Journal of Pedagogical Research
Volume 5, Issue 3, 2021
https://doi.org/10.33902/JPR.2021371303

Research Article

Differential item functioning analysis of a high stake test in terms of statistical regions of Turkey

Yeşim Özer Özkan[1] and Meltem Acar Güvendir[2]

[1]*Gaziantep University, Faculty of Education, Turkey (ORCID: 0000-0002-7712-658X)*
[2]*Trakya University, Faculty of Education, Turkey (ORCID: 0000-0002-3847-0724)*

Large scale assessment is conducted at different class levels for various purposes such as identifying student success in education, observing the impacts of educational reforms on student achievement, assessment, selection, and placement. It is expected that these tests and their items are used in education do not display different traits with regard to the responses of individuals at the same ability level but from different groups. The purpose of this study was to identify whether the test includes Differential Item Functioning (DIF), and if it does, to identify whether the items are biased or not. The Rasch model was performed using the Winsteps package software to determine if the items contain DIF. DIF was identified in eight items during the analysis. Expert opinion was sought to determine if the difference between the regions in item is due to DIF or item effect. Based on the feedback from the experts, no bias was observed in the items with regard to regions.

Keywords: Differential item functioning; High stake test; Rasch model; Bias, Region

Article History: Submitted 12 March 2021; Revised 14 July 2021; Published online 21 August 2021

1. Introduction

Large-scale assessments are undertaken at various class levels for a range of goals, including measuring student achievement in school, studying the effects of educational reforms on student success, and assessing, selecting, and placing students. The results of these exams, which are used to transition from elementary to secondary school and from secondary to higher education, serve as a basis for making some decisions regarding students, depending on the aim of application (Baykul, 2000).

National large scale exams are implemented by the Ministry of National Education [MoNE] for determining the achievements of primary school students in Turkey. In addition to determining the level of achievement, Free Boarding and Scholarship Exam [FBSE] is also conducted for students who wish to continue their primary and secondary education with free boarding and scholarship. FBSE is conducted "during each academic year to select and place the students who want to continue their education with free boarding and scholarship; with central system for the 5th, 6th and 7th grades of the primary schools, 9th and 10th grades of secondary schools and the

Address of Corresponding Author

Yeşim Özer Özkan, PhD, Gaziantep University, Faculty of Education, Department of Educational Measurement and Evaluation, 27310, Gaziantep, Turkey.

✉ yozer80@gmail.com

How to cite: Özer-Özkan, Y. & Acar-Güvendir, M. (2021). Differential item functioning analysis of a high stake test in terms of statistical regions of Turkey. *Journal of Pedagogical Research, 5*(3), 122-134. https://doi.org/10.33902/JPR.2021371303

9th, 10th and 11th grades of the four year secondary school institutions". FBSE is implemented on students with financial difficulty at different class levels and for different courses. This exam measures the Turkish, Mathematics, Science and Social Sciences abilities of primary school students. Students are ranked based on their acquired scores, and finally, students who earned the right for scholarship are identified (MoNE, 2014).

Since FBSE is a selection and placement exam that aims to provide free boarding and scholarship to children of families with financial difficulties, it is critical that it yields valid, reliable and fair results. The fact that students with different demographic but similar socioeconomic traits (low socioeconomic level) from every region of Turkey take part in this exam increase the importance of the right decisions to be taken based on the results. For this reason, what is expected in the exam results is that the items are not affected by variables such as gender, socio-cultural level or region other than the students' abilities and that they do not work for or against one of these subgroups.

It is expected that the tests and their items used in education do not display different traits with regard to the responses of individuals at the same ability level but from different groups. In other words, it is expected that exams used for selection and placement purposes do not provide advantages or disadvantages to any students at the same class level but with different gender, socio-economic and statistical region traits (Öğretmen & Doğan, 2004). Lower or higher responses by a group to test items due to their certain traits compared with the other group is bias. Bias reflects on measurement results as "systematic error". This has an impact on validity and it is very important for validity analysis to identify biased items (Acar, 2011).

An examination of the literature shows that in gender (Akalın, 2014; Bakan Kalaycıoğlu & Kelecioğlu, 2011; Fincan, 2017; Kan et al., 2013; Karakaya, 2012; Satıcı & Özer Özkan, 2016; Türkan & Çetin, 2017) school type (Bekçi, 2007; Karakaya & Kutlu, 2012; Şenferah, 2015); educational regions and cultures (Gümüş Özyıldırım, 2018; Özmen, 2014; Yurdugül & Aşkar, 2004) based DIF and bias studies have been carried out frequently for large scale tests applied at different years. Ardıç and Gelbal (2017) conducted a study examining measurement invariance and DIF among groups for interest and motivation related items in the mathematics teaching section of the PISA 2012 student survey with regard to gender, school type and territorial units for statistics. It was observed that the comparison with regard to territorial units for statistics of the explained model was statistically significant. It was discussed whether the observed difference between the comparisons of territorial units for statistics is due to the actual situation. Uyar and Doğan (2011) carried out a study testing a model on the learning strategies in the learning to learn section of the student survey for the PISA 2009 Turkey sample group and the invariance of the model was tested in the gender, school type and territorial units for statistics (12 NUTS). The model was found to be equal in the territorial units for statistics sub-groups. It was concluded that any difference between the groups with regard to learning strategies is not related with the applied scale. Berberoğlu and Kalender (2005) carried out a geographical region based comparison in their study as a result of which it was concluded that even though the Eastern and Southeastern Anatolia regions along with the Black Sea region displayed relatively lower results in both the University Entrance Examination (UEE) and PISA results, the differences do not have a major practical meaning. In other words, regional differences were not as high as expected.

The purpose of this study is to examine the fit of the data to the Rasch model and to identify whether the test includes DIF and if it does, to identify whether the items are biased or not.

2. Method

Turkey Nomenclature of Territorial Units for Statistics [NUTS] (Ministry of Development, n.d.) is used to determine the comparisons to be performed in the study. This classification provides the opportunity to examine Turkey under 12 regions. The DIF findings of the items in the FBSE mathematics subtest according to the regions were analyzed according to the Level I classification. The distribution of students by regions is given in Table 1.

Table 1
Distribution of Students by Regions

Region	*Number of student*
İstanbul	918
Western Marmara	334
Aegean	1600
Eastern Marmara	534
Western Anatolia	1107
Mediterranean	3550
Central Anatolia	926
Western Black Sea	624
Eastern Black Sea	383
Northeastern Anatolia	409
Central Eastern Anatolia	1053
Southeastern Anatolia	3498
Total	14936

The population, geography, regional development plans, fundamental statistical indicators, and socioeconomic development ranking of Turkey's 12 provinces have been classified as one of the NUTS level 1. The research was based on the mathematics exam responses of 14.936 5th grade students who took part in the PBYS in 2014. Table 2 shows the distribution of students in terms of type of booklet and gender.

Table 2
Student Distribution in the Test According to Booklet Type and Gender

	Booklet A	*Booklet B*	*Total*	*%*
Female	3971	4122	8093	54.18%
Male	3439	3404	6843	45.81%
Total	7410	7526	14936	100%

Table 2 shows that female students account for 54.18 percent of the 5th grade students in the study's sample, while male students account for 45.81 percent.

2.1. Instrument

The Free Boarding and Scholarship Examination (FBSE) is designed for primary school students in the 5th, 6th, and 7th grades, secondary school students in the 9th and 10th grades, and four-year secondary school students in the 9th, 10th, and 11th grades, and is administered through a central system. Primary school students' knowledge in Turkish, Mathematics, Science, and Social Studies is assessed in this exam, and students who are eligible for scholarships are decided by a ranking based on the results (MoNE, 2014). The items from the 5th grade FBSE mathematics subtest administered in 2014 were used as data in this study. There are 25 multiple choice questions in the test.

2.2. Data Analysis

Whether the items contain DIF or not was determined by the Rasch model through the Winsteps (n.d.) package program. ConQuest (Wu et al., 2007), Facets (Linacre, 2009) and Winsteps (Linacre, 2010) package programs are used for DIF analysis detection in Rasch models (Karami, 2012). Winsteps program is one of the most preferred applications for DIF detection in Rasch model. Two methods are suggested to find the substance containing DIF in the Winsteps. One of them is the Mantel-Haezensel (MH) chi-square statistic and the other is the Welch t-test. In cases where $p < .05$ in the MH and Welch t-test, the items are considered to contain DIF. DIF items were flagged when the MH probability value was less than .05 and classified as negligible, moderate or large DIF

based on the DIF size suggested by Educational Testing Service DIF category (Zwick et al., 1999). Before determining whether items contain DIF, it was examined whether the data provided the unidimensionality and the local independence characteristics of the test items in order to meet the assumptions of the Rasch model.

A common assumption in IRT models is that a set of items in a test measure only one ability (Hambleton et al., 1991). In order to get reliable results from a test, the assumption of unidimensionality must be checked (Gao, 1997). Unidimensionality is one of the necessary conditions for estimating item and ability parameters in studies based on Rasch model. Confirmatory Factor Analysis (CFA) is used to verify compliance between the current analysis data and the factor structure. It can be assumed that the data providing model fit is unidimensional. Confirmatory factor analyzes were carried out using the *lavaan* package in the R program (Rosseel, 2012). CFA analyzes in the study were performed using the Weighted Least Squares (WLS). WLS method is used in studies where the data is not normally distributed or in which there are dichotomous items.

Local independence pertains to sufficiency of an IRT model for the data (Embertson & Reise, 2013). Local independence means that when the abilities influencing test performance are held constant, examinees' responses to any pair of items are statistically independent (Hambleton et al., 1991). In other words, students with similar performance levels are expected to give the same answer to the same question. Note in particular that local independence follows automatically from unidimentionality; because of local independence, which is guaranteed by unidimentionality (Lord, 1980, p.19, 54) Therefore, since local independence is to explain the relationships between items with only one ability, meeting the unidimensionality assumption can be interpreted as providing the local independence assumption. Also Yen (1984) revealed the Q3 statistics as the local item independence index. R studio program "sirt" package was used in the calculation of Q3 statistics (Robitzsch et al., 2020).

For model fit, each item must make a contribution to the structure. The mean square fit statistics are used to monitor the compatibility of the data with the model (Bond & Fox, 2007, p.239). In general, its expectation is 1.0. Values substantially less than 1.0 indicate overfit greater than 1.0 indicate underfit. MNSQ can be interpreted as follows: If MNSQ >2.0, distorts or degrades the measurement system; 1.5 < MNSQ ≤ 2.0 unproductive for construction of measurement, but not degrading; 0.5< MNSQ ≤ 1.5 productive for measurement; MNSQ <<0.5 Less productive for measurement, but not degrading (Linacre, 2002). Wright and Linacre (1994) suggest that MNSQ values less than 1.2 are acceptable for multiple-choice tests that are high stakes. Point-measure correlation coefficient (PTMEA CORR) is the correlation between scored responses and ability measures. We expect the highest category will have a strong positive correlation with ability, and the lowest category to have a strong negative correlation with ability (Linacre, 2012).

3. Results

3.1. Fit of the Data to the Latent Trait Model

Item difficulty level for 22 items in the mathematics test, standard error of the model, infit z-standardized (ZSTD) and Mean-square (MNSQ), outfit z-standardized (ZSTD) and mean square (MNSQ) and PTEMEA CORR is given in Table 3.

According to Table 3, the item difficulty level (2.25) was the highest the 24th item, while the item difficulty level (-1.41) was the lowest item 1. When the model fit coefficients of the items are examined, it is seen that these values are between 0.5 ment (Linacre, 2002).

Table 3
Analysis of Item Fit and Person-Item Summary Statistics

Raw Score	*Total Score*	*Measure*	*Model S.E.*	*Infit*		*Outfit*		*PT MEACORR*	*Item*
				MNSQ	*ZSTD*	*MNSQ*	*ZTSD*		
1430	13506	2.25	.03	1.03	1.5	1.46	9.9	.30	m24
2224	12692	1.62	.03	1.17	9.7	1.51	9.9	.25	m5
3230	11706	1.07	.02	.94	-4.8	.96	-1.7	.47	m19
3855	11080	.78	.02	1.14	9.9	1.16	8.7	.34	m2
4355	10581	.57	.02	1.11	9.9	1.17	9.9	.36	m3
4371	10565	.56	.02	.94	-5.8	.93	-4.9	.49	m18
4849	10087	.37	.02	1.05	5.3	1.11	7.4	.40	m13
5235	9701	.23	.02	.96	-4.4	.95	-4.3	.48	m25
5932	9004	-.03	.02	1.14	9.9	1.20	9.9	.33	m22
6121	8814	-.09	.02	.93	-8.7	.90	-9.0	.50	m23
6143	8793	-.10	.02	.98	-3.2	.95	-4.0	.46	m20
6182	8753	-.11	.02	.92	-9.9	.89	-9.9	.51	m14
6255	8681	-.14	.02	1.04	4.8	1.04	3.4	.41	m15
6339	8597	-.17	.02	.91	-9.9	.88	-9.9	.51	m11
6357	8579	-.17	.02	.97	-4.6	.94	-5.1	.47	m10
6548	8387	-.24	.02	.93	-9.9	.90	-8.9	.50	m6
6554	8381	-.24	.02	.88	-9.9	.84	-9.9	.54	m16
6583	8353	-.25	.02	1.13	9.9	1.22	9.9	.33	m21
7024	7911	-.40	.02	1.00	.4	1.00	.0	.43	m7
7236	7699	-.47	.02	.99	-1.8	.97	-2.5	.44	m9
7567	7368	-.58	.02	.90	-9.9	.85	-9.9	.51	m8
8335	6601	-.83	.02	.91	-9.9	.87	-9.9	.49	m17
8944	5992	-1.04	.02	1.14	9.9	1.32	9.9	.27	m4
9409	5526	-1.19	.02	.90	-9.9	.85	-9.9	.48	m12
10021	4914	-1.41	.02	1.00	.4	.99	-.8	.38	m1

3.2. Unidimensionality

The second assumption, unidimensionality, was tested using CFA. The modification index values were analyzed as a consequence of the CFA analysis, and the elements that altered the model's fit were eliminated. The first, second, and fourth questions from the mathematics test questions for the fifth grade, which totaled 25 items, were eliminated from the model, keeping the CFA with 22 items. According to the goodness of fit indices (Schermelleh-Engel & Moosbrugger, 2003) NFI=0.97, RMSEA=0.03, SRMR= 0.02, CFI=0.97, GFI=0.97 and p= .00 show a good fit. DFA final results are given in Table 4.

Table 4 shows that the factor loadings of the items ranged between .20 and .54. The table also shows items with a factor loading of less than .30. Considering the good fit of the model, the item was not dropped just becaouse of the factor load.

Table 4
Mathematics test CFA results

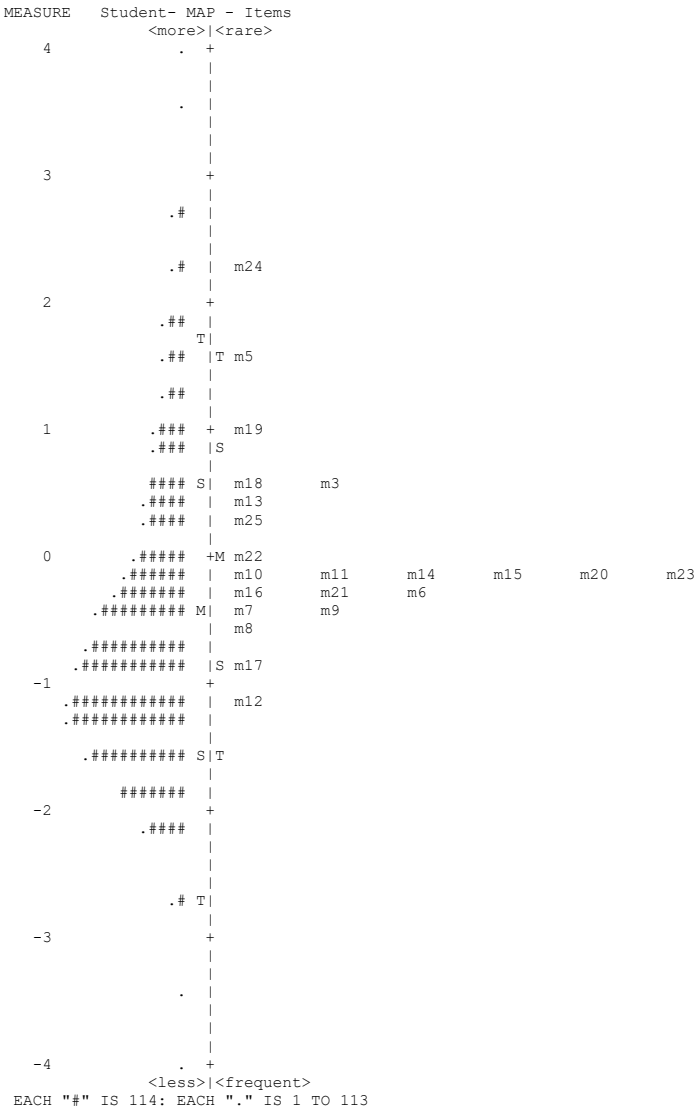
Items	*Factor Loadings*	*Std. Error*	*z-value*	*p*	*Lower*	*Upper*
m3	.300	.002	69.278	< .001	.133	.140
m5	.197	.002	45.296	< .001	.067	.074
m6	.474	.002	109.034	< .001	.231	.240
m7	.391	.002	91.123	< .001	.191	.199
m8	.497	.002	114.986	< .001	.244	.252
m9	.409	.002	95.276	< .001	.200	.208
m10	.437	.002	100.885	< .001	.212	.221
m11	.497	.002	113.757	< .001	.241	.250
m12	.436	.002	103.576	< .001	.206	.214
m13	.350	.002	80.465	< .001	.160	.168
m14	.501	.002	114.363	< .001	.243	.251
m15	.364	.002	84.663	< .001	.176	.184
m16	.536	.002	122.070	< .001	.262	.270
m17	.455	.002	106.737	< .001	.222	.230
m18	.479	.002	107.864	< .001	.214	.222
m19	.443	.002	99.268	< .001	.179	.186
m20	.434	.002	99.972	< .001	.209	.218
m21	.270	.002	63.522	< .001	.130	.138
m22	.262	.002	61.307	< .001	.124	.132
m23	.470	.002	107.791	< .001	.227	.235
m24	.244	.001	54.569	< .001	.069	.074
m25	.443	.002	101.140	< .001	.207	.216

3.3. Local Independence

The final assumption tested was local independence. In a test where the unidimensionality is provided, it can be interpreted as the local independence assumption is also fulfilled (Lord, 1980). Yen (1984) revealed the Q3 statistic as the local item independence index. The value of the Q3 statistic greater than .20 indicates that the local independence assumption cannot be achieved for the relevant item pair. The results regarding the Q3 statistics of Yen (1984) used in testing the local independence assumption are given in Appendix 1. The table in Appendix 1 shows that the statistics of all possible item pairs are less than .20. Accordingly, it was concluded that the test also provided the local independence assumption. Figure 1 shows the distribution of 25 items related to the mathematics test according to difficulty and individuals. In Figure 1, each "#" sign represents 114 people.

While the easiest item is represented at the bottom in Figure 1, the most difficult item is shown at the top. Accordingly, while the most difficult item was item 24 (m24) (item 1 was the most difficult item in the first analysis, item 24 turned out to be the most difficult item since this item was excluded in the CFA), the easiest item was found to be the 12th item (m12). As the difficulty level of the items increases, the response rate decreases and the item becomes a difficult item. On the contrary, as the difficulty level of the items decreases, the response rate increases and the item becomes an easy item.

Figure 1
Map of persons and items



3.4. Statistical Region DIF Investigation

DIF values of the items in the mathematics test were analysed according to 12 statistical regions determined by Turkish Statistical Institute. DIF contrast value and DIF change graph of the items in the mathematics test according to the statistical regions were included. In this section, according to the statistical regions, eight items with moderate and higher DIF findings from only DIF present substances and regions are included. DIF-flagged items are given Table 5.

Eight items had the MH probability value of less than .05 and therefore, were flagged as DIF items in Table 5. The result showed the moderate to large DIF exists between Western Black Sea and Northeastern Anatolia on item 5 (DIF contrast= .85 logit, t(957) = 4.31, p= .0000) and Eastern Black Sea and Northeastern Anatolia on item 22 (DIF contrast= .70 logit, t(784) = 4.28, p= .0000). The other six items exist slight to moderate DIF. DIF contrast values for these six items are in the range from -.43 to .62.

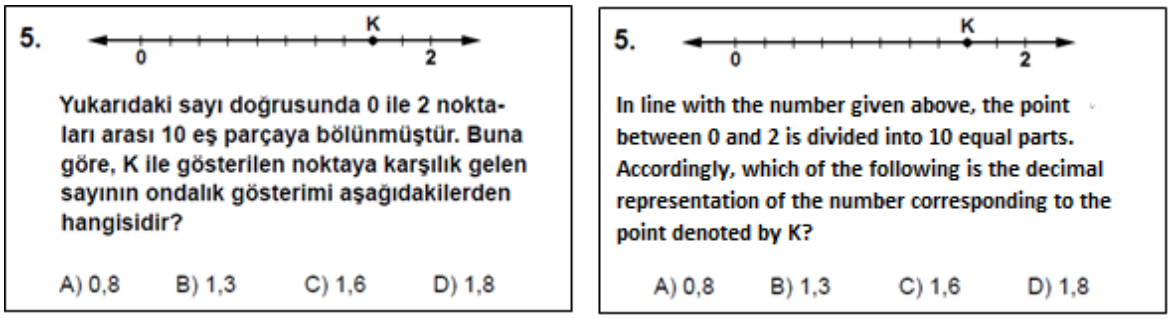
Table 5
DIF-flagged Items Based on IRT

Item	*Group*	*DIF Measure 1*	*Group*	*DIF Measure 2*	*DIF Contrast*	*Welch t*	*p*	*MH x2*	*MH p*	*DIF Size*
5	Western Black Sea	1.96	Northeastern Anatolia	1.11	.85	4.31	.0000	6.69	.0097	C
	Western Anatolia	1.88	Northeastern Anatolia	1.11	.77	4.40	.0000	5.53	.0186	C
	Western Black Sea	1.96	Central Eastern Anatolia	1.34	.62	3.84	.0001	5.96	.0146	B
	Mediterranean	1.67	Northeastern Anatolia	1.11	.57	3.61	.0003	3.73	.0535	B
	Western Anatolia	1.88	Southeastern Anatolia	1.33	.55	5.09	.0000	4.46	.0347	B
	Western Anatolia	1.88	Central Eastern Anatolia	1.34	.54	4.04	.0001	5.51	.0188	B
11	İstanbul	-.39	Western Black Sea	.04	-.43	-3.61	.0003	10.13	.0015	B
14	Northeastern Anatolia	.22	Eastern Black Sea	-.30	.52	3.14	.0017	51.95	.0226	B
	İstanbul	-.24	Northeastern Anatolia	.22	-.46	-3.26	.0012	59.45	.0148	B
	Eastern Marmara	-.22	Northeastern Anatolia	.22	-.45	-2.87	.0042	40.46	.0443	B
16	İstanbul	-.57	Western Anatolia	-.13	-.43	4.27	.0000	166.19	.0000	B
17	Northeastern Anatolia	-1.09	Eastern Black Sea	-.51	-.58	-3.63	.0003	9.89	.0082	B
21	İstanbul	.03	Northeastern Anatolia	-.51	.54	4.08	.0000	67.49	.0094	B
22	Eastern Black Sea	.36	Northeastern Anatolia	-.34	.70	4.28	.0000	64.45	.0111	C
	Eastern Black Sea	.36	Central Eastern Anatolia	-.22	.58	4.25	.0000	88.80	.0029	B
	Eastern Black Sea	.36	Southeastern Anatolia	-.19	.55	4.42	.0000	91.14	.0025	B
	Northeastern Anatolia	.34	Eastern Marmara	.19	.53	-3.53	.0004	38.20	.0506	B
	Western Marmara	-.14	Eastern Black Sea	.36	-.50	-2.93	.0035	62.48	.0124	B
	Western Black Sea	.17	Northeastern Anatolia	-.34	.50	3.44	.0006	45.57	.0328	B
	Western Anatolia	.11	Northeastern Anatolia	-.34	.45	3.41	.0007	51.86	.0228	B
	Aegean	-.07	Eastern Black Sea	.36	-.43	-3.31	.0010	85.63	.0034	B
25	İstanbul	.03	Eastern Black Sea	.50	-.47	-3.36	.0008	101.65	.0014	B
	İstanbul	.03	Western Anatolia	.47	-.44	-2.93	.0035	86.78	.0032	B

4. Discussion and Conclusion

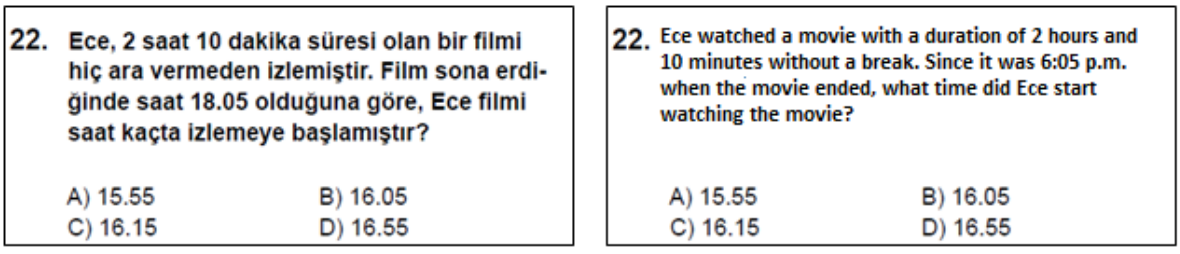
The results of large-scale exams have a significant impact on the shareholders of education. Education is structured based on the assumption that individual and social differences do not generate biased results in the applied tests. Hence, it is expected that the tests applied in the field of education are free of all kinds of bias as they are developed with an unbiased structure. The use of measurement tools before eliminating the impact of the demographic characteristics (gender, region, school type etc.) of the individuals may lead to the generation of erroneous results in large-scale examinations (Reise et al., 1993). The present study aims at investigating the validity of a high-stakes test in general and to considering the role of region as a source of bias in the FBSE. It was tested when analyzing the data whether the IRT and Rasch assumptions have been met or not. IRT based Rasch model was used to determine whether the items included in the study include DIF or not. DIF was identified in eight items during the DIF analysis. (It can be interpreted that out of the 22 items, 8 items display DIF-flagged items). Çelik and Özer Özkan (2020) reported DIF finding at a statistically significant level in all items as a result of the analysis based on the territorial units for statistics of the items in the PISA 2015 mathematics sub-test which is a large-scale international exam. The two items that display the highest DIF have been presented below. Expert opinion was consulted to determine whether the difference between the territorial units for statistics in the 5th item (see Figure 2) is doe to DIF or item effect.

Figure 2
5th item in the test



Based on the feedback from the experts, it was observed that there is no bias in the item with regard to the territorial units for statistics. The experts have stated that the "point" expression used in the question may lead to misunderstandings and that it will be more clear and understandable if expressed as "numbers between 0 and 2 were divided in 10 parts. ….". Item 22 (see Figure 3) is the second item with the highest DIF.

Figure 3
22th item in the test



Similar to the fifth item, no bias with regard to territorial units for statistics was determined, according to the expert feedback. Berberoğlu and Kalender (2005) carried out a comparison on the basis of geographical regions as a result of which it was concluded that despite the relatively lower results observed in the Eastern and Southeastern Anatolia and the Black Sea regions for both UEE and PISA results, they did not have a major practical significance.

When the results of the present study are compared to the results in existing literature, it can be concluded that the lack of bias is an important indicator in an examination applied in Turkey for ensuring social justice and eliminating inequalities when selecting free boarding and scholarship students. Education has an important role in eliminating social class discrimination through the efforts of individuals. It is also a social sub-system that has priority in the redistribution of status. Therefore, while ensuring education equality contributes to the development of human resources, it can also be an important point for social reconciliation. Thus, it is important that unbiased results can be attained from measurement and assessment as a sub-dimension of education which plays a critical role with regard to the transformation of individual efforts into a final product.

The present study is an indication that FBSE which is one of the large-scale exams in Turkey does not generate bias on the basis of regions. Gümüş Özyıldırım (2018) observed that the transition from basic education to the secondary education exam [TBESE] mathematics subtest contained negligible DIF according to geographical regions. According to Gümüş Özyıldırım (2018) the reason for not observing the substance containing DIF according to the geographical regions in the TBESE mathematics subtest is due to the fact that the mathematics course has a structure which is at the lowest level from culture to culture or person to person and logical rules in itself. Ardıç and Gelbal (2017) conducted a study it was observed that the comparison with regard to territorial units for statistics of the explained model was statistically significant. It was discussed whether the observed difference between the comparisons of territorial units for statistics is due to the actual situation. Expert evaluation revealed that the items that were statistically found to be biased did not include bias in practice as a result of the study.

Declaration of conflicting interest. The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Acar, T. (2011). Sample size in differential item functioning: an application of hierarchical linear modeling. *Educational Sciences: Theory & Practice*, *11*(1), 284-288.

Akalın, Ş. (2014). *The examination of the KPSS general ability test in terms of item bias* [Unpublished Doctoral Dissertation]. Ankara University, Ankara, Turkey.

Ardıç, E. Ö., & Gelbal, S. (2017). Cross-group equivalence of interest and motivation items in PISA 2012 Turkey sample. *Eurasian Journal of Educational Research*, *68*, 221-238. https://doi.org/10.14689/ejer.2017.68.12

Bakan Kalaycıoğlu, D., & Kelecioğlu, H. (2011). Item bias analysis of the university entrance examination. *Education and Science*, *36*(161), 3-13.

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: klasik test teorisi ve uygulanması* [Measurement and Evaluation in Education and Psychology: Classical test theory and practice]. OSYM Publications.

Bekçi, B. (2007). *Examining differential item functions of the elementary school student selection and placement examination according to gender and school type* [Unpublished Master's Thesis]. Hacettepe University, Ankara, Turkey.

Berberoğlu, G., & Kalender, İ. (2005). Investigation of student achievement across years, school types and regions: The SSE and PISA analyse. *Educational Sciences and Practice, 4*(7), 21-35.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. (2nd Ed.). Lawrence Erlbaum Associates Publishers.

Çelik, M., & Özer Özkan, Y. Ö. (2020). Analysis of Differential Item Functioning of PISA 2015 Mathematics Subtest Subject to Gender and Statistical Regions. *Journal of Measurement and Evaluation in Education and Psychology, 11*(3), 283-301. https://doi.org/10.21031/epod.715020

Embretson, S. E., & Reise, S. T. (2013). *Item response theory for psychologists*. Lawrence Erlbaum Associates Publishers.

Fincan, B. (2017). *An investigation of math subtests item bias in 2014 free boarding and scholarship examination in terms of gender variable* [Unpublished Master's Thesis]. Gaziantep University, Gaziantep, Turkey.

Gao, F. (1997). *DIMTEST enhancements and some parametric IRT asymptotics* [Unpublished Doctoral Dissertation]. University of Illinois at Urbana- Champaign, United States.

Gümüş Özyıldırım, F. (2018). *Examining TEOG mathematic sub-test exam in terms of differential item functioning based on geographical regions* [Unpublished Master's Thesis]. Hacettepe University, Ankara, Turkey.

Hambleton, R K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publication.

Kan, A., Sünbül, Ö., & Ömür, S. (2013). 6.-8. Sınıf seviye belirleme sınavları alt testlerinin çeşitli yöntemlere göre değişen madde fonksiyonlarının incelenmesi. [Examining differential item functioning of 6th-8th grade Level Determination Examinations (LDE) subtests to various methods] *Mersin University Journal of the Faculty of Education, 9*(2), 207-222.

Karakaya, İ., & Kutlu, 0.(2012). An Investigation of item bias in Turkish sub tests in level determination exam. *Education and Science, 37*(165), 348-362.

Karakaya, İ. (2012). An investigation of item bias in science & technology subtests and mathematic subtests in Level Determination Exam (LDE). *Educational Sciences: Theory & Practice, 12*(1), 215-229.

Karami, H. (2012). An introduction to differential item functioning. *The International Journal of Educational and Psychological Assessment, 11*(2), 59-76.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2009). *FACETS Rasch-model computer program* (Version 3.66.0) [Computer software]. Winsteps.com.

Linacre, J. M. (2010). *Winsteps*®(Version 3.70.0) [Computer Software]. Winsteps.com.

Linacre, J. M. (2012). *A User's Guide to WINSTEPS Rasch-model Computer Programs*. MESA Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Erlbaum.

Ministry of Development (n.d.). *İstatistiki Bölge Birimleri Sınıflandırması.[Statistical Region Units Classification].* Retrieved from January 6, 2018, from http://www3.kalkinma.gov.tr/PortalDesign/PortalControls/WebIcerikGosterim.aspx?Enc=83D5A6FF03C7B4FCC26F032470459B0B

Ministry of National Education [MoNE]. (2014). *Parasız yatılılık ve bursluluk sınavı (PYBS) başvuru ve uygulama e-Kılavuzu* [Free Boarding and Scholarship Exam (FBSE) reference and application e-Guide]. Retrived from October 10, 2016 from http://www.meb.gov.tr/

Öğretmen, T., & Doğan, N. (2004). OKÖSYS matematik alt testine ait maddelerin yanlılık analizi [Differential item functioning analysis for the Mathematics subtest of the High Schools Entrance Examination (HSEE)]. *İnönü University Journal of the Faculty of Education, 5*(8), 61-76.

Özmen, D.T. (2014). A study on PISA 2009 reading test items in terms of bias PISA 2009. *Educational Sciences and Practice, 13*(26), 147-165

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance, *Psychological Bulletin, 114*(3), 552-566. https://doi.org/10.1037/0033-2909.114.3.552

Robitzsch,A., Kiefer, T., & Wu, M. (2020). *TAM: Test analysis modules*. [R package]. https://CRAN.R-project.org/package=TAM

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Satıcı, K.. & Özer Özkan, Y. (2017). Investigation of item bias in 2014- November transition exam from primary to secondary education in terms of gender. *Mersin University Journal of the Faculty of Education, 13*(1), 254-274. https://doi.org/10.17860/mersinefd.305954

Schermelleh-Engel, K., & Moosbrugger, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23-74.

Şenferah, S. (2015). *An investigetion of differential item functioning and item bias for the mathematics subtest of level determination test in 2010* [Unpublished Doctoral Dissertation]. Gazi University, Ankara, Turkey.

Şengül, Ü., Eslemian, S., & Eren, M. (2013). Economic activities of regions of level 2 according to statistical regional units classification (NUTS) in Turkey determining by using DEA and Tobit model application. *Journal of Administrative Sciences, 11*(21), 75-99.

Türkan, A., & Çetin, B. (2017) Study of bias in 2012-placement test through Rasch Model in terms of gender variable. *Journal of Education and Practice, 8*(7), 196-204.

Uyar, Ş., & Doğan, N. (2011). An investigation of measurement invariance of learning strategies model across different groups in PISA Turkey sample. *International Journal of Turkish Education Science, 3,* 30-43.

Winsteps (n.d.). *Winsteps. Rasch Analysis and Rasch Measurement Software*. Retrieved April 13, 2018, from https://www.winsteps.com/winsteps.htm.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2: Generalized item response modeling software* [Computer software]. Australian Council for Educational Research.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125-145.

Yurdugül, H. & Aşkar, P. (2004). The investigation of the student selection and placement examination for secondary education with respect to student settlement region in terms of differential item functioning. *Hacettepe University Journal of Education, 27*, 268-275.

Zwick, R., Thayer, D.T., & Lewis, C. (1999) An empirical bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*(1), 1-28. https://doi.org/10.1111/j.1745-3984.1999.tb00543.x.

Appendix 1. Q3 Correlation Matrix

	m3	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16	m17	m18	m19	m20	m21	m22	m23	m24	m25
m3	–																					
m5	-0.005	–																				
m6	-0.064	-0.075	–																			
m7	-0.044	-0.024	-0.020	–																		
m8	-0.066	-0.081	-0.004	-0.052	–																	
m9	-0.048	0.024	-0.075	-0.051	-0.052	–																
m10	-0.066	-0.091	-0.003	-0.043	0.011	-0.092	–															
m11	-0.071	-0.093	-0.005	-0.046	-0.008	-0.022	-0.030	–														
m12	-0.052	-0.073	-0.028	-0.027	0.010	-0.042	-0.043	0.005	–													
m13	-0.052	-0.028	-0.042	-0.040	-0.057	-0.031	-0.036	-0.062	-0.065	–												
m14	-0.026	-0.020	-0.044	-0.043	-0.032	-0.015	-0.040	-0.007	-0.025	-0.071	–											
m15	-0.043	-0.044	-0.065	-0.058	-0.029	-0.048	-0.052	-0.039	-0.034	-0.043	-0.045	–										
m16	-0.053	-0.033	-0.035	-0.047	-0.028	0.007	-0.044	0.019	-0.000	-0.071	0.037	-0.032	–									
m17	-0.067	-0.080	-0.007	-0.026	0.000	-0.053	-0.022	-0.031	0.024	-0.078	-0.029	-0.053	-0.021	–								
m18	-0.027	0.035	-0.069	-0.045	-0.012	0.041	-0.079	-0.044	-0.033	-0.044	0.038	-0.036	0.019	-0.063	–							
m19	-0.051	-0.069	-0.005	-0.038	-0.015	-0.044	0.018	-0.013	-0.044	-0.011	-0.046	-0.057	-0.020	-0.023	-0.018	–						
m20	-0.049	-0.016	-0.049	-0.045	-0.068	0.003	-0.070	-0.021	-0.037	-0.019	0.009	-0.077	0.017	-0.050	0.016	-0.057	–					
m21	-0.014	0.075	-0.119	-0.046	-0.071	-0.009	-0.100	-0.108	-0.041	-0.067	-0.002	-0.042	-0.037	-0.060	0.067	-0.080	-0.007	–				
m22	-0.053	-0.116	0.009	-0.064	-0.066	-0.098	-0.012	-0.015	-0.065	-0.040	-0.102	-0.057	-0.089	-0.036	-0.146	-0.041	-0.081	-0.153	–			
m23	-0.075	-0.117	0.016	-0.041	0.018	-0.082	-0.000	-0.004	-0.012	-0.049	-0.060	-0.057	-0.028	0.021	-0.086	0.000	-0.064	-0.136	-0.002	–		
m24	-0.008	0.015	-0.047	-0.041	-0.048	-0.039	-0.032	-0.050	-0.073	-0.017	-0.037	-0.031	-0.056	-0.053	0.014	0.032	-0.055	-0.039	-0.036	-0.015	–	
m25	-0.075	-0.120	0.015	-0.054	0.005	-0.110	0.026	-0.023	-0.040	-0.007	-0.070	-0.020	-0.061	-0.009	-0.100	0.027	-0.077	-0.158	0.041	0.046	0.016	–