Full Length Research Paper

DIF analysis across genders for reading comprehension part of English language achievement exam as a foreign language

Tuncay Öğretmen

Faculty of Education, University of Ege, Bornova, İzmir, Turkey.

Received 09 May, 2015; Accepted 1 June, 2015

The purpose of this study is to carry out differential item functioning (DIF) analysis for content areas of a reading comprehension subtest using four area indices within Item Response Theory (IRT) framework. The differences in the magnitudes of the area indices were compared based on the subject areas. The DIF analysis was carried out across gender groups only. The item level data of the English reading comprehension subtest were gathered from English Language Achievement exam done in School of Foreign Languages, Ege University, Turkey, in 2013. A sample of 2,117 examinees (1,011 males and 1,116 females) was randomly selected. For the DIF analysis, (a) an IRT model for the item characteristic curves was specified; (b) model-data-fit was investigated for the selected IRT model; (c) Item characteristic curves were separately computed for each group on a common scale; finally, (d) indices indicating the degree of DIF on each item were computed. The results of the study indicated that both un-weighted and weighted area indices showed non-uniformity in DIF in the item characteristic curves in reading comprehension subtest in most cases. A significant correlation was observed between un-weighted and weighted area indices.

Key words: Item characteristic curve, item bias, differential item functioning (DIF).

INTRODUCTION

A number of researchers have indicated that there are gender differences in achievement of verbal ability on items requiring inference or application, particularly for science-related content (Lawrence and Curley, 1989; Lawrence et al., 1988). For instance, males are likely to perform better than their female counterparts on verbal items related to natural sciences or technical content, whereas females are more successful than males on items in the social sciences, arts, or humanities. In

addition, based on the results of the several studies, it can be concluded that reading comprehension items related to technical aspects of science is gender based (Lawrence and Curley, 1989; Carlton and Harris, 1992; Curley and Schmitt, 1993). In addition, some studies reported gender differences in content categories of structure, vocabulary, and reading in second or foreign language testing (Bügel and Buunk, 1996; Chavez, 2001; Ryan and Bachman, 1992; Gafni, 1991; Brantmeier, 2001,

E-mail: togretmen@gmail.com, tuncay.ogretmen@ege.edu.tr.

Authors agree that this article remain permanently open access under the terms of the [Creative Commons Attribution License 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

2003; Payne and Lynn 2011). Payne and Lynn (2011) showed that females performed significantly better than males in second language reading comprehension when they were matched in all these variables. This suggests that females have a stronger module for second language processing than males do. Moreover, Pae (2004b) analyzed and showed that gender differences in reading comprehension items classified as logical inference were highly likely to favor males. As stated by Pae (2004a, 2012), conducting gender DIF studies of language tests will encourage language researchers to analyze various components that seem to contribute to the apparent gender difference in language tests. This will thereby provide valuable information for classroom teachers and their students, curriculum design of the educators, students' assessment of foreign language instructors, and ultimately administrators' decisions on language policy.

For reasons explained above, identifying test items which show differential item functioning in favor of any gender group- is an important issue. This is more so especially in the selection and placement of decisions for the purpose of allowing equal opportunity for persons who have equal ability. These persons may belong to a disadvantaged group, which is not taken into consideration in most cases. Since tests are increasingly used as devices for evaluation and placement, the test constructors must make every effort to remove any possible bias from the items. Statistical differential item functioning analysis techniques are frequently used to detect items that include DIF in a test or in a pool of questions in order to increase the test validity. Traditionally, item analysis procedures have been used to determine empirically whether items function as intended or not. Differential item functioning analysis detection techniques serve the same purpose but specifically focus on the validity of items for different subgroups that take the test (Ironson and Craig, 1982; Shepard, Camilli and Williams, 1985; Maller, 2001). On the other hand, detection of DIF is one of the most important issues in test development and evaluation efforts. There have been many attempts made to minimize the DIF in test by the use of several statistical techniques. If the test is used for heterogeneous groups, studying item bias or DIF becomes more important and questions arise as follows:

1. Is the test free of DIF across different gender groups?
2. Is the test free of DIF across different SES groups?
3. Is the test free of DIF across different regions of the country?
4. Is the test free of DIF across different school types?
5. It is obvious that test items should not favor one of the groups in the population tested.

Various methods that have been proposed for detecting DIF in tests include the Classical Test Theory (CTT) techniques and Item Response Theory (IRT) techniques, generally (Ironson and Craig, 1982; Hambleton and Rogers, 1988; Seong and Subkoviak, 1987; Lim and Drasgow, 1990). Comparison of (p) values (the mean item difficulty level) and item discrimination parameters (r) (that is, the extent at which an item is able to differentiate among individuals possessing different levels of the ability or trait underlying the test or scale), which are based on CTT, confounds DIF and true sub-population differences in the trait. This is because this may lead to a high rate of incorrect conclusion that items exhibit DIF when they truly do not (Shepard et al., 1985; Lim and Drasgow, 1990). Furthermore, significant differences in p values may be a result of psychometrically desirable item discrimination power rather than real DIF. Finally, the use of p values may lead to a large number of type II errors by failing to detect items that truly function differently across groups (Lim and Drasgow, 1990). The other method that falls in the CTT techniques is chi-square approach (Scheuneman, 1987). This method assumes that examinees from different groups but with roughly equivalent total test scores are expected to have equal probabilities of success on the item.

A theoretically more appropriate framework for the study of DIF, which avoids the serious deficiencies inherent in CTT approaches, could be found in item response theory (IRT) (Hambleton and Rogers, 1988; Hambleton et al., 1991; Shepard et al., 1984; Devine and Raju, 1982). Their close connection to the most widely accepted definition of item bias has caused IRT-based methods to become popular and some researchers consider these methods "theoretically preferred" (Shepard et al., 1985; Lim and Drasgow, 1990). This definition states that an item is biased or shows DIF if examinees of the same ability but different sub-groups do not have the same probability of correct response to the item (Adams and Rowe, 1988; Lim and Drasgow, 1990; Mellenbergh, 1989; Hambleton et al., 1991; Raju et al., 1993). Central to IRT is a functional relationship between the probability $P_i(\theta)$, that individuals with specific ability θ will respond correctly to an item, and certain item characteristics, or item parameters (Lim and Drasgow, 1990). IRT models often use an item discrimination parameter (a), an item difficulty parameter (b), and a pseudo chance level (c). The graph of $P_i(\theta)$ as a function of θ is known as an item characteristic curve (ICC). Related to the DIF research, the most important advantage of IRT over CTT is the sub-population-invariant property of the ICC and item parameters. Analogous to the coefficients of a simple linear regression equation, these

parameters are invariant when sub-population with different θ distributions is formed. Consequently, IRT-based item parameters do not confound DIF with sub-population differences in ability (Lim and Drasgow, 1990). On the other hand, from an IRT perspective, DIF exists when individuals from different subgroups, but possessing identical levels of the latent trait (θ), have unequal probabilities of correctly answering an item. For an item to be labelled as measuring identically across groups, ICCs, item parameters must be identical (within sampling error) across different sub-populations (Lim and Drasgow, 1990). Thus, the study of DIF within an IRT framework is a matter of comparing the item characteristic curves (ICCs) for the two sub-groups (that is male vs. female) (Hambleton and Swaminathan, 1985; Mellenbergh, 1989).

In this study, ICCs were compared across gender groups to provide evidence on whether one of the groups had an advantage in syntax, connecting and synthesizing, vocabulary, extracting explicit information questions in the content areas of reading comprehension subtest of the English Language Achievement exam of School of Foreign Languages at Ege University, Turkey. More specifically, the present study was designed to accomplish the following purposes: (i.) to examine the direction of DIF for each test item in the reading comprehension subtest across males versus females, (ii.) to compare the magnitudes of the DIF indices between syntax, connecting and synthesizing, vocabulary, extracting explicit information items in the reading comprehension subtest and (iii.) to compare the different area indices used in the DIF analysis.

METHODOLOGY

Subjects

The item level data of the English Reading Comprehension subtest were gathered from English Language Achievement Exam of School of Foreign Languages at Ege University, Turkey, in 2013. A sample of 2,117 examinees (1,011 males and 1,116 females) was randomly selected from all examinee population. About 52.7% were females, and their ages ranged from 17 to 19.

Instrument

The English Reading Comprehension Achievement subtest consists of a total of 40 items. It focuses on the measurement of the students' achievements for understanding and evaluating what is read in English. The items subtest can be classified into four types: (a) syntax, (b) connecting and synthesizing, (c) vocabulary, (d) extracting explicit information. All of the 40 English Reading Comprehension items are based on a multiple choice format with four item options. Reliability estimation for the sample was

assessed using the KR-20 Coefficient. KR-20 Coefficient for the 40 English Reading Comprehension items was 0.82. Moreover, the factor structure of the English Reading Comprehension subtest was examined in order to know if the data met the unidimensionality assumption of the IRT models. A scree test indicated that there was a sharp decrease from the first eigenvalue to the second in the total sample (5.93; 1.32; 1.11; 1.03, respectively). The structure observed in the factor analyses indicates that the English Reading Comprehension subtest includes a strong primary factor in general. This finding constitutes evidence for the claim that the data on the English Reading Comprehension subtest met the unidimensionality assumption of the IRT scaling.

DIF analysis procedure

The following steps were followed in DIF analysis:

1. A model was specified for the item characteristic curve,
2. Model-data-fit was investigated for the selected IRT model,
3. ICCs were separately computed on a common scale for each group,
4. Indices were computed indicating the degree of DIF on each item (Mellenbergh, 1989).

In the present study, IRT DIF area indices were used. A major advantage of this method is that the various ICC theory models describe the item parameters independently of the samples used for estimation (Devine and Raju, 1982). This means that the item parameters are independent of the distributional characteristics of the sample. Therefore, under any IRT model, parameters from different samples should be equal. Items whose parameters are notably unequal violate the assumptions of the model, and consequently are said to be biased because they may be measuring something different for a particular group (Adams and Rowe, 1988).

DIF indices can be classified as weighted and un-weighted. If the number of examinees in two groups is not considered, then it is un-weighted. In order to see the direction of the DIF, the signed area index is used. If un-weighted area index is used, then it could be signed or unsigned. In the unsigned area index the DIF appears, but the direction cannot be seen.

The un-weighted unsigned area index is:

$$B_{\text{unsigned}} = \int_p^q |P_{1i}(\theta) - P_{2i}(\theta)| d\theta$$

where $P_{1i}(\theta)$ and $P_{2i}(\theta)$ are the item characteristic curves in the first and second groups and p and q are the range of abilities of interest.

The un-weighted signed area index is:

$$B_{\text{signed}} = \int_p^q \{P_{1i}(\theta) - P_{2i}(\theta)\} d\theta$$

The signed index shows the direction of DIF. If $P_{1i}(\theta)$ is usually larger than $P_{2i}(\theta)$, the measure is positive, indicating that the item shows DIF in favor of the first group. A negative index implies the DIF is in favor of the second group (Mellenbergh, 1989).

If the two groups are compared with different numbers of examinees, then the area between the two ICCs is weighted as Linn et al. proposed (Cited in Shepard et al., 1985). When the index is weighted, it gives more weight to parts of the ability scale where

most data are concentrated and where the difference in two probabilities at θ_j has a small variance. So, if weighted area index is used, then it could be signed or unsigned.

The weighted unsigned index used in the present study is:

$$B_{\text{unsigned}} = \sum_{j=1}^N \left\{ P_{1i}(\theta_j) - P_{2i}(\theta_j) \right\}^2 / N s_i^2(\theta_j)$$

where; N is the total number of examinees in the two groups and $s_i(\theta_j)$ is the variance error of the difference in the two probabilities at θ_j .

The corresponding signed index is:

$$B_{\text{signed}} = \sum_{j=1}^N \left\{ P_{1i}(\theta_j) - P_{2i}(\theta_j) \right\} \left| P_{1i}(\theta_j) - P_{2i}(\theta_j) \right| / N s_i^2(\theta_j)$$

The interpretation of area indices should be based on non-uniform and uniform DIF. The shape of the item characteristics curves can be the same in two groups when the curves do not coincide, as in the cases when the item is more difficult in male group than in the female group. This is called uniform DIF. It is also possible to have ICCs crossing each other on the ability continuum, which may indicate the non-uniform DIF on the item. In this case ICCs cross ability point $\theta = 0$. For abilities below 0, the item is more difficult in female group and for abilities above 0 the item is more difficult in male group. In the signed area indices, if there is non-uniform DIF on the item, the areas below and above the crossed points can cancel each other and DIF may not be identified even though the ICCs do not coincide. Therefore, un-weighted (signed and unsigned) area indices were considered in order to obtain evidence about the uniformity and non-uniformity of the DIF in the comparisons.

Kruskal-Wallis test was used to evaluate the magnitudes of the area indices in the reading comprehension subject areas of syntax, connecting and synthesizing, vocabulary, extracting explicit information items. To compare the area indices among the subject areas, correlation coefficients were used.

In the first step of the analysis, the fit of one, two and three parameter IRT models was calibrated in each of the gender groups. As a result of these analyses, it was decided to use the two-parameter Birnbaum model (1968) in the ICC comparisons since it served better overall fit statistics compared to the other models.

For the DIF analysis, two package programs were used. The BILOG computer program (Mislevy and Bock, 1986) with Bayesian modal estimation was used for test calibration. For the DIF analysis, a package program Calcbias developed by Oort (1992) was used.

RESULTS

Before the DIF analysis, the reading comprehension subtest was rescaled in order to equate the mean θ to 0 and the standard deviation to 1 across the gender groups for the purpose of establishing a comparable unit for the DIF analysis. Therefore, the comparisons of ICCs were made in the form of female versus male. Four area indices were calculated for each subject area of reading

comprehension subtest as unweighted signed area "SA", unweighted unsigned area "USA", weighted signed area "WSA" and weighted unsigned area "WUSA" indices. The signed area indices indicated the direction of DIF. If the area index was positive in value, it was interpreted as DIF against the female group. A negative index implies the DIF is against the male group. In order to determine whether there are significant mean rank differences among the magnitudes of the area indices across different subject areas of the items with respect to gender groups, the Kruskal-Wallis non-parametric test was carried out. The significance levels of the tests are all evaluated as $p=0.05$. The results of the analyses were presented separately for each subtest.

After the first scaling, 40 reading comprehension items were rescaled for each sex, in order to compute the ICCs separately for each group on a common scale. Finally, indices were calculated indicating the degree of DIF on the items.

In the reading comprehension subtest, 4 items are in the syntax, 17 items are in the connecting and synthesizing, 13 items are in the vocabulary and 6 items are in the extracting explicit information sections with respect to subject areas. The area indices with respect to the gender groups are given in Table 1.

When the values of the un-weighted indices are evaluated there are 4 indices (10.0%) identified as jump outs in the un-weighted signed area (SA), but there are 12 indices (30.0%) identified as jump outs in un-weighted unsigned area (USA). However, when weighted indices are considered, there are no jump out indices observed in either WSA or WUSA.

On the other hand, when both un-weighted and weighted area indices (SA and USA) were compared, it was seen that the values of the unsigned area indices are larger than the signed area indices. This is the evidence of the existence of non-uniformity in the ICCs across the gender groups.

As it is shown in Table 1, DIF has been observed with respect to both weighted and un-weighted area indices. There are 21 items (52.5%) which showed DIF against males according to un-weighted signed area (SA) indices. On the other hand, 13 of 40 items (32.5%) indicate that DIF is against males according to weighted signed area indices (WSA).

As indicated in Table 1, 19 of 40 items (47.5%) showed DIF is against females with respect to the un-weighted signed area indices (SA). However, 27 items (67.5%) indicated DIF is against females with respect to the weighted signed area indices (WSA). On the other hand, according to the weighted area indices there are more items, which showed DIF in favor of males.

Table 2 points out the DIF comparisons among the

Table 1. Area indices in female versus male comparisons of reading comprehension subtest items.

Item no.	Content areas	Un-weighted		Weighted		DIF in favor of Gender
		SA	USA	WSA	WUSA	
1	connecting and synthesizing	0,4240	0,4540	0,0515	0,0515	Male
2	connecting and synthesizing	0,2960	0,4147	0,0532	0,0535	Male
3	connecting and synthesizing	-0,8980	1,7602	-0,0087	0,0087	Female
4	connecting and synthesizing	0,0940	1,4422	0,0162	0,0313	Male
5	connecting and synthesizing	0,0760	0,0898	0,0191	0,0191	Male
6	vocabulary	-0,7500	1,4595	0,0255	0,0390	Female
7	vocabulary	0,3160	0,4266	0,0748	0,0748	Male
8	vocabulary	-0,0370	0,6618	0,0104	0,0347	Female
9	vocabulary	0,0790	0,0790	0,0171	0,0171	Male
10	connecting and synthesizing	-0,0280	0,4798	0,0509	0,0576	Female
11	extracting explicit information	-0,0730	0,1919	-0,0037	0,0205	Female
12	connecting and synthesizing	-0,4180	0,4750	-0,0221	0,0345	Female
13	connecting and synthesizing	-0,0170	0,1512	-0,0083	0,0136	Female
14	extracting explicit information	-0,1630	0,2715	0,0069	0,0152	Female
15	vocabulary	-0,1450	0,2547	-0,0304	0,0304	Female
16	vocabulary	0,0170	0,1736	0,0051	0,0117	Male
17	vocabulary	0,1330	0,1904	0,0373	0,0379	Male
18	vocabulary	0,0630	0,1698	0,0130	0,0191	Male
19	syntax	0,1390	0,1391	0,0346	0,0346	Male
20	syntax	-1,2210	1,8292	-0,0056	0,0089	Female
21	syntax	-0,0460	0,1414	0,0009	0,0044	Female
22	syntax	1,3580	1,6190	0,0087	0,0172	Male
23	vocabulary	-0,0190	0,1020	0,0057	0,0121	Female
24	vocabulary	0,3300	1,5862	0,0295	0,0371	Male
25	vocabulary	-0,0940	0,9995	0,0140	0,0141	Female
26	connecting and synthesizing	0,2390	0,8067	0,0163	0,0297	Male
27	extracting explicit information	-0,0030	0,2643	-0,0095	0,0223	Female
28	connecting and synthesizing	0,1370	0,2562	0,0145	0,0181	Male
29	connecting and synthesizing	0,3020	0,6061	0,0318	0,0364	Male
30	connecting and synthesizing	-0,0730	0,2007	-0,0069	0,0219	Female
31	extracting explicit information	0,1050	0,1891	0,0033	0,0048	Male
32	extracting explicit information	-0,2240	0,2663	-0,0202	0,0202	Female
33	vocabulary	-0,0380	0,2095	0,0008	0,0009	Female
34	vocabulary	0,3000	0,5012	0,0359	0,0374	Male
35	connecting and synthesizing	-0,1490	0,6107	-0,0009	0,0156	Female
36	connecting and synthesizing	-0,0600	0,0659	-0,0229	0,0229	Female
37	extracting explicit information	-0,1640	0,2424	-0,0173	0,0227	Female
38	connecting and synthesizing	-0,0600	0,0710	-0,0132	0,0134	Female
39	connecting and synthesizing	0,0670	0,0670	0,0216	0,0216	Male
40	connecting and synthesizing	0,2130	0,4864	0,0114	0,0115	Male

Table 2. Percentages of items indicated DIF in favor of gender groups with respect to subject areas of reading comprehension subtest.

		Syntax		Connecting and synthesizing		Vocabulary		Extracting explicit information	
Sex		SA	WSA	SA	WSA	SA	WSA	SA	WSA
Female	# of Item	2	1	8	7	6	1	3	3
	% of Item	50.00	25.00	47.06	41.18	46.15	7.70	50.00	50.00
Male	# of Item	2	3	9	10	7	12	3	3
	% of Item	50.00	75.00	52.94	58.82	53.85	92.30	50.00	50.00

subject areas with respect to the un-weighted signed area indices and weighted signed area indices across gender groups. 2 of 4 items (50.0%) showed DIF in favor of the male group in the syntax section with respect to un-weighted SA indices, whereas 3 items (75.0%) showed DIF in favor of the male group in WSA indices. 9 of 17 items (52.94%) showed DIF in favor of the male group, whereas 10 items (58.82%) showed DIF in favor of the same group in the connecting and synthesizing section with respect to WSA indices. On the other hand, when the female group is considered for this section with respect to the SA there are 8 items (47.06%) showing DIF in favor of this group, whereas 7 items (41.18%) showed DIF in favor of the same group. As for the vocabulary section, 6 items (46.15%) showed DIF in favor of females with respect to the un-weighted signed area (SA) indices; in the weighted signed area (WSA) indices only 1 item (7.69%) indicated DIF in favor of the same group (Table 1).

On the other hand, 5 items (83.33%) showed DIF in favor of females according to the un-weighted signed area (SA) indices with respect to extracting explicit information section. For the weighted signed area (WSA) indices, 13 items (50.0%) showed DIF in favor of the female group with respect to vocabulary section (Table 2).

In order to compare the magnitudes of the area indices with respect to subject areas of the reading comprehension subtest items, the non-parametric Kruskal-Wallis test was used.

According to the results, no significant mean rank differences have been observed among the area indices across the items in the subject areas of syntax, connecting and synthesizing, vocabulary and extracting explicit information.

The correlations among the four area indices in the Reading Comprehension subtest comparisons indicated that there were two significant correlations (Table 3).

The first one is observed between the un-weighted signed area (SA) indices and the weighted signed area

Table 3. Correlation among the area indices in gender comparison.

	SA	USA	WSA
USA	-.191		
WSA	.407*	.085	
WUSA	.251	.065	.710*

*Significant at $p < 0.01$.

(WSA) indices since the correlation coefficient, $r = .407$ is significant at $p = .000$. The second correlation was observed between the weighted signed (WSA) indices and the weighted unsigned (WUSA) area indices which produced $r = .71$ that is significant at $p = .000$.

DISCUSSION AND CONCLUSION

When the DIF analysis of the reading comprehension subtest is considered, the items do not function as intended since they are not free from DIF. In the subtest, most of the items (21 items) favored the female group with respect to the un-weighted signed area indices. In contrast, with respect to the weighted signed (WSA) area indices, 27 of the 40 cases indicated DIF in favor of the male group.

When the mean theta levels obtained for each group are compared, a difference is observed across the gender groups that accounts for the weighting of the ICCs in the DIF analysis. For this reason, considering the weighted area indices might be more reasonable for the Turkish subtest. Therefore, if the weighted indices are considered, males have a greater advantage in responding to the reading comprehension items. The result of the study also indicated that the differences between the magnitudes of the un-weighted SA and the un-weighted USA indices showed non-uniform DIF in almost all of the cases when ICCs were compared. This finding implies that the students from different sub-groups

with identical levels of ability have unequal probability of answering the items correctly. In this subtest, jump outs were observed only in the un-weighted indices.

Analysis also revealed that there is a significant correlation observed between the un-weighted SA indices and the weighted signed area WSA indices in the data set of the reading comprehension subtest. However, it is not a high relationship. The correlation coefficient of $r=.407$ is significant at $p=.000$. But it is not surprising to obtain low correlation among the different area indices because there are some differences in the calculations of the four area indices. Furthermore, the weighted area indices (WSA and WUSA) have a significant and high relationship among themselves. When the different DIF indices were considered, the opposite results were obtained. For this reason, there was a problem in deciding which area indices were more meaningful in the DIF analysis. If the mean theta level was equal across ICCs, it should be weighted. Otherwise, the un-weighted indices might be useful.

With respect to the signed area (SA) indices, the female group had an advantage in responding comprehension items. However, with respect to the weighted signed area (WSA) indices, the male group seemed to have a greater advantage. In the subtest, the un-weighted indices indicated non-uniformed DIF for almost all of the items in the ICCs compared. In this subtest jump outs were observed only in the un-weighted indices (SA and USA) also. This implies that the students from different subgroups with identical levels of ability have unequal probability of answering the items correctly as indicated by Adams and Rowe (1988), Lim and Drasgow (1990), Mellenbergh (1989), Hambleton et al. (1991) and Raju et al. (1993).

Further, when assessing reading comprehension as a second language, instructors should be more sensitive to the gender differences in the performance of specific reading comprehension items that favor one gender group against another, which will contribute to a fair assessment of reading comprehension as indicated by Pae (2004b). On the other hand, the design of a reading comprehension curriculum should include the information about the differential performance by gender, as reported in this study. Applying this way, teachers can take proactive steps to minimize gender differences in higher levels of thinking skills that are critical to efficient reading comprehension.

RECOMENDATIONS FOR FURTHER RESEARCH

Further DIF studies can be made by including examinees coming from different school types, branches and

curricula. Statistical properties of the subtests should be examined in a more detailed way to analyze the reasons for the DIF observed in the Reading Comprehension subtest and special emphasis should be given to analyze whether selection decisions are affected by DIF or not.

Conflict of Interests

The author has not declared any conflict of interests.

REFERENCES

- Adams RJ, Rowe KJ (1988). Item Bias.in Keeves, J.P.(ed.) Educational research, methodology, and measurement: An international handbook. Oxford: Pergamon Press.
- Birnbaum A (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.K. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Brantmeier C (2001). Second language reading research on passage content and gender: Challenges for the intermediate-level curriculum. *Foreign Language Annals* 34 (4): 325–333.
- Brantmeier C (2003). Beyond linguistic knowledge: Individual differences in second language reading. *Foreign Language Annals* 36(1):33–43.
- Bügel K, Buunk BP (1996). Sex differences in foreign language text comprehension: The role of interests and prior knowledge. *Modern Language J.* 80:15–31.
- Carlton ST, Harris AM (1992). Characteristics associated with Differential Item Functioning on the Scholastic Aptitude Test: Gender and Majority/Minority Group Comparisons. ETS Research Report, 92-64, New Jersey.
- Chavez M (2001). *Gender in the language classroom*. McGraw Hill, New York.
- Curley W, Schmitt AP (1993). Revising SAT-Verbal Items to Eliminate Differential Item Functioning. College Board Report No. 93-2.
- Devine PJ, Raju NS (1982). "Extent of overlap among four item bias methods." *Educ. Psychol. Measure.* 42:1049-1066.
- Gafni N (1991). Differential Item Functioning: Performance by Sex on Reading Comprehension Tests. ERIC Document No. ED 331844.
- Hambleton RK, Rogers HJ (1988). Apr. Detecting biased test items: Comparison of the IRT area and Mantel-Haenszel Methods. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 300 398)
- Hambleton RK, Swaminathan H (1985). *Item Response Theory: Principles and applications*. Boston: Kluwer-Nijhoff Publishing.
- Hambleton RK, Swaminathan H, Rogers HJ (1991). *Fundamentals of item response theory*. London: Sage Publication.
- Ironson GH, Craig R (1982). Item bias techniques when amount of bias is varied and score differences groups are presented. University of South Florida, Tampa. Dept of Psychology. (ERIC Document Reproduction Service No. ED 227 146)
- Lawrence IM, Curley WE (1989). Differential Item Functioning for Males and Females on SAT-Verbal Reading Subscore Items: Follow-up Study. Educational Testing Service Research Report 89-22, New Jersey.
- Lawrence IM, Curley WE, McHale FJ (1988). Differential Item Functioning for Males and Females on SAT Verbal Reading Subscore Items. College Entrance Examination Board Report No. 88-4, New York.
- Lim Rodney G, Drasgow F (1990). "Evaluation of two methods for

- estimating item response theory parameters when assessing differential item functioning." *J. Appl. Psychol.* 75(2):164-174.
- Maller JS (2001) Differential Item Functioning in the THE WISC-III: Item Parameters for Boys and Girls in the National Standardization Sample, *Educ. Psychol. Measure.* 61(5):793-817
- Mellenbergh GJ (1989). "Item bias and item response theory." *International Journal of Educational Research: Applications of Item Response Theory.* 13(2):123-144.
- Mislevy RJ, Bock DR (1986). *PC-BILOG: Item analysis and test scoring with binary logistic models.* Scientific Software Inc.
- Oort F (1992). Computer program which computes area indices. University of Amsterdam. Faculty of Psychology.
- Pae TI (2004a). DIF for examinees with different academic backgrounds. *Language Testing.* 21:53-73.
- Pae TI (2004b). Gender effect on reading comprehension with Korean EFL learners. *System.* 32:265-281.
- Pae TI (2012) Causes of gender DIF on an EFL language test: A multipledata analysis over nine years. *Language Testing* 29(4):533-554
- Payne TW, Lynn R (2011) Sex differences in second language comprehension. *Personality and Individual Differences.* 50(3):434-436
- Raju NS, Drasgow F, Slinde JA (1993). "An empirical comparison of the area methods, Lord's chi-square test, and the Mantel-Haenszel technique for assessing differential item functioning." *Educ. Psychol. Measure.* 53:301-314.
- Ryan K, Bachman LF (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing.* 9(1):12-29.
- Scheuneman JD (1987). "An experimental, exploratory study of causes of bias in test items." *J. Educ. Measurement.* 24(2):97-118.
- Seong Tae-Je, Subkoviak MJ (1987). Apr. A comparative study of recently proposed item bias detection methods. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington. (ERIC Document Reproduction Service No. ED 281 883)
- Shepard LA, Camilli G, Williams DM (1985). "Validity of approximation techniques for detecting item bias." *J. Educ. Measure.* 22(2):77-105.

Appendix A: Examples of test items**PART 3: READING COMPREHENSION**

Choose the option which best completes the meaning of the following paragraphs. Then fill the correct space on your answer sheet.

A. Questions 1-5 (5 points)

1. Water is the simplest compound of hydrogen and oxygen, elements that are essential to all living things. The body is made up of about 70 percent of water, with certain tissues such as the grey matter in the brain containing up to 85 percent. _____. And in return, it transports waste to certain parts such as the liver, kidneys, lungs and skin to be eliminated from the body.

- a. This can result in disorientation, hallucination and eventually death
- b. Water also comprises 80 percent of blood, which takes food to the tissues
- c. Water acts as the medium for all chemical processes that take place in the body
- d. Every day, the body loses around 2 litres of water just through breathing alone

2. Over half of the world's rain forests are in South and Central America. _____.

There are over 40 different types of rain forests, each with its own variety of plant and animal life. The largest rain forest in the world is in South America. It is almost as big as Australia.

- a. All around the world, large areas of rain forest are disappearing
- b. Unless there is a change, there will be no rain forests left by the year 2050
- c. The remainder can be found in Africa, Asia and Australia
- d. People all over the world use rain forest products

3. Gymnastics, one of the nine original sports of the first modern Olympics, is highly competitive and demanding. _____. By doing so, they can ensure fitness, strength and stamina. To become a top-class athlete is a real achievement and requires years of dedication.

- a. As a result, it is essential for would-be gymnasts to start training at a very early age
- b. The best known competitions are the world championships and the Olympics
- c. Gymnasts try to achieve excellence by performing a combination of these difficult moves
- d. Therefore, the precision of every movement is vital as the slightest mistake counts

.....