Jurnal Teknologi, 50(E) Jun 2009: 21–27

THE ANALYSIS OF IRAN UNIVERSITIES' 2003-2004 ENTRANCE EXAMINATION TO DETECT BIASED ITEMS

IBRAHIM MOHAMMAD POUR[1] & MOHAMED NAJIB ABDUL GHAFAR[2]

Abstract. Item bias or differential item function (DIF) refers to the situation in which the probability of correct responses to an item for examinees with equal ability measured by test but belong to different groups are not equal. The existence of bias in items decreases the validity of the test. In this study the range of item difficulty among surveyed groups, has been used as a method for detecting the item bias in Persian literature subtest as part of the Entrance Examination to Universities of Iran in 2003-2004. For this purpose, report cards of 5000 (each group of 1000 examinees) participants in this examination from three provinces i.e. Yazd, Azerbaijan Sharghi and Kurdistan as sample groups were analyzed using the computerized program, BILOG-MG. Out of 25, two items, numbers 9 and 10 showed bias between gender groups and both were in favour of female group and were identified as biased items. Of this number, four items numbers 2, 7, 9, and 12 showed bias among linguistic groups.

Keywords: University's entrance examination; differential item function (DIF); BILOG-MG

Abstrak. *Bias item* atau *differential item functioning* (DIF) merujuk kepada situasi di mana keberangkalian respons yang betul untuk pengambil peperiksaan yang mempunyai kebolehaan yang sama, tetapi dari kumpulan yang berbeza, apabila diukur melalui ujian, memberi keputusan yang tidak sama dan berbeza. Kewujudan bias dalam item-item mengurangkan kesahan ujian. Dalam kajian ini, turutan kesukaran item di antara kumpulan yang ditinjau digunakan untuk mengenal pasti bias item dalam ujian Sastera Persia, iaitu sebahagian daripada ujian kemasukan ke universiti di Negara Iran bagi tahun 2003-2004. Untuk tujuan ini, data kad laporan dari 5000 peserta (setiap kumpulan terdiri daripada 1000 pengambil ujian) dari tiga daerah, iaitu Yazd, Azerbaijan Sharghi dan Kurdistan yang digunakan sebagai sampel kajian telah dianalisis menggunakan program komputer BILOG-MG. Dapatan kajian menunjukkan daripada 25 item ujian, sebanyak dua item (nombor 9 dan 10) adalah bias jantina dan kedua-duanya berpihak kepada kumpulan perempuan. Sebanyak empat item (nombor 2, 7, 9 dan 12) menunjukkan bias bahasa dan berpihak kepada bahasa Persia (nombor 7), Kurdish (nombor 12) atau Turki (nombor 2 dan 9). Adalah dicadangkan pembentukan item masa hadapan mengambil kira ujian bias dan menolak item yang berkenaan.

Kata kunci: Ujian kemasukan universiti; fungsi pembezaan item (DIF); BILOG-MG

1.0 INTRODUCTION

Tests are divided into two groups: criterion-referenced tests and norm-referenced test. In criterion-referenced tests, examinee's performance is evaluated with respect to a

[1&2] Faculty of Education, Universiti Teknologi Malaysia, UTM Skudai, 81310 Johor
Tel: +60133012654. Email: [1]ebrahim47@gmail.com
Tel: +60197013783. Email: [2]p-najib@utm.my

criterion which is often predetermined educational objectives. In norm-referenced tests, examinee's performance in a test is not evaluated with a predetermined criterion but in comparison with other examinees. University's entrance examination is a good example of norm-referenced tests (Saif, 1998).When the objective of conducting a test is to compare examinee's performance with each other, providing equal condition for all group members is inevitably necessary so that they can show their knowledge and abilities.

The world consists of extensive kind of races, cultures, languages, tribes and diverse religions. Problems concerning student's education in relation to those elements are as diverse and complex as the issue of equal access to educational opportunities and the practice of assessing their aptitudes and abilities. People are still concerned students' problems which are related to differences in cultures, languages and tribes, but most of these differences have been ignored (Lotfabady, 1996).

Different factors can affect the result of any test. These factors are divided in two groups: The first consists of factors including the itemisation, examinee's motivation, fatigue, examinee's attitudes toward the test, which were called random errors. These usually change when the time of conducting the test changes and randomly affect the performance of examinees. The second group which is called systematic errors in the process of measurement consists of factors such as gender, race and language. They are the characteristics of examinees and cannot be changed and bias is often the result of these factors (Axman, 1990).

The concept of bias or differential item functioning (DIF) refers to different performance of an item for members of different groups that are equal in the ability which are measured by the test (e.g. reading ability). In other words, item bias refers to the conditions in which the probability of responding correctly to an item for different groups of examinees after controlling their abilities are not equal (Shepard, Camili and Averill, quoted by Gierl, *et al*. 1999). Tests are conducted and designed for different goals including placement of examinees in different academic grades, employment and prescribing medicine.

One of the functions of tests in Iran is for universities entrance examination. The tests were conducted monotonously for all participants to select students for universities and institutes of higher education. More than one million and three hundred thousand males and females from different cultures and languages, urban and rural areas compete for selection into universities and the only criterion are their scores in this test. Therefore, the result of the test may have a permanent and deep effect on examinee's and their family's morale and future destiny. It is important to carry out a study on the analysis of the items of entrance examination in universities in Iran to detect item bias due to the effect of the examination on their life, language and culture. The main question of this study is to determine the selection fairness of examinees within the context of the examination. Each item of the test can be analyzed for two aspects which are the items fairness for all groups of subjects and their state of bias.

There are different approaches to detect item bias. One of the approach is based on the classical test theory (CTT) and the other is from the item response theory (IRT). Methods based on the item response theory can provide a stronger theoretical framework for reference in comparison with the classical test theory (Boiteau, Bartran and Sient Anj, 2001). This research will use the item response theoretical approach to detect DIF. There are different models in this theory to explain the relationship between examinee's ability and response to each item. The first common model is the one-parameter model that can be used to evaluate item's bias only on the difficulty parameter of item with the assumption that the discrimination parameter of all items are equal and the examinee's pseudo-guessing is not a factor in test performance. The second model is the two parameters model where the model item bias is evaluated on the basis of the difficulty parameter (b) and the discrimination parameter (a).The model also assumes that examinee's pseudo-guessing is not a factor in test performance. The third model which is three-parameters model is a model that evaluate item bias on the basis of three parameters, i.e., the difficulty parameter (b), the discrimination parameter (a) and the examinee's pseudo-guessing parameter (c). The three parameters model is preferred over the other models to evaluated item bias including in multiple choice format because bias from actual differences among examinees can be identified (Lord, Peterson, 1977, quoted by Shepard, Camili and Wiliams, 1985). Furthermore, the three parameters model can be used to identify uniform and non-uniform bias and can be applied as a standard model for comparison with others (Beak, 1997). Rudner (1978), using the three parameters model has analyzed Monte Carlo items and came to a conclusion that the three parameters model can provide satisfying results in comparison with other methods under any conditions (Rudner, 1978, quoted by Osterlind, 1983).

With respect to the above argument, the three parameters model is preferred over the other models. The examinee's pseudo-guessing for a correct response to an item can also be evaluated and is important in the analysis of multiple-choice items. It is possible for an examinee to respond to an item correctly without being aware that the correct response was selected only by pseudo-guessing.

2.0 RESEARCH OBJECTIVE

The main goal of this study is to analyse the subtest items of the Persian Literature for universities entrance examination for the year 2003-2004 to detect items which are biased for male-female gender groups and Persian, Turkish and Kurdish language groups.

3.0 RESEARCH QUESTIONS

The following questions were investigated.

(1) Do the Persian literature items in universities entrance examination have different functions for male and female groups?

(2) Do the Persian Literature items in universities entrance examination have different functions for Persian, Turkish and Kurdish language groups?

4.0 RESEARCH METHOD

Data used in this study were responses from 5000 examinees with a diploma in the area of natural sciences in 2003-2004 universities entrance examination. A total of 25 items of subtest of Persian Literature were analysed with the permission of assessment organization. Data was first entered in SPSS software in terms for examinees gender and language. After arranging the data in terms of their raw scores, response for examinees with all items correct and who did not answer any of the items were omitted. In both cases, it was not able to be identify the DIF. Based on the percentage of scores for all range of scores, the responsiveness trends of 5000 examinees for all items were analysed with the use of computerized program BILOG-MG.

5.0 POPULATION AND SAMPLING

The population in this study was all participants of natural science group in the Yazd province as representating the Persian, Azerbaijan Sharghi of the Turkish group and Kurdistan of the Kurdish language group. All three groups were from the randomly selected mathematics, natural science and human science fields.

The sample was randomly selected from 17175 participants in universities entrance examination from Yazd**,** Azerbaijan Sharghi and Kurdistan provinces from the unit of machinery services in Assessment of Education Organization (AEO). Out of this number, the participants selected were those whose place of birth and residence were the same when they first took the examination as a mean of controlling the mother tongue language variable. After arranging the scores, 5000 participants (each group of 1000) were selected. The conducting of the model of three parameters of item response theory for estimating item parameters requires for at least 1000 examinee in each group (Farzad, 1992).

Table 1 Sample characteristics

Groups	**male**	**female**	**Persian**	**Turkish**	**Kurdish**
Number	1000	1000	1000	1000	1000
Mean	47/5	5/47/5	44/6	44/6	44/6
Standard error of measurement	8/59	8/59	6/93	6/93	6/93
Deviation of Standard	271/82	271/82	219/08	219/08	219/08

In studying bias items, the next step after selecting sub-groups is to match the examinees characteristics. Items can be taken as biased when the probability of a correct respond to an item for examinees with equal ability, but belonging to different

groups is unequal. It is necessary to match the examinees on the basis of a validated criterion. This criterion is often the total score of test or the subtest. The total score of the test is the most common factor to match examinees because standardized tests are designed in such a way that it can measure only one unidimentional construct. However, the score of the total test may not be the most reliable criterion (Ackerman, 1992; Clauser, Nungester and Swaminathan, 1996; Hamilton, 1997, quoted by Nhuan.Le.vi, 1999). Conducted past studies showed that the number of biased items can be reduced to one third (ibid) when the score of subtest, instead of the total test score, were used for matching examinees. Therefore, the same criterion was used for matching the examinees in this study, i.e., the subtest score.

6.0 RESEARCH FINDING

The results as follows are presented accordingly to the research questions. The first one was stated as 'Do the Persian literature item in universities entrance examination have different functions for male and female groups?'

The results showed that out of the 25 items of this subtest, two items, i.e., items 9 and 10 between male and female group have different functions. The value of difficulty difference of item 9 between the two groups equals to 0.242 and the value of difficulty difference of item 10 between the two groups equals to 0.175. The value of difficulty difference for both items at the α level of 0.05 was statistically significant and both items were easier for the female than the male group.

The second research question was 'Do the items of Persian Literature in universities entrance test have different functions for Persian, Turkish and Kurdish language groups?

The result of this study shows that out of the 25 subtest items, four items, i.e., items 2, 7, 9 and 12 among the language groups have different functions. Items 2 and 9 for the Persian and the Turkish language groups also have different functions. The difficulty difference value for item 2 is equal to 0.276 and difficulty difference for item 9 is equal to 0.372. These values were statistically significant at α level equals to 0.05. The two items were easier for the Turkish language group than for the Persian.

Items 7 and 12 have different functions for Persian and Kurdish language groups. The difficulty difference value for item 7 for the two groups was 0.153 and was statistically significant at α level equals to 0.05. The item was easier for the Persian language group than for the Kurdish.

The difficulty difference value for item 12 between the Persian and Kurdish language groups was 0.201 which was statistically significant ($\alpha = .05$). The item was easier for the Kurdish language group than for the Persian.

7.0 DISCUSSION

The result of this study shows that there was bias in universities entrance examination items. One of the methods used to evaluate item bias is by content analyzing the items

which were identified as biased based on statistical methods. In this study, the biased items were content analyzed after specifying for different groups by fifty experienced Persian literature teachers. Their views were as follows.

Items 9 and 10 which favour the female group than the male were based on poems and the contents were relevant to female emotions and feelings. The teachers stated that since comparison was made with prose and poetry, females which enjoy literary subtlety were better in understanding complicated statements in different phenomena. Based on their experience, they also suggested that females as compared to males were better at memorizing, understanding and interpreting the poems. Probably, this was the reason for the two different performances in the items.

Items 2 and 9 favoured the Turkish language group than the other groups. The contents were about history and literature on the Ghaznavid Dynasty and was originally Turkish. The probability is that the Turkish students were more knowledgeable on the issue which relates to their culture and history. The subjects were in their curriculum and as such they get better results.

Item 7 favoured the Persian language group compared with others. The content of this item was about grammatical point, and since the medium of instruction and the mother tongue of this group was the same, it follows that they achieved better results than the other language groups.

Item 12 which favoured the Kurdish language group was about the history of literature It was about Victor Hugo, the French novelist and poet who lived in the 19th century. The item did not mention directly the name of Victor Hugo but the social, political, timely circumstances and his works were mentioned. Kurdish students in Iran differed in many ways from the other Iranians regions in cultural, social and religious aspects. The role of Hugo in criticizing the French situation (before the revolution, 1987), the description of the poor people's life, and the political battles against central government were similar to the life of the Kurdish people (before 1978) with the French before their revolution. Victor Hugo was well-known among Kurdish people where his works were routinely performed in theatres in different parts of Kurdistan. Due to this circunstances, the Kurdish group achieved better results than the others.

Biased analysis was increasingly practiced to detect items which are not comparable due to language and cultural differences. Researchers who investigated the psychometric characteristics of tests pointed out that the level of bias in tests which were translated from one language or culture to another were usually high. For example, Gierl, Rogers and Kinger 1999, (quoted by Judow and Ekerman, 2000) reported that 26 out of 50 items (51%) in students' social studies achievement test at sixth grade in Canada that have been translated from English to French showed a DIF. Ercikan (1999) concluded that there was bias in 58 out of 140 TIMMS science items (41%) when comparison was made between English and French students. Allaouf *et al.,* (1999) mentioned that 42 verbal items out of 125 (34%) in a psychometric entrance test were biased when comparison made between Russian and Israeli subjects (ibid). The items of TIMMS

science in Australia were analyzed to review the differences between students' performances (girls and boys). The content included items on geology, biology, physics, chemistry, natural resources and sciences' nature. The item analysis was conducted using the classical test theory and item response theory. The findings showed that some items were easier for a certain group than the others. The biology items were easier for female students than the males while the items of physics were easier for males than females.

Beak (1997) investigated item bias in entrance examination to universities of Iran for the biology subtest for male and female groups, Persian and Turkish groups and Muslim and non-Muslim groups using the three methods of chi-quire, one parameter and three parameters models of item response theory. He concluded that two items in male and female groups and one item in Muslim and non-Muslim groups showed bias. Based on the findings of this study, it is suggested that the Assessment of Education Organization should evaluate the items that show DIF for different Iranian groups. It is also suggested that items with high probability of bias should be continuously evaluated and decisions make to reduce those in subsequent years.

REFERENCES

Axman, R. C. 1990. *Gender Bias and Fairness*. Practical Assessment, Research & Evaluation. Vol. 2(3).

Beak. M. 1997. *Investigation of Item Bias in Entrance Examination to Universities of Iran in Subtest of Biology for Male and Female Groups, Persian and Turkish Groups and Muslim and non-Muslim Groups*. Thesis (M.A) in Allameh Tabatabai University.

Farzad. V. 1992. Methods of Evaluated Item Bias. *Journal of Educational Research Institute.* Teacher Training University. No; 1 and 2.

Gierl, Mark, *et al*. 1999. *Gender Differential Item Functioning in Mathematics and Science: Prevalence and Policy Implications.* Centre for research in Applied Measurement and Evaluation University of Alberta. Paper presented at the Symposium entitled improving Large-Scale Assessment in Education at the Annual Meeting of the Canadian Society for the Study of Education, Sherbrook, Quebec, CANADA.

Lotfabady. H. 1996. Assessment and Measurement in Education and Psychology. Second Edition. Samt Press.

Nhuan, L. v. 1999. *Identifying Differential Item Functioning on the NELS: 88 History Achievements Test*. Centre for the Study of Evaluation National Centre for Research on Evaluation, Standard, and student Testing Graduate School of Education & Information Studies University of California, Los Angeles, CA 90095-1522(310) 206-1532.

Saif, A. 1998. *Educational Measurement, Assessment and Evaluation.* Second Edition. Tehran Duran Press.

Shepard, L. A., G. Camilli & D. M. Willams. 1985. Validity of Approximation Techniques for Detecting Item Bias. *Journal of Education Measurement*. Vol. 22(2).

Zimowski, M., E. Muraki, R. Mislevy & D. Bock. 2003. *Belong-MG.* Scientific Software International.