Journal of Education and Practice
ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)
Vol.8, No.7, 2017

Study of Bias in 2012-Placement Test through Rasch Model in Terms of Gender Variable

Res. Assist. Azmi TURKAN
Yildiz Technical University

Assoc.Prof. Dr. Bayram CETİN
Gazi University

Abstract
Validity and reliability are among the most crucial characteristics of a test. One of the steps to make sure that a test is valid and reliable is to examine the bias in test items. The purpose of this study was to examine the bias in 2012 Placement Test items in terms of gender variable using Rasch Model in Turkey. The sample of this study was 216363 participants that randomly selected out of 1075546 students who took Placement Exam (SBS) in 2012. A stratified sampling was used for this study. Winsteps 3.8.1 package program was used in data analysis. Through statistical procedures it has been established that two items in Maths subtest and one item in Social Sciences subtest, namely three items in total, have Differential Item Functioning (DIF). According to expert opinion regarding the bias in these three items no bias was available in the items. Consequently, it had been established that there was no biased question in Turkish language, Mathematics, Science and Technology and Social Sciences subtests within the 2012 Placement Test and they were valid and reliable.
Keywords: Placement test, Item Bias, Differential Item Functioning, Rasch Model

1. Introduction

Main goal of measurement and assessment in education is to make right and appropriate decisions for individuals. The most important assessment tool is tests while these decisions are taken. Tests need to have some features since they are used to determine skills, achievements, interests and characteristics of individuals. According to Micheel and Karnes (1950) tests should be comprehensive, distinguishing, objective, practical, valid and reliable. Tekin (2000), Yilmaz (1998) and Turgut (1995) argue that tests be practical, valid and reliable. These properties includes being affordable and serving purposes. Apart from these concept of error is also among these features. If an assessment tool is free from errors, the reliability of the scale increases (Turgut & Baykul, 2010; Tekin, 2000; Demircioglu, 2009). Errors depending upon the rater or chance might come up in the scale. In addition to these errors, when the assessment tool offers an advantage to a certain group, in other words has a bias, this impacts its validity and reliability.

Since assessment is used to make some predictions about individuals, it is highly important that tests be valid and reliable. Determining item bias is a way of making a test more valid and reliable. Item bias appears when individuals having same level of skill have different possibilities of answering an item due to their certain characteristics (Zumbo, 1999). It is especially important in exams for picking and placing people into a particular school since it affects the future of people during exams (Bekci, 2007).

Bias is directly relevant to the concept of validity in test points. Therefore, those points should not be in favor of particular group. It seems crucial that studies on bias should be carried out for the exams that have a huge impact on the future of individuals.

When a test becomes in favor of or against a particular group as a result of test features or conditions that are not suitable for the main objective of the test, then this leads to test bias for people of the same level of skill (Zumbo, 1999, Schumacker, 2005; Hambleton & Rogers, 1995). As can be seen in the definition as well, bias is regarded as a systematic error because differing in favor of a particular group or within a particular rule is observed. In this context test bias has a great impact on the validity and reliability of a test (Kristjansson, Alyeswort, Mcdowell & Zumbo, 2005).

Test bias became apparent in IQ tests 1900s demanding ethnic origin and high level of language skill. These tests turned against Afro-American people and it has been established that there is ethnic and cultural bias in tests (Cole & Zieky, 2001). In addition, studies on the bias in standard tests became a topical issue when Binet and Simon introduced the first versions of their own IQ tests in 1905. As a result of this study done on children of working class and upper class the difference between groups became more evident. This cultural bias in IQ tests was spotted later and it was restrained by eliminating the items that reveal social class differences. Psychologist Kenneth Eells was the one who laid down the foundations of contemporary studies on bias with his study in 1951. Focusing on the test items in his study, he examined the IQ differences between various cultural groups. A study of his, namely "An Investigation of Item Bias" is another contemporary study on item bias and it was published in "*Educational and Psychological Measurement",* which is a prestigious magazine on psychometrics. This study tries to detect item bias using Anova technique (Cleary & Hilton, 1968). Although it

has methodological errors, today ANOVA strategy is still frequently used in studies on item bias. Scheuneman came up with another important method for item bias. Scheuneman recommended chi-square method in her study. However, this method has been criticized for its deficiencies. By carrying out a series of studies Berkeley from California University improved Chi-Square method by Scheuneman to detect item bias and made a significant contribution to this statistical method (Osterlind, 1983). These methods were made to make up the shortages in earlier studies and provide more advanced methods. However, it was considered in later studies that not only bias but also genetic and environmental factors may also affect the differences between IQ tests which measure intelligence. In this regard, today studies on bias are carried out to determine the cause of the difference. The first stage of the studies to determine bias is to determine whether the items in the scale contain Differential Item Functioning (DIF).

Differential Item Functioning (DIF) can be defined as statistically determining the difference between the answers to a test. (Yurdugul, 2003; Gierl, Khaliq & Boughton, 1999; Taylor & Lee, 2012). DIF is the state of being more difficult or easier for a particular group between two groups (focal and reference) in terms of all item difficulty levels (Kristjansson, & et al., 2005). DIF is a statistical process as we can see from the definition. First of all, items that contain DIF in subgroups with the same level of skill are spotted (Gok, Kelecioglu & Dogan, 2010; Roever, 2005). In other words, it is statistically detected that items give a certain group advantage. Zumbo (1999) considers that the difference between groups may stem from item impact or bias. Expert opinion indicates whether DIFs stem from assessment procedure (item impact) or from the structure of individuals who make up the group. If the difference stems from the groups in this stage, this DIF is called bias (Camilli & Shepard, 1994; Zumbo, 1999). In other words, DIF is a statistical step of study on bias.

All statistical procedures for determining DIF and bias aim to increase validity and reliability of the tests. In this regard, it seems crucial that studies on bias should be carried out for the exams that have a huge impact on the future of individuals. Many exams are held to pick and place people in Turkey. Students take The Transition to Higher Education Exam (YGS) and Undergraduate Placement Exam (LYS) to study in a university after high school. Apart from these exams, Ministry of National Education held Secondary Education Exam (OKS) to assess students in the 8th grade and place them into high schools till 2007. After 2008, students in the 6th, 7th and 8th grade began to take Placement Exam (SBS) to be able to attend certain leading secondary schools (Anatolian High Schools, Science High Schools, Anatolian Teacher Training High Schools, etc.). After 2010, SBS was gradually abolished in the 6th and 7th grades (MEB, 2011). SBS includes main courses, such as Turkish, Maths, Science and Technology, Social Sciences (Religious Culture and Ethics) and Foreign Language. The goal of these main courses is that students can learn to make interpretations, analysis, critical thinking, problem solution and logical connections. The result is yielded from their success in the exam and grade point average that comes from all courses taken at school (Annual Success Point). Students are placed to the secondary schools according to these results. As these exams may have a great impact on people's lives, it is highly important that they have no bias.

Several methods are used to detect bias. Among them are Logistic Regression (LR), Mantel-Haenszel (MH) and Chi-square. Another one is the method of determining bias according to Rasch model. In Rasch model, each examinee has a skill (β) and each item has a difficulty (δ). If a test contains DIF, item difficulty for the reference group and for the focal group will be different. Rasch analysis makes an estimation about skills of each participant in each group within an equally spaced scale.

Many international studies on bias are available. (Hanna, 1986; Linacre & Wright, 1987; Zwick & Erikcan 1989; Gamer & Engelhard, 1999; Le, 1999; Zenisky, Hambleton & Robin, 2004; Ariffin, Idris & Ishak, 2010; Taylor & Lee, 2012). However, there are fewer studies on bias in Turkey (Kurnaz, 2006; Bekci, 2007; Dogan & Ogretmen, 2008; Gok, Kelecioglu & Dogan, 2010; Kalaycioglu & Kelecioglu, 2011). When we regard that there are limited number of studies on bias, lack of studies to detect bias through Rasch model and the importance of test reliability in Turkey, this study aims to examine bias in subtest items in 2012 Placement Exam through Rasch model in terms of gender. In this context, the following research questions were answered;

1. Do items of subtests in the 2012 Placement Exam (SBS) contain Differential Item Functioning (DIF) in terms of gender according to Rasch model?
2. Are there any bias in items containing Differential Item Functioning (DIF) based on expert opinions?

2. Method

2.1. Research Design

The study aims to determine whether Turkish, Maths, Science and Technology and Social Sciences subtests in 2012 SBS which 8th grade students took has a bias or not. In this study the main goal has been to present a circumstance as it is without making any change. Since it aims to describe a circumstance without changing or impacting it, it is descriptive survey research (Karasar, 2008).

2.2. Participants

Population of the study are all individuals who take SBS-2012. 1075546 participants took the exam. 527978 of them are female and 547568 are male. For analysis of data there was available data regarding all of the population. However, sample has been formed through stratified sampling method by taking %20 of each booklet and gender to prevent problems in the analysis. Stratified sampling is an appropriate sampling method for studies in social sciences, especially in non-homogenous populations (Baykul, 1997). Proportional selection from stratified sampling method has been preferred in this study. This sample has been created by using simple random sampling method. In this method, all elements in the populations have equal and independent chance to be selected (Karasar, 2008).

Statistical information composing the sample of the study has been given in the Table 1 according to booklet and gender type.

Table 1. Student sample distribution according to booklet and gender type

	A Booklet	**B Booklet**	**Total**
Female	53785	52620	106405
Male	56776	53182	109958
Total	110561	105802	216363

It can be seen in the Table 1 that 216363 students make up the sample of the study and they are similarly distributed according to booklet type and gender.

2.3. Data Collection and Process

Items in Turkish, Maths, Science and Technology and Social Sciences subtests in 2012 SBS that 8th grade students were taken into consideration. Turkish test includes 23 items. Science and Technology test and Social Sciences test each include 20 items. Study data has been yielded from answers of the subtests in 2012 SBS. Data used to find an answer to the question in the study has been acquired with the written permission (IRB#62927161-300) of the Ministry of National Education Directorate General of Educational Technologies Department of Measurement and Assessment.

2.4. Data Analysis

Raw data from the Ministry of National Education Directorate General of Educational Technologies Department of Measurement and Assessment has been transformed into a format for analysis with Excel, SPSS-20. For analysis of data SPSS-20, Amos and Winsteps 3.8.1 were used.

In Rasch models ConQuest (Wu, Adams, Wilson & Haldane, 2007), Facets (Linacre, 2009) and Winsteps (Linacre, 2010) are used for DIF analysis (Karami, 2011). Winsteps one of the most frequently used programs for DIF analysis in Rasch model (Wang & Chen, 2005). In Winsteps program we look at DIF Contrast value to detect DIF. According to Lai and Eton (2002) if DIF Contrast value is 0,5 logit value, this is a critical value for likert scales. Furthermore, Pallant and Tennant (2007) consider that if DIF value is below 0,5 logit value in terms of gender, this indicates that the item has no DIF. Negative DIF Contrast value shows that the item is easy for the focal group and it is advantageous for the focal group.

After Differential Item Functioning (DIF) is detected through Winsteps program, expert opinion has been taken to determine the bias. These experts are subtest experts, Guidance and Psychological Counseling experts and measurement and assessment experts. They are 10 experts in total, namely 2 associate professor, 3 assistant professors and 5 research assistants.

3. Findings

Turkish, Maths, Social Sciences and Science and Technology subtests have been examined in terms of normal distribution, uni-dimensionality and local independence of test items to look at whether analysis is suitable for Rasch model. Kurtosis and skewness values have been examined to determine that data shows a normal distribution. For uni-dimensionality DFA fit index values have been looked at. As the results of statistical procedures indicated, it has been established that tests have a normal distribution and one-dimensional structure. Local independence is to define the relations in item set by looking at only one skill. Therefore, if uni-dimensionality assumption is realized, this can be interpreted as realizing local independence as well (Lord, 1980; Hambleton & Swaminathan, 1985). After statistical procedures, it has been established that Turkish and Science-Technology items have no DIF. In this regard, findings about Maths and Social Sciences subtests having DIF have been mentioned in this section.

Findings about 2012-SBS Maths Subtest

Contrast value of Differential Item Functioning (DIF) about 2012-SBS Maths subtest is given in Table 2.

Table 2. Contrast values of differential item functioning (DİF) about maths subtest

Focal Group	DIF Measurement	Reference Group	DIF Measurement	DIF Contrast	T	Item
Male	.07	Female	-.03	.10	8.24	1
Male	-.13	Female	-.39	.26	24.36	2
Male	-.42	Female	-.42	.00	.00	3
Male	-1.37	Female	-1.91	.54	45.94	4
Male	.46	Female	.66	-.20	-13.1	5
Male	.35	Female	.27	.07	5.90	6
Male	-.20	Female	-.16	-.05	-4.43	7
Male	-.21	Female	-.12	-.09	-8.02	8
Male	.37	Female	.39	-.02	-1.83	9
Male	1.69	Female	2.03	-.34	-22.6	10
Male	.13	Female	.13	.00	.00	11
Male	.17	Female	.05	.11	9.73	12
Male	.21	Female	.09	.12	8.97	13
Male	.61	Female	.61	.00	.00	14
Male	-1.19	Female	-1.26	.07	6.09	15
Male	-1.29	Female	-1.21	-.08	-7.31	16
Male	-.05	Female	-.05	.00	.00	17
Male	.90	Female	.90	.00	.00	18
Male	-.99	Female	-.43	-.56	-50.8	19
Male	.87	Female	.94	-.07	-4.43	20

DIF measurement values have been given in the Table 2 according to the answers of participants to the 20 items of Maths subtest in relation to their gender. DIF measurement values of male participants, DIF values of female participants, total DIF measurement value, t values and item numbers have been given.

Table 2 indicates that DIF measurement values of Maths subtest of male students changes between -1.37 and 1.69 logit. The lowest DIF value of male students in Maths subtest is in the 4th item with -1.37 logit and the highest is in the 10th item with 1.69 logit. DIF measurement values of female students in Maths subtest changes between -1.91 and 2.03 logit. The lowest DIF value of female students in Maths subtest is in the 4th item with -1.29 logit and the highest is in the 10th item with 1.69 logit.

When DIF values of Maths subtest are examined in terms of gender, it is indicated that DIF contrast values of items change between -0.56 and 0.54. The lowest DIF value in Maths subtest is in the 19th item with -0.56 logit and the highest is in the 4th item with 0.54 logit. The fact that DIF contrast values are negative values shows that bias is in favor of the focal (first) group. Ordinary value of DIF is between 0.5 and -0.5. Moreover, it is indicated that t value of the items in Maths test is between -50.8 and 45.94. The lowest t value is in the 19th item with -50.8 and the highest is in the 4th item with 45.94.

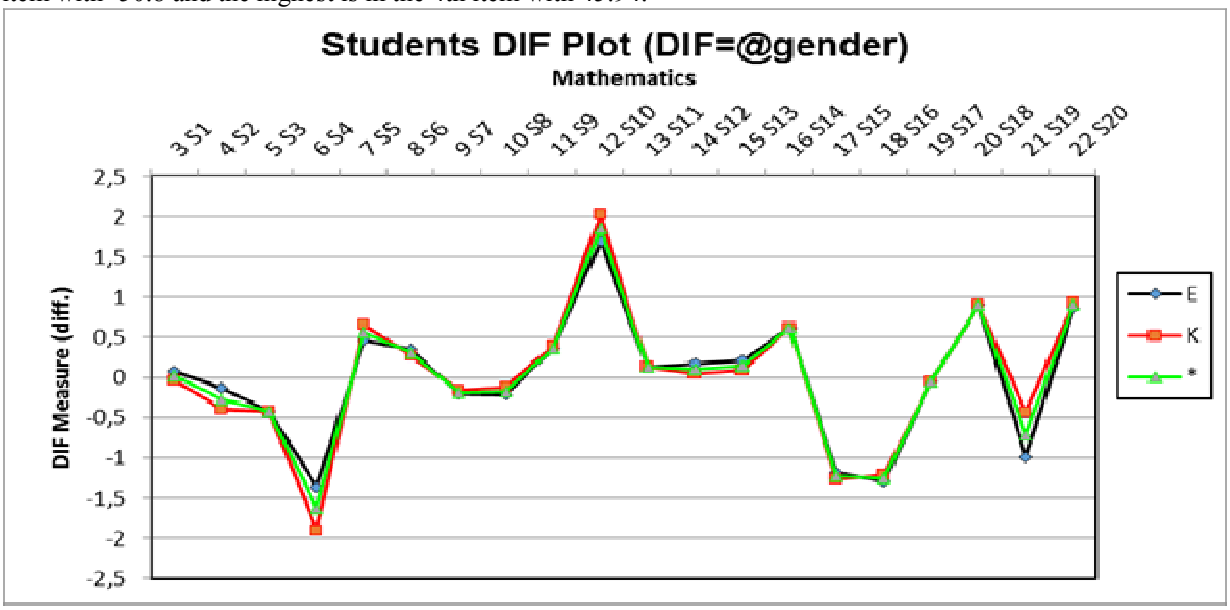

Figure 1. Change of Maths Items In Terms of Gender

It has been indicated how DIF changes in Maths subtest in terms of gender in the Figure 1. The DIF value in the green line shows average DIF value whereas the red line indicates the DIF value of female students in Maths subtest items. The black one, on the other hand, indicates DIF value of male students in Maths subtest items. The fact that the ratio of change between the red and black line is over 0.5 indicates that items have a significant level of DIF. As can be seen in Figure-1, 19th and 4th has a DIF contrast value over 0.5. In all other items it is apparent that there is a change below 0.5. When Table 2 and Figure 1 are examined, it is found out that 19th and 4th items have DIF.

Findings about 2012-SBS Social Sciences Subtest

Contrast Values of Differential Item Functioning (DIF) about Social Sciences Subtest is given in Table 3.

Table 3. Contrast values of differential item functioning (DIF) about social sciences subtest

Focal Group	*DIF Measurement*	*Reference Group*	*DIF Measurement*	*DIF Contrast*	*T*	*Item*
Male	-.89	Female	-.60	-.29	-25.4	1
Male	-.66	Female	-.47	-.19	-16.1	2
Male	.85	Female	1.47	-.62	-57.5	3
Male	2.08	Female	2.00	.08	7.13	4
Male	-.84	Female	-1.07	.23	18.01	5
Male	-.64	Female	-.46	-.19	-16.6	6
Male	-.48	Female	-.48	.00	.00	7
Male	.32	Female	.20	.12	11.12	8
Male	-.87	Female	-1.35	.48	38.38	9
Male	.07	Female	-.02	.10	8.96	10
Male	1.94	Female	1.94	.00	.00	11
Male	.04	Female	.31	-.27	-25.6	12
Male	1.01	Female	1.38	-.37	-34.2	13
Male	-.74	Female	-1.22	.47	36.91	14
Male	-.81	Female	-1.12	.31	25.51	15
Male	-.72	Female	-.99	.27	22.63	16
Male	.68	Female	.50	.18	17.25	17
Male	.70	Female	.70	.00	.00	18
Male	.32	Female	.51	-.20	-19.1	19
Male	-1.21	Female	-1.59	.37	27.84	20

DIF measurement values have been given in the Table 3 according to the answers of participants to the 20 items of Social Sciences subtest in relation to their gender. DIF measurement values of male participants, DIF values of female participants, DIF contrast measurement value, t values and item numbers have been given.

Table 3 indicates that DIF measurement values of Social Sciences subtest of male students changes between -1.21 and 2.08 logit. The lowest DIF value of male students in Social Sciences subtest is in the 20th item with -1.21 logit and the highest is in the 4th item with 2.08 logit. DIF measurement values of female students in Social Sciences subtest changes between -1.59 and 2.00 logit. The lowest DIF value of female students in Social Sciences subtest is in the 20th item with -1.59 logit and the highest is in the 4th item with 2.00 logit.

When DIF values of Social Sciences subtest are examined in terms of gender in the Table-3, it is seen that DIF contrast values of items change between -0.62 and 0.48. The lowest DIF value in Social Sciences subtest is in the 3th item with -0.62 logit and the highest is in the 9th item with 0.48 logit. The fact that DIF contrast values are negative values shows that bias is in favor of the focal group (first). It is expected that DIF measurement value should be between 0.5 logit as absolute value. Moreover, it is indicated that t value of the items in Social Sciences test is between -33.2 and 25.06. The lowest t value is in the 3th item with -57.5 and the highest is in the 9th item with 38.38.

In the Figure 2 below measurement distribution of Differential Item Functioning (DIF) in terms of gender in Social Sciences subtest has been given.

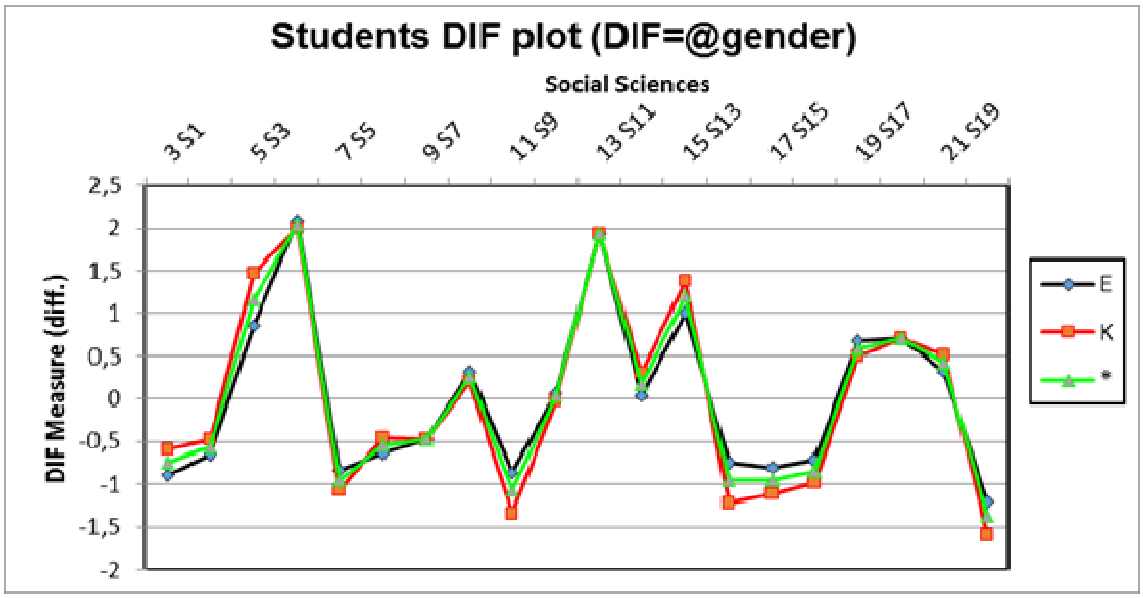

Figure 2. Change of Social Sciences Items In Terms of Gender

It has been indicated how DIF changes in Social Sciences subtest in terms of gender in the Figure 2. The green line shows average DIF value whereas the red line indicates the DIF value of female students in Social Sciences subtest items. The black one, on the other hand, indicates DIF value of male students in Social Sciences subtest items. If the variation between DIF ratios in red and black is over 0.5, this indicates that those items have a significant amount of DIF. It is seen in the figure that only 3th item has a value approximately over 0.5. It is seen in the figure that the difference between all other items is below 0.5. When Table 3 and Figure 2 are examined, it is found out that 3th item has DIF.

Findings Yielded from Expert Opinion about Items Having Differential Item Functioning (DIF):

Through statistical procedures it has been established that two items in Maths subtest and one item in Social Sciences subtest, namely three items in total, have Differential Item Functioning (DIF). Expert opinion has been taken in order to determine whether items with DIF stem from item impact or bias. On 4th item in Maths subtest two of the experts stated that the item is in favor of males. Another expert said it is in favor of females and the other experts said it stems from item impact. Similarly, when expert opinions about the 19th item in Maths subtest, only one participant tended to think that the item may be in favor of males. Other participants stated that the item may stem from *item impact*. When 3th item in Social Sciences subtest is examined, two of the experts claimed that the item may stem from *item bias.* Other experts stated that it stems from *item impact.* According to expert opinion, there is no bias in mathematics and social sciences items.

4. Discussion

This study examined item bias in SBS subtests in terms of gender through Rasch model. Item parameters are independent from the group and group skills de not depend upon the test. Tests have been examined in terms of normal distribution, uni-dimensionality and local independence of test items to look at whether analysis is suitable for Rasch model. Analysis indicated that data is suitable for the analysis of Rasch model.

Contrast Values of DIF in Maths Subtest in 2012 SBS have been examined. The fact that DIF contrast values are negative values shows that bias is in favor of the focal group (first group). While it is expect that DIF value is zero, 0.5 as absolute value is an acceptable value. When results of the analysis regarding Maths subtest are examined, it is seen that some of the items have negative contrast value whereas some others have positive contrast value. When we look at results of the analysis, we see that the item end up in an unexpected value. It is found out that one of these items is in favor of males with a negative value whereas the other is in favor of females with a positive value. Consequently, when we examine Table-2 and Figure-1, we see that 4th and 19th items in Maths subtest are between 0.54 logit ad -0.56 logit and they contain DIF whereas others have none. It has been concluded that these items have no bias according to expert opinion.

When we examine the 4th item in Maths test in 2012-SBS, we see that it is a question that demands geometry skills It has been detected that this item is in favor of females. A study explains that female students are better at mathematics and courses based on verbal language skills in primary school. However, the same study indicates that male students are better at geometry beginning from high school (Amrein & Berliner, 2002). This finding contradicts the study carried out by Zenisky, Hambleton and Robin (2003). Zenisky, Hambleton and Robin (2003) argue that items demanding visual-spatial intelligence, in other words the ones with tables, figures, graphics, are in favor of males. 19th item in Maths test is another item having DIF. This item is a verbal

item and it has been established that it is in favor of males. This finding is compatible with the study by Yurdugul and Askar (2004). The study by Yurdugul and Askar indicates that when mathematics problems are expressed verbally, they are in favor of males (Yurdugul & Askar, 2004). Willingham and Cole (1997) stated that female students are better at especially mathematical operations and measurements whereas male students are better at problem solving. Since the item is related to problem solving, this finding bears similarity to the finding of Willingham and Cole.

Statistical analysis shows that 3th item in Social Sciences subtest in 2012-SBS contains DIF. It has been established that this item is advantageous for males but expert opinion showed that this advantage does not stem from bias. Since the 3th item in Social Sciences test in 2012-SBS is about war, agreement, etc., it is considered to be in favor of males. This finding is compatible with a finding from a study by Kalaycıoglu and Kelecioglu. The study examined whether items in 2005-Student Selection Examination contains DIF and it found out that an item about history in Social Sciences test had DIF. Researchers emphasized that questions about politics and war in History test may become advantageous for males (Kaaycıoglu & Kelecioglu, 2011). Similarly, Zwick and Ercikan concluded in their study that some items in History test may have DIF in favor of males (Zwick & Ercikan, 1989). Le (1999) confirmed that history questions are in favor of males. This finding is compatible with Le's finding.

5. Conclusion

Through statistical procedures it has been established that two items in Maths subtest and one item in Social Sciences subtest, namely three items in total, have Differential Item Functioning (DIF). Expert opinion has been taken in order to determine whether items with DIF stem from bias or item impact. Expert opinion indicates that items does not stem from bias but from item impact. In this regard, it has been established that 2012 Placement Exam is a valid and reliable test in terms of gender bias.

6. Recommendations

This study examines the gender bias in Turkish, Maths, Science and Social Sciences subtests in 2012 Placement Exam (SBS). It is beneficial to examine the local bias in nationwide exams.

It will contribute a lot to do studies on bias in undergraduate entrance exams in terms of school type and graduation status of students.

The body of literature indicates that many methods have been used to detect DIF according to Classical Reaction Theory and Item Reaction Theory. In body of literature several methods used to determine DIF have been compared. Items having DIF have been detected through Rasch model in this study. It will beneficial to compare Rasch model and other methods to determine DIF.

Contributions

This paper has been produced from first author's Master thesis.

REFERENCES

1. Ariffin, S. R., Idris, R. & Ishak, N. M. (2010). Differential item functioning in Malaysian generic skills instrument (MyGSI). *Jurnal Pendidikan Malaysia*, *35*(1), 1-10.
2. Amrein, A. L. & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10, 18.* 05.04.2013 tarihinde*, http://epaa.asu.edu/epaa/v10n18/ adresinden alınmıştır.*
3. Baykul, Y. (1997). *İstatistik Metotlar ve Uygulamalar.* Ankara: Anı Yayıncılık, 2.Baskı.
4. Bekci, B. (2007). Ortaöğretim Kurumları Öğrenci Seçme Sınavının Değişen Madde Fonksiyonlarının Cinsiyete ve Okul Türüne Göre İncelenmesi. *Yayınlanmamış Yüksek lisans Tezi.* Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü.
5. Camilli, G. & Shepard L. A. (1994). *Methods for identifying biased test items*. London: Sage Publications.
6. Cleary, T. & Hilton, T. L. (1968). An investigation of item bias. *Educational and Psychological Measurement*.
7. Cole, N. S. & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement*, 38(4), 369-382.
8. Demircioglu, G. (2009). Geçerlik ve Güvenirlik. *Ölçme ve Değerlendirme*. Karip. E. (Ed.) Pegem Yayıncılık, Ankara.
9. Dogan, N., & Ogretmen, T. (2010). Değişen Madde Fonksiyonunu Belirlemede Mantel - Haenszel, Ki - Kare ve Lojistik Regresyon Tekniklerinin Karşılaştırılması. *Eğitim ve Bilim*, *33*(148), 100-112.
10. Gamer, M. & Engelhard Jr, G. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education*, *12*(1), 29-51.

11. Gierl, M., Khaliq, S. N. & Boughton, K. (1999). Gender differential item functioning in mathematics and science: Prevalence and policy implications. In*Annual Meeting of the Canadian Society for the Study of Education, Sherbrooke*.
12. Gok, B., Kelecioglu, H., & Dogan, N. (2010). Değişen madde fonksiyonunu belirlemede Mantel–Haenszel ve Lojistik Regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim*, *35*(156), 3-16.
13. Hambleton, R. K., & Rodgers, J. (1995). *Item bias review*. ERIC Clearinghouse on Assessmentand Evaluation, the Catholic University of America, Department of Education.
14. Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer Academic Publishers.
15. Hanna, G. (1986). "Sex Differences in The Mathematics Achievement Of 8th Graders in Ontorio." *Journal for Reaserch in Mathematic Education*Vol. 17, p. 231-237
16. Kalaycıoglu, D. B. & Kelecioglu, H. (2011). Öğrenci Seçme Sınavı'nın Madde Yanlılığı Açısından İncelenmesi. *Eğitim ve Bilim*, *36*(161), 3-13.
17. Karami, H. (2011). An Introduction to Differential Item Functioning. *The International Journal of Educational and Psychological Assessment September* 2012, Vol. 11(2).
18. Karasar, N. (2008). Bilimsel Araştırma Yöntemleri. *Ankara, Nobel Yayıncılık*.
19. Kristjansson, E., Aylesworth, R., Mcdowell, I., and Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement*, *65*(6), 935-953.
20. Kurnaz, B. (2006). Peabody Resim Kelime Testinin Madde Yanlılığı Açısından İncelenmesi. *Yayınlanmamış Yüksek Lisans Tezi.* Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü.
21. Lai, J.S. & Eton, D.T. (2002*). Clinically Meaningful Gaps. Rasch Measurement Transactions 15(4): 850.*
22. Le, V. N. (1999). *Identifying Differential Item Functioning on the NELS: 88 History Achievement Test*. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.
23. Linacre, J. M. (2009). FACETS Rasch-model computer program (Version 3.66.0) [Computer software]. Chicago, IL: Winsteps.com.
24. Linacre, J. M. (2010) Winsteps®(Version 3.70.0) [Computer Software]. Beaverton, Oregon:Winsteps.com.
25. Linacre, J. M. & Wright, B. D. (1987). *Item bias: Mantel-Haenszel and the Rasch model* (Memorandum No. 39). Chicago: MESA Psychometric Laboratory.
26. Lord, F.M. (1980). *Application of item response theory to practical testing problem.* New Jersey: Lawrence Erlbaum Associates Publishers.
27. MEB. (2011). " "Türk Milli Eğitim Sisteminin Örgütlenmesi-2011" (Türkçe) (Ağ sayfası, metin, .html). http://sgb.meb.gov.tr/eurydice/index.htm" 11.10.2013 tarihinde adresinden alınmıştır. Milli Eğitim Bakanlığı Strateji Geliştirme Başkanlığı. Türkiye.
28. Micheels, W. J. & Karnes, M. R. (1950). *Measuring educational achievement*. McGraw-Hill.
29. Osterlind, S. J. (Ed.). (1983). *Test item bias* (Vol. 30).Sage.
30. Pallant, J.F. & Tennant, A. (2007*). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS).* Br J Clin Psychol 46: 1-18.
31. Roever, C. (2005). *"That's not fair!" Fairness, bias, and differential item functioning in language testing*. 18.05.2013 tarihinde Hawai'i Üniversitesi Sisteminden alınmıştır. Web site: http://www2.hawaii.edu/~roever/brownbag.pdf
32. Schumacker, R. E. (2005). Test Bias and Differential İtem Functioning. *Applied Measurement Associates*, 1-2.
33. Taylor, C. S. & Lee, Y. (2012): Gender DIF in Reading and Mathematics Tests With Mixed Item Formats. *Applied Measurement in Education*, 25:3, 246-280
34. Tekin, H. (2000). *Egitimde Ölçme ve Degerlendirme*(16. Baskı). Ankara:Yargı Yayınevi.
35. Tekin, H. (2000). *Egitimde Ölçme ve Degerlendirme*(16. Baskı). Ankara:Yargı Yayınevi.
36. Turgut, M.F. (1995). *Eğitimde Ölçme Değerlendirme Metotları*. Yargıcı Yayıncılık, 10.Baskı, Ankara
37. Turgut, M.F. & Baykul, Y. (2010). *Eğitimde Ölçme Değerlendirme*. Pegem Yayıncılık, Ankara
38. Wang, W. C. & Chen, C. T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the WINSTEPS program for the family of Rasch models. *Educational and Psychological Measurement*, *65*(3), 376-404.
39. Willingham, W. W. & Cole, N. S. (Eds.). (1997). *Gender and fair assessment*. Psychology Press.
40. Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). ACER ConQuest Version 2: Generalized item response modeling software [computer program]. Camberwell: *Australian Council for Educational Research*.

41. Yılmaz, H. (1998). *Eğitimde ölçme ve değerlendirme*. Mikro Yayınevi, Konya.
42. Yurdugul, H. (2003). *Ortaöğretim kurumları öğrenci seçme ve yerleştirme sınavının madde yanlılığı açısından incelenmesi*, Yayınlanmamış Doktora tezi, Hacettepe Üniversitesi, Ankara.
43. Yurdugul, H. & Askar, P. (2004). Ortaöğretim Kurumları Öğrenci Seçme ve Yerleştirme Sınavı'nın cinsiyete göre madde yanlılığı açısından incelenmesi.*Eğitim Bilimleri ve Uygulama*, *3*(5), 3-20.
44. Zenisky, A. L., Hambleton, R. K., & Robin, F. (2003). Detection of differential item functioning in large-scale state assessments: A study evaluating a two-stage approach. *Educational and Psychological Measurement*, *63*(1), 51-64.
45. Zenisky, A. L., Hambleton, R. K., & Robin, F. (2004). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment*, *9*(1-2), 61-78.
46. Zumbo, B. D. (1999*). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.
47. Zwick, R. & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement*, *26*(1), 55-66.