c is pseudo-guessing parameter, D is a scaling factor ($=1.7$) that is devised to approximate the IRT models to a cumulative normal curve, and e is a transcendental number whose value is 2.718.

a - Listening Comprehension

There were 80 items (40 items in each test) in the Listening Comprehension Modules of 2003 and 2005 IELTS tests. Based on the results of a prior Mantel-Haenszel DIF statistics, six items were flagged to be performing differently and thus eliminated from the set of the remaining items to render the purified matching criterion.

To further examine DIF by means of the IRT model, each item was fitted with a modified 3 PL IRT model, where a prior distribution was imposed on the c parameter (Thissen *et al.* 1988). Thus, a total of 80 items with 160 parameters were studied for DIF. Tables 1 and 2 below summarize these results.

Table 1 - IRT-DIF Items for Males and Females in the Listening Comprehension

DIF items	G^2 $a = a^a$	P a=a	$a_{males} - a_{females}^b$	G^2 $b = b^c$	P b=b	$b_{males} - b_{females}^d$	$RMSD^e$
7	5.5	.0190	.21	9	.0111	-.05	.0495
13	.2	.7618	-.02	13.7	.0003	.13	.0389
38	.0	1	-.02	7.1	.0095	-.12	.0543

NOTES: $H_0^a : a_{males} - a_{females}$. G^2 is the difference in G^2 for constrained a and free parameter models.

$a_{males}^b - a_{females}$ is the difference in a parameters for the males and females obtained in the free model.

$H_0^c : b_{males} - b_{females}$, G^2 is the difference in G^2 for constrained a and b parameter models.

$b_{males}^d - b_{females}$ is the difference in b parameters for the males and females obtained in the free models.

$RMSD^e$ is the mean squared difference in probabilities for the males and females estimated in the free model.

Table 2 - IRT-DIF Items for Sciences and Humanities in the Listening Comprehension

DIF items	G^2 $a = a^a$	P a=a	$a_{humanities} - a_{sciences}^b$	G^2 $b = b^c$	P b=b	$b_{humanities} - b_{sciences}^d$	$RMSD^e$
52	4.5	.0334	.09	9.2	.0102	.02	.0289
74	7.2	.0082	.16	14.7	.0007	-.022	.6490
N							

OTES: $H_0^a : a_{humanities} - a_{sciences}$. G^2 is the difference in G^2 for constrained a and free parameter models.

$a_{humanities}^b - a_{sciences}$ is the difference in a parameters for the males and females obtained in the free model.

$H_0^c : b_{humanities} - b_{sciences}$, G^2 is the difference in G^2 for constrained a and b parameter models.

$b_{humanities}^d - b_{sciences}$ is the difference in b parameters for the males and females obtained in the free models.

$RMSD^e$ is the mean squared difference in probabilities for the males and females estimated in the free model.

As shown in these tables, five items were flagged for gender and major DIF at the .05 significance level. Two items exhibited uniform DIF, and three items (i.e., Item 7, 52 and 74) non-uniform DIF. Specifically, three items (Items 7, 38 and 74) were differentially more difficult for the females and Sciences, whereas two items (Items 13 and 52) were differentially easier for the females and Sciences. The three items (Items 7, 52 and 74) flagged for non-uniform DIF were more discriminating for males and Humanities. All reported RMSD values ranged between .0289 and .6490. When the entire DIF detection procedure was repeated with the sample randomly divided in half – which determines a baseline estimate of Type I DIF error rate (i.e., DIF error rate due to chance) – it was found that one item (Item 52) and one parameter showed DIF due to chance at alpha level of .05.

b - Reading Comprehension

The reading comprehension subscale had 160 items (80 General Reading items in two tests, and 80 Academic Reading items). The Mantel-Haenszel DIF procedure flagged 31 DIF items overall across different focal and reference groups. The MH-DIF items were left aside to help form the purified matching criterion with the remaining items amounting to 129.

To investigate DIF, each item (including the once MH-DIF flagged items) was fitted with a modified 3 PL IRT model, where a prior distribution was imposed on the c parameter (Thissen *et al.* 1988). Therefore, a total of 160 items with 320 parameters were subjected to DIF investigation.

Tables 3 to 5 below report DIF information for each studied item in the General and Academic Reading Comprehension modules. For the Reading Comprehension Subscale, 23 items were flagged for DIF at the .05 significance level. RMSD statistics for the studied items ranged from .0180 to .8901. When the entire DIF detection was repeated, with the sample randomly divided in half, it was found that two items and three parameters showed DIF due to chance at alpha level of .05.

Table 3 - IRT-DIF Items for Males and Females in the General Reading Comp.

DIF items	G^2 $a = a^a$	P a=a	$a_{males} -$ $a_{females}^b$	G^2 $b = b^c$	P B=b	$b_{males} -$ $b_{females}^d$	$RMSD^e$
4	7.1	.0066	.11	7.3	.0260	.32	.0180
11	9.1	.0023	.22	9.1	.0106	0	.0330
16	2.4	.1312	.21	14.7	.0001	.10	.0345
25	0	1	.02	17	0	.10	.0295
27	11.9	.0007	.35	14.8	.0007	.06	.0503
33	4.4	.0359	.18	11	.0041	-.12	.0479
48	8	.0047	.10	10.9	.0043	.12	.0493
51	9.3	.0023	.23	9.9	.0071	.02	.0506
59	.5	.4795	-3.06	6.8	.0091	-.03	.0981
63	5.4	.0201	.20	5.4	.0639	0	.0757
67	4.1	.0429	.16	4.2	.1287	.01	.0696
77	.3	.5839	.21	7.8	.0049	.07	.1708
79	5	.0253	3.54	12.3	.0019	-.38	.0876

NOTES: See Table 1

Table 4 - IRT-DIF Items for Males and Females in the Academic Reading Comp.

DIF items	G^2 $a = a^a$	P a=a	$a_{males} -$ $a_{females}^b$	G^2 $b = b^c$	P b=b	$b_{males} -$ $b_{females}^d$	$RMSD^e$
9	10.1	.0024	.23	10.1	.0172	1	.0479
21	6	.0354	4.65	13.1	.0120	-.49	.0725
34	.4	.6943	.32	8.1	.0051	.08	.0349
40	2.4	.1413	.24	15.6	.0003	.11	.0567
61	.6	.5674	-4.01	6.8	.0092	-.04	.8901
75	0	2	.3	17	0	.11	.3987

NOTES: See Table 1

Table 5 - IRT-DIF Items for Sciences and Humanities in the Academic Reading Comp.

DIF items	G^2 $a = a^a$	P a=a	$a_{humanities} -$ $a_{sciences}^b$	G^2 $b = b^c$	P B=b	$b_{humanities} -$ $b_{sciences}^d$	$RMSD^e$
15	4.4	.0156	.12	8	.0123	-.04	.0434
33	.1	.6789	-.01	12.8	.0005	.12	.0566
68	0	1.001	-.02	6.4	.0081	-.10	.0232
72	3.5	.0356	.08	9.3	.0102	.01	.2567

NOTES: See Table 2

Among the 23 DIF flagged items 10 items exhibited uniform DIF, and 13 items non-uniform DIF (Items 4, 16, 25, 27, 48, 51, 63, 67, and 79 in the General Reading Comp. module, and items 21, 40, 72, and 75 in the Academic Reading module). Within the General Reading comprehension module, items (4, 16, 25, 27, 48, 51, and 77) were differentially easier for females.

5 - 2 - Mantel-Haenszel Procedure

In addition to IRT, the present study made use of the Mantel-Haenszel chi-square (MH- χ^2) method to detect DIF (Mantel and Haenszel 1959) because it has been widely used and a focus of research (e.g., Scheuneman and Gerritz 1990; Schmitt, Hatrup, and Landis 1993).

This method asks: Given two groups of candidates matched on the amount of an attribute, do the two groups differ significantly in the rate at which they endorse each item that measures that attribute? The MH tests the null hypothesis that the odds ratio is 1 for an item with no DIF. Significant deviations from 1 are typically associated with a significant MH- χ^2 value. A significant MH- χ^2 value, reflecting an association between the classification variable (e.g., sex or academic background) and the rate of item endorsement, is taken as evidence of DIF for the studied item. A significance level of .05 was applied to all statistical tests. The purified subscale of items (based on a previous MH-DIF detection) was used as the matching variable. In each module (Listening Comprehension versus Reading Comprehension of IELTS 2003 & 05), the reference group (men or Sciences) was compared against the focal group (women or Humanities) using the MH-DIF procedure. Table 6 summarizes the results.

Table 6 - MH-DIF Items for Each Reference and Focal Group

DIF Category	Gender						Major						Total
	Males			Females			Sciences			Humanities			
	A	B	C	A	B	C	A	B	C	A	B	C	
Listening Comprehension	<i>1</i>					2		1			1		6 (7.5%)
General Reading Comprehension	1	1				9	3						16 (20%)
Academic Reading Comprehension	<i>1</i>	1	2			7	1	2		2	2		15 (18.75%)

NOTE: The italicized numbers indicate diversions from the number of items flagged to be functioning differentially via IRT.

After screening DIF items through IRT and the Mantel-Haenszel statistical procedures, whose results greatly mapped onto each other, the multi-variant sources of such bias were determined. At last, it was decided whether such differences were statistically significant to endanger the tests' construct validity thus rendering them biased.

6 - Discussion

Merely detecting DIF does not identify the element in the item that causes it. Generally, it is not always clear which element(s) in an item causes DIF. One reason for DIF may be the context material, for example, a text, a graph, or a drawing. Other sources might be: the question asked in the item, the four response options, or the interaction between the various item elements. In the second phase of the study, the principle aim was to find the possible sources of DIF, in rather judgmental terms with resort to expert comments.

In order to undertake a more targeted screening of the item elements to find the possible sources of DIF, a large scale literature search was carried out by the researcher. One of the most important findings was that in the international literature only minimal attention was paid to – and very little seemed to be known about – the possible causes of DIF. Schmitt *et al.* (1993) mention three

possible reasons for this. First of all, they observe that DIF research is still in its infancy. Until recently, most attention has been paid to the statistical procedures for detecting DIF. Secondly, the detection of DIF causes presupposes a theory about why items would show DIF for different groups of examinees. And, thirdly, the detection of DIF is a very complex activity, since a number of factors may be at work on an item at the same time.

After the detection of DIF items, the researcher tried to investigate the sources of such bias across various categories of say examinee, content material, item structure/ format, etc. The work here was rather judgmental by nature. However, to reduce the subjectivity of the findings, the outcomes were checked and re-checked with those of other such studies. Also, specialists who had extensive knowledge of the content areas measured by the tests as well as the knowledge and cognitive skills required by examinees to solve the test items were asked to verify the accuracy of the results.

6 - 1 - Content analyses of the DIF items

Once DIF items were detected, it was crucial to investigate the sources of DIF across the groups. DIF items are not necessarily biased items. As Clauser and Mazor (1998) point out, DIF is necessary but not a sufficient condition for demonstrating item bias. Results of the DIF analysis should lead to the investigations of the potential sources of bias.

It has been reported by test developers that they are often confronted by DIF results that they cannot understand; and no amount of deliberation seems to help explain why some perfectly reasonable items have large DIF values. Several investigators (e.g., Cole 1981; Linn 1986; Plake 1980; Tittle 1982) noted that the judgment of bias is generally unreliable. This is not surprising; the judgment of item difficulty by itself is not highly reliable (see, e.g., Thorndike 1982), and the judgment that an item will or will not be differentially difficult is expectedly even less reliable.

In general, however, theories about why items behave differentially across groups can be described only as primitive. Part of the problem, as the researcher sees it, is that the very notion of differential item functioning by groups implies a homogeneous set of life experiences on the part of the focal groups that are qualitatively different from the reference groups.

In this section at first, test items were categorized into groups based on their type (Listening Comprehension versus Reading Comprehension), format (MC, Flow chart, summary completion, True/False/Not Given, etc.), content (Science, Humanities), and the kind of psychological process the examinees needed to undergo in order to respond to the items (inferencing, referencing, scanning, skimming).

Next, a group of three IELTS practitioners were consulted to comment on the probable sources of item DIFs. Since the judges themselves were primarily teachers, it was quite safe to assume that they would be able to provide more

accurate estimates of DIF than individuals who are not practitioners. Further, they received a 45 minute training session on item bias. Referring to Uiterwijk and Vallen (2005), the possible linguistic causes of DIF in test items (see Tables 4-8) were listed as below and distributed among the judges.

Table 7 - Possible Linguistic Causes of DIF in Test Items

1) Word Level
<ul style="list-style-type: none"> ◦ Exact meanings of words; ◦ Low-frequency words; ◦ Words in contexts for example in a text) that do not give explicit information about the meaning of the word; ◦ "Abstract" words; ◦ Ambiguous words in texts, where the context contains no explicit information about the meaning of the word.
2) Sentence Level
<ul style="list-style-type: none"> ◦ Negative sentences; ◦ Passive sentences; ◦ Metaphorical language, idiomatic expressions.
3) Text Level
<ul style="list-style-type: none"> ◦ Texts that require a great deal of memorization; ◦ Texts with content that is not plausible at first sight; ◦ Texts with complex references; ◦ Texts with unfamiliar or incorrect clues with respect to the structure of the text.
4) Metalinguistic Competence
<ul style="list-style-type: none"> ◦ Test items requiring the detection of grammatical mistakes in a text; ◦ Test items requiring the correction of language in a general sense.

Uiterwijk and Vallen 2005

6 - 2 - Results of the DIF Analyses

One type of potential bias that is relevant to investigate is that relating to the sex of the test-taker. According to Tittle's (1982: 25), there is evidence that 'a test or exam can favor female or male testees in three possible ways: "Content", "Format" and "Type"'.

a - General Reading Comprehension

In the General Reading Comprehension subsection, 7 out of the 13 DIF flagged items favored females and 6 proved much easier for males. Upon further scrutiny,

areas of difficulty (DIF causes) were investigated and grouped into content, format, type, and wording categories. As for the causes of the DIF items, the True/False/Not Given item types rendered more difficult for females among all other formats. Format was not a determining cause of DIF as far as males were concerned.

Another significant cause of DIF for females was recognized to be the 'wording' of the item. Here, at the 'word level' (Uiterwijk and Vallen, 2005), the DIF items contained difficult or unknown vocabulary.

As for the males, the reason behind DIF items seemed to lie in the 'type' of the underlying psychological process. Males' DIF items lied where 'inferencing' was needed (as opposed to direct referencing) to provide the correct response to an item.

b - Academic Reading Comprehension

Within the Academic Reading Comprehension module, out of the 10 DIF flagged items, 6 displayed significant gender DIF. Results from the further analyses of the flagged DIF items indicated that, when women were compared to a matched group of men, they typically performed less well than men on reading comprehension items with science-related content. In other words, items based on science passages were generally differentially more difficult for women than for the matched group of men, and items based on social science and humanities passages were generally differentially easier for women than the matched group of men, as also supported by (Scheuneman & Gerritz 1990; Wild & McPeck 1986). It was also found that for sentence completion, gender DIF appeared to be associated with science vs. non-science item content, and non-technical science items were easier for females.

The science results for the reading comprehension items may well reflect differences in attitudes about science for men and women. Differences in the proportions of men and women taking courses in science and planning careers in science have been widely documented (e.g., Ramist & Arbeiter 1986). It may be that there are differences between the groups in their interest in science topics, their confidence in their abilities to understand scientific subject matter, and their comfort level with science passages.

The format of the items and their wordings also seemed to raise problems for females. Females had problems with multiple-choice items. This was while, their comparative male group had difficulties in summary completion and flow-charts. In terms of the linguistic causes of DIF, there were two instances illustrating the fact the wording of the item (i.e., the ways item elements were put together) was the area of difficulty. Within this module, the remaining 4 items displayed significant 'major' differential functioning. It seemed that Science-major candidates were more successful with science content (e.g., *Spider Silk Cuts Weight of Bridges*, the first passage in the 2003 Academic Reading Comprehension subsection). This was while, the Humanities were more successful with social sciences' topics and content (e.g., *Teaching in Universities*, second passage in the 2003 Academic Reading Comprehension subsection).

C - Listening Comprehension

In the Listening Comprehension subsection, it was really much more difficult to comment on the causes of DIF as the judges also confessed. The candidates themselves identified item format and topic along with the item's speeded-ness as sources of difficulty and differential functioning.

As for the females, Multiple Choice and True/False/Not Given items proved more difficult. They specifically seemed to have problems with the Not Given part. With regard to DIF on content, as in the Reading Comprehension module, science topics did not interest females so much unless they belonged to the Science focal group.

Overall, the results of the analyses suggested that the content of Listening and Reading Comprehension items appeared to be a major cause of DIF. At the same time, however, the study also implied that DIF was based on various factors, given the relatively small proportion explained by the content characteristics alone. Findings from the content analysis across the two subscales suggested that items dealing with science-related topics, data analysis, and number counting were differentially easier for the Sciences, whereas items about human relationships were differentially easier for the Humanities. This pattern provided an interesting comparison to the findings reported from gender DIF studies, where females tended to perform better than a matched group of males on reading items related to human relationships, whereas males were likely to outperform a comparable group of females on science-related topics, as also supported by the present study (e.g., Lawrence *et al.* 1988; Lawrence and Curley 1989; Scheuneman and Gerritz 1990; Gafni 1991; Carlton and Harris 1992; Curley and Schmitt 1993; O'Neill and McPeck 1993; O'Neill *et al.* 1993; Maller, 2001). Item format and type proved to be playing roles of differing degrees as well.

7 - Conclusions

The present research was an attempt to detect and investigate the possible causes of DIF in two versions of IELTS (2003 and 05), Listening and Reading Comprehension sub-sections. The study was carried out in three steps. First, DIF items were identified using the Item Response Theory Likelihood Ratio Approach and the Mantel-Haenszel statistical procedure. The second stage was an investigation of the multi-variant sources of DIF for different reference and focal groups of candidates and whether it worked to their advantage or disadvantage. At the third stage, it was decided (based on tests of significance, experts' comments and the students' think-aloud experiments) whether a possible source of DIF was relevant or irrelevant to the construct the tests claimed to measure.

Items in the Listening Comprehension module showed gender and major bias 7.5%. Biased items in the General Reading Comprehension sub-test amounted to almost 20% of the whole. And the rate of bias in the Academic Reading Comprehension module was 18.75%. Item format, content, and type, along with word or sentence-level complexities and ambiguities were

recognized as possible sources of bias in items flagged as DIF. It seems the two studied versions of IELTS suffered from significant DIF patterns.

Such findings do pose serious questions about what the specifications for a test should be, however. What proportion of the passages should involve technical science/female-oriented, etc material? How technical/ female-oriented should such passages be? Should the proportion of such passages be controlled tightly from form to form of the test? ...

Enough information has accumulated to make it timely to begin the process of rethinking test specifications and the role that DIF results should have in determining those specifications. Future DIF analyses and refinements of techniques surely will provide additional grist for that mill, but that will continue to be true for some time and substantial changes in test specifications cannot take place overnight anyway. Thus, it is not too early for the Testing Society and its major clients to begin the process of opening up test specifications for a major review. Experienced test development staff with a hand in DIF research should have an opportunity to have a significant voice in the process.

In the present study, the two statistical methods employed in the DIF detection procedure (the Item Response Theory Likelihood Ratio Approach and the Mantel-Haenszel statistical procedure) might have yielded similar results because they share a common principle. Other researches could take place using other ways of assessing DIF such as the Rasch Logistic Response Model, the Ordinal Logistic Regression Method, the Mixed Rasch Model, etc. to cross-validate such findings.

The results of the current study are mainly based on the 2003 and 05 IELTS Listening and Reading Comprehension item pool. Speaking and Writing modules were put aside due to their highly polytomous nature. Further studies can take place with specific attention paid to such sub-sections. And, in differential item functioning studies, examinees are usually classified into a reference group, or a focal group, or are ignored as members of other groups, based on their responses to a group identification question. Some examinees fail to answer this question and so introduce missing data that can affect the inferences that are based on DIF measures. Because DIF often is concerned with a comparison of an item's performance for a relatively small subpopulation (the focal group) versus its performance for a much larger subgroup (the reference group), even a small amount of missing data may have profound effects on the inferences made (Little and Rubin 1987). Future studies which would address this issue and investigate the effect of missing data on DIF are highly appreciated.

References

- Angoff, W. H. 1993. "Perspectives on Differential Item Functioning Methodology". PP. 3-23, in *Differential Item Functioning*. Holland, P. W. & Wainer, H., editors Hillsdale, NJ: Lawrence Erlbaum.

- Bond, L.. 1993. "Comments on the O'Neill & McPeck paper". PP. 277-280, in *Differential item functioning* Hillsdale, P.W. Holland, & H. Wainer (eds.), NJ: Lawrence Erlbaum.
- Brown, J. D. 1999. "The relative Importance of Persons, Items, Subtests & Languages to TOEFL Test Variance". *Language Testing* 16 (2): 217-38.
- Camilli, G. & Shepard, L. 1994. *Methods for identifying biased test items*. Thousand Oaks. CA: Sage.
- Carlton, S. T. & Harris, A. M. 1992. *Characteristics associated with Differential Item Functioning on the Scholastic Aptitude Test: gender and majority/minority group comparisons*. ETS Research Report, 92-64. Princeton, NJ: ETS.
- Chen, Z., & Henning, G. 1985. "Linguistic and cultural bias in language proficiency tests". *Language Testing* 2: 155-63.
- Clauser, B. E. & Mazor, K. M. 1998. "Using statistical procedures to identify differentially functioning test items". *Educational Measurement: Issues and Practice* 17, 31-47.
- Cole, N. S. 1981. "Bias in testing". *American Psychologist* 36: 1067-1077.
- Curley, W. & Schmitt, A. P. 1993. *Revising SAT-verbal items to eliminate Differential Item Functioning*. College Board Report 93-2. New York: College Entrance Examination Board.
- Dorans, N. & Holland, P. W. 1993. "DIF detection and description: Mantel-Haenszel and standardization", pp 35-66. in Holland, P.W. and Wainer, H., editors, *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum.
- Englehard, G., Hansche, L., & Rutledge, K. E. 1990. "Accuracy of bias review judges in identifying differential item functioning on Teacher Certification Tests". *Applied Measurement in Education* 3, 347-60.
- Gafni, N. 1991. *Differential Item Functioning: performance by sex on reading comprehension tests*. ERIC Document ED 331-844. Rockville, MD: Educational Resources Information center.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. 1991. *Fundamentals of item response theory*. Thousand Oakes, CA: Sage.
- Holland, P. W. & Thayer, D. T. 1988. "Differential Item Performance and Mantel-Hawnszel Procedure", PP.129-45, in Wainer, H. and Braun, H., editors, *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum.
- Lawrence, I. M., Curley, W. E. 1989. *Differential item functioning for males and females on SAT-Verbal Reading subscore items: follow-up study*. Educational Testing Service Research Report 89-22. Princeton, NJ: ETS.

- Lawrence, I. M., Curley, W. E. & McHale, F. J. 1988. *Differential item functioning for males and females on SAT verbal reading subscore items*. Report No. 84-88. New York: College Entrance Examination Board.
- Linn, R. L. 1986. Bias in college admissions. pp. 80-86, in *Measures in the college admissions process: A College Board colloquium* New York: The College Entrance Examination Board.
- Little, R. J. A., & Rubin, D. B. 1987. *Statistical analysis with missing data*. New York: Wiley.
- Maller, S.J. 2001. "Differential item functioning in the WISC-III: item parameters for boys and girls in the national and standardization sample". *Educational and Psychological Measurement* 61: 793-817.
- Mantel, N. & Haenszel, W. 1959. "Statistical aspects of the analysis of data from retrospective studies of disease". *Journal of the National Cancer Institute*, 22- 719, 748.
- McPeck, W. M. & Wild, C. L. 1992. *Identifying differentially functioning items in the NTE Core Battery*. Princeton, NJ: TES.
- Millsap, R. E., & Everson, R. D. 1993. *Item and test scoring with binary logistic models*. Mooresville, IN: Plenum.
- O'Neill, K. A., & McPeck, W. M. 1993. "Item and Test Characteristics that are associated with differential Item Functioning", p. 255-276, in Holland, P.W. and Wainer, H., editors, *Differential Item Functioning*, Hillsdale, NJ: Lawrence Erlbaum.
- Plake, B. S. 1980. "A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process". *Educational and Psychological Measurement* 40: 397-404.
- Ramist, L., & Arbeiter, S. 1986. *Profiles, college-bound seniors 1985*. New York, NY: College Entrance examination Board.
- Ryan, K., & Bachman, L. F. 1992. "Differential item functioning on two tests of EFL Proficiency". *Language Testing* 9: 12-29.
- Scheuneman, J. D. & Gerritz, K. 1990. "Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics". *Journal of Educational Measurement* 27: 109-31.
- Schmitt, A. P., Mazzeo, J., & Bleistein, C. A. 1993. *Sex-related performance differences in constructed-response and multiple-choice sections of Advanced Placement Examinations* (College Board Report No. 92-7). New York: College Entrance Examination Board.
- Stout, W. 1987. "A non-parametric approach for assessing latent trait unidimensionality". *Psychometrika* 52: 589-617.

- Thissen, D., Steinberg, L., & Wainer, H. 1988. "Use of item response theory in the study of group differences in trace lines", p. 147-169, in *Test Validity*, H. Wainer and H. Braun (eds.), Hillsdale, NJ: Erlbaum.
- Thissen, D., & Wainer, H. 1993. *Confidence envelopes for monotonic functions: Principles, derivations, and examples* (Technical Rep. No. 82-37). San Antonio, TX: National Academy Library.
- Thorndike, R. L. 1982. "Item and score conversion by pooled judgment", PP. 309-317, in *Test equating*, P.W. Holland & D.B. Rubin (eds.), New York: Academic.
- Tittle, C. K. 1982. "Use of judgmental methods in item bias studies", PP. 31-63, in *Handbook of methods for detecting test bias*, R.A. Berk (ed.) Baltimore: John Hopkins University Press.
- Uiterwijk, H., & Vallen, T. 2005. "Linguistic sources of item bias for second generation immigrants in Dutch tests". *Language Testing* 22 (2): 211-234.
- Wild, C. L., & McPeck, W. M. 1986. *Performance of the Mantel-Haenszel statistic in identifying differentially functioning items*. Paper presented t the annual meeting of the American Psychological Association, Washington, DC.