ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Araştırma Makalesi (Research Article); 2022; 13(1); 1-14

Gender-based Differential Item Functioning Analysis of the Medical Specialization Education Entrance Examination

Dilara BAKAN KALAYCIOĞLU *

Abstract

The Medical Specialization Education Entrance Examination is a national high-stake test for the placement of medical graduates in medical specialization training in Turkey. The purpose of this study is to determine whether the Medical Specialization Education Entrance Examination items display gender-related differential item functioning (DIF) by using Mantel-Haenszel and logistic regression methods. To determine the presence of item bias, content experts reviewed items. The analyzes were conducted on the answers of 11,530 physicians to the Basic Medical Sciences and Clinical Medical Sciences tests of the 2017 Medical Specialization Education Entrance Examination spring term. According to the Mantel-Haenszel method, there were eleven out of 234 items identified as showing B level gender-related DIF. While six of the items functioned in favor of male physicians, five of them were in favor of female physicians. Since the number of items in favor of each gender is close, DIF cancellation occurs. According to content areas, one histology and embryology, one internal medicine, and three gynecology and obstetrics items were in favor of female physicians, one physiology, two medical pharmacology, one pediatrics, and two surgical items were in favor of male physicians. To the experts' reviews, there are no biased items. The medical specialty preferences of the physicians and content area of the displaying differential item functioning items overlapped.

Keywords: medical specialization education entrance examination, residency, differential item functioning, logistic regression, Mantel-Haenszel

Introduction

Medical graduates (physicians) admitted to medical specialty training (residency) at medical faculties of universities and the Ministry of Health education research hospitals in Turkey with The Medical Specialization Education Entrance Examination (MSE, [Tıpta Uzmanlık Eğitimi Giriş Sınavı, TUS]) scores. The MSE has been administered by the Assessment, Selection and Placement Center [Ölçme Seçme ve Yerleştirme Merkezi, ÖSYM], twice a year, in spring and autumn terms since 1987. The MSE consists of Basic and Clinical Medical Sciences Tests. The Basic Medical Sciences Test (BMST) is designed to assess core medical science knowledge, whereas The Clinical Medical Sciences Test (CMST) core clinical knowledge. The MSE is a very competitive examination only 27.7% of physicians placed in a medical specialization training program in the 2017 spring term. The MSE is not the only concern of examinees but also other stakeholders such as medical schools and the Ministry of Health Medical Specialty Board.

All kinds of examinations, especially high-stake examinations where life-altering decisions are made concerning career progression, need to be fair to all test takers regardless of age, gender, disability, race, or other personal characteristics; otherwise, validity can be compromised (American Educational Research Association, 2018). Item bias is one of the threats to the validity of test score interpretation (Downing, 2002). In this context, differential item functioning (DIF) studies are important to provide validity evidence for proposed interpretations of test scores. According to Dorans and Holland (1992), "DIF refers to a difference in item performance between two comparable groups of examinees, that is groups, construct being measured by the test" (p.3). While determining the DIF items, it is made under the assumption that individuals in different subgroups are equal or equal in terms of the characteristics measured by the test. The aim is to distinguish between true group differences (item impact) and bias in

* Assoc. Prof., Gazi University, Gazi Education Faculty, Ankara-Turkey, dilarabakan@gmail.com; ORCID ID: 0000-0003-1447-6918

To cite this article:
Bakan Kalaycıoğlu (2022). Gender-based differential item functioning analysis of the medical specialization education entrance examination. *Journal of Measurement and Evaluation in Education and Psychology, 13*(1), 1-13. https://doi.org/10.21031/epod.998592

Received: 21.09.2021
Accepted: 12.12.2021

measurement. While doing this, individuals with the same ability level in different groups are matched so that individuals from different groups can be compared. Then, it is checked whether the performance of these individuals on the test items is the same. Displaying DIF items can be categorized as either uniform or nonuniform depending on how group membership interacts with ability. When there is no interaction between ability level and group membership, DIF is uniform. On the other hand, DIF is nonuniform when there is an interaction between ability level and group membership. DIF analysis is useful for identifying potentially biased items. However, DIF is a necessary but not sufficient condition for identifying a bias item. Zumbo (1999) recommends follow-up item analysis such as content analysis or empirical evaluation to determine the presence of item bias.

A range of procedures has been proposed based on different theories for analyzing DIF, including Mantel-Haenzsel (MH) (Holland & Thayer, 1986), Logistic Regression (LR) (Swaminathan & Rogers, 1990), Restricted Factor Analysis (RFA) (Oort, 1992), Item Response Theory Log-Likelihood Ratio (IRT-LR) (Thissen et al., 1993), Multiple Indicator Multiple Causes (MIMIC) model (MacIntosh & Hashim, 2003; Muthen, 1988) and others (Camilli & Shepard, 1994). These procedures have been studied fairly extensively in terms of their ability to correctly identify DIF items (Finch, 2005; Gomez-Benito & Navas-Ara, 2000; Güler & Penfield, 2009; Uğurlu & Atar, 2020). Since these procedures' assumptions and approaches to modeling the data are distinct, they may identify different items as displaying DIF. For this reason, it is recommended to use more than one procedure in DIF studies (Hambleton, 2006).

A recent systematic review of published studies that have analyzed DIF detection methods concluded that MH and LR procedures are the most widely studied using simulated data under various conditions Berrío et al. (2020). In terms of identifying the presence of uniform DIF, one of the most prominent and widely used methods of DIF detection is the MH procedure (Diaz et al., 2021; Gomez-Benito & Navas-Ara, 2000; Guilera et al., 2013). Another approach that has been discussed both in terms of uniform and nonuniform DIF is LR (Narayanan & Swaminathan, 1996; Swaminathan & Rogers, 1990) procedure. Although the observed score is an inadequate indicator as a substitute for a latent trait, the MSE subtests are constructed based on the Classical Test Theory. Thus, conventional DIF detection methods for uniform DIF MH procedure and nonuniform DIF LR procedure are preferred.

The MH is a non-parametric method designed to determine if the uniform DIF exists for the different sub-groups (Camilli & Shepard, 1994). An advantage of the non-parametric method is that there are few model assumptions so that DIF is not confounded with lack of model fit; however, such methods required larger samples (Teresi, 2006). MH divides the data into a focal group and a reference group and compares each group's performance matched on the skill level usually taken as the total raw score (Holland & Thayer, 1986). MH procedure is one of the most commonly used methods to detect DIF (Wainer & Sireci, 2005). MH tests the null hypothesis that there is no difference in item performance between the focal group and the reference group when controlling for ability. If a significant difference in group performance is found, the effect size is computed using the measure described by Zieky (1993). If $|\Delta MH| < 1$, DIF level is A (negligible); if $1 < |\Delta MH| < 1.5$, DIF level is B (medium); and if $|\Delta MH| \geq 1.5$, DIF level is C (high).

The LR is one of the most effective and recommended methods among various methods for determining DIF (Camilli & Shepard, 1994). Swaminathan and Rogers (1990) first applied the LR method to DIF detection studies. With this method, both uniform and nonuniform DIF can be detected. In the LR, variables are included in the model hierarchically. The variable to be limited in Model 1, namely the total score, takes its place in the model. In Model 2, the group variable (gender) is added, and in the third model, the interaction variable is included in the model in addition to the previous variables. The item shows uniform DIF when the group coefficient is statistically significant, and the interaction coefficient is not. The item shows nonuniform DIF when the interaction coefficient is statistically significant (Zumbo, 1999). The difference in Negelkerke R^2 values obtained from the third model and the first model includes both uniform and nonuniform DIF. Proposed effect size measures for LR procedure are, A level or negligible DIF, $\Delta R^2 < .035$, B level or moderate DIF, $.035 \leq \Delta R^2 \leq .070$, and C level or large DIF $\Delta R^2 \geq .070$ (Jodoin & Gierl, 2001). In this study, the Negelkerke R^2 value differences obtained from the LR analyses and Zumbo's (1999) $\Delta R^2 = R^2$ (step3)-R^2 (step1) formula was calculated.

Finch and French (2007) and Uğurlu and Atar (2020) reported in their studies that the power of the LR method increased as the sample size increased. Hidalgo and Lopez-Pina (2004) stated in their simulation study that the LR method is more effective than the MH method when determining nonuniform DIF, and the MH method is more effective when determining uniform DIF, and the results support the study of Swaminathan and Rogers (1990).

Although various gender fairness DIF studies were carried out on educational tests (Akcan & Atalay Kabasakal, 2019; Bakan Kalaycıoğlu & Berberoğlu, 2011; Çelik & Özer Özkan, 2020; Çepni & Kelecioğlu, 2021; Grover & Ercikan, 2017; Khorramdel et al., 2020; Kıbrıslıoğlu Uysal & Atalay Kabasakal, 2017) and health-related tests (Crane et al., 2004; Edelen et al., 2006; Sunderland et al., 2010) the number of studies on medical education tests was quite limited (Clauser et al., 1996a; Hope et al., 2018; Swanson et al., 2002). Clauser et al. (1996a) analyzed responses of medical students to pediatrics, surgery, obstetrics-gynecology, and medicine subtests using MH and LR methods. Fifty-six items out of 266 items were identified as exhibiting DIF between gender groups. MH and LR methods produced very similar results, and the percentage of items identified by both methods is 89%. Hope et al. (2018) investigated the performance of 13,694 candidates taking the Membership of the Royal Colleges of Physicians (MRCP UK) Examination using the LR method. Gender-related DIF analyses demonstrated that only eight items were identified as showing DIF out of 2,773 items. They emphasized that a panel of clinician assessors identified no plausible explanations for displaying DIF items. Swanson et al. (2002) used a hierarchical LR model to identify sources of gender-related DIF. They analyzed responses of 6,581 examinees to the clinical component of the United States Medical Licensing Examination (USMLE) and concluded that a potential explanation for gender DIF is an interaction between examinee gender and the medical discipline.

To the best of the researcher's knowledge, no DIF study has been conducted on the MSE items, which is why this study is essential. Moreover, by identifying possible causes of gender-related DIF or biased items, the study has the potential to provide support for the validity of the MSE scores. Fairness supports not only the validity but also the defensibility of the MSE scores, which are used for selection purposes. The psychometric properties of the MSE items must be adequate for both genders. Since the process of validation involves accumulating relevant evidence to provide information for score interpretations, this will be the main contribution of the present study.

In this study, we investigated the performance of 11,530 physicians taking the 2017 MSE spring term, a high-stake postgraduate assessment for medical specialty, and we compared males against females. The focus of this article is detecting DIF items in the MSE. We used MH and LR DIF procedures to test 234 items and report the results of DIF analysis alongside the expert reviews of displaying DIF items.

Purpose of the Study

The purpose of the study is to determine whether the MSE items display gender-related DIF by using MH and LR methods. Following DIF analysis, displaying DIF items were reviewed by a group of expert physicians to evaluate the possible causes of DIF and to decide item bias.

In this study, answers to the following research questions will be sought.

1. Are the items in the BMST and the CMST display gender-related DIF?
2. What are the possible causes of items displaying DIF?
3. If there are any items displaying DIF, do they indicate the presence of bias?

Method

Within the scope of this study, among the items in the 2017 MSE spring term test, those showing gender-related DIF were determined. This study is exploratory in terms of items identified as having DIF. Detailed information about participants, the data collection instruments, DIF detection methods were presented.

Participants

The participants were 11,530 physicians who took the 2017 MSE spring term. Of them, 5,841 (50.7%) were male and 5,689 (49.7%) were female physicians. The analyses were carried on the entire national data set (population). Non-medical professionals also take the MSE, but they are only placed in the quota determined for them. For this reason, only the data of medical faculty graduates were included in the scope of the study.

Data Collection Instruments

The Medical Specialization Education Entrance Examination

The MSE data were obtained from Assessment, Selection and Placement Center [Ölçme Seçme ve Yerleştirme Merkezi, ÖSYM]. The MSE is a national postgraduate medical examination that consists of the Basic Medical Sciences Test (BMST) and the Clinical Medical Sciences Test (CMST). Both subtests consist of 120 multiple-choice items, with five options and 150 minutes allotted to answer each test. Mostly due to security-related problems, items are cannot be field-tested and can be cancelled according to the objections of the examinees after the exam. In the 2017 MSE spring term, the ÖSYM scientific committee has cancelled three items from each test. Table 1 provides information on the content area of the tests and the number of items analyzed (ÖSYM, 2017).

Table 1
The Content Area and the Number of Items

Tests	Content Area	Number of Items
BMST	Anatomy	13
	Histology and Embryology	8
	Physiology	9
	Medical Biochemistry	21
	Medical Microbiology	22
	Medical Pathology	22
	Medical Pharmacology	22
CMST	Internal Medicine	41
	Pediatrics	29
	Surgical	35
	Gynecology and Obstetrics	12
Total		234

Physicians are placed in their preferred medical specialization solely based on their MSE scores. When calculating the MSE scores, correct and incorrect answers given by the candidates to the items in BMST and CMST are collected separately, and the raw scores of BMST and CMST are obtained by subtracting one-fourth of the number of correct answers from the number of wrong answers. These scores converted into standard scores with a mean of 50 and a standard deviation of 10 for each test (ÖSYM, 2017).

Expert Review Form

Expert reviews were obtained for displaying DIF items. Five experts reviewed the eleven items to evaluate the possible causes of DIF and to identify biased items. Except for one expert who has a Ph.D. in measurement and evaluation and an associate professor in the medical education department, all other four experts are physicians. Among these four physicians, two of them are professors, and two of them are Ph.D. students in the department of medical education and informatics. Items displaying DIF were

sent to the experts, but they only knew that DIF was present, not favoring gender. First, they were asked which items might work in favor of which gender, and then the possible causes of the DIF items were obtained via open-ended questions.

Data Analysis

The data were coded by marking correct answers as 1 and incorrect or missing answers as 0. Three items from each subtest were not included in the analyses as the ÖSYM scientific committee has canceled them. Before DIF analyses, descriptive statistics, subtest scores, and Cronbach's α coefficients were calculated for gender groups. A unidimensional measurement model was tested through confirmatory factor analysis. Then, MH and LR DIF detection methods were used. Following DIF analysis, expert reviews were obtained as a part of the process of possible causes of DIF and the item bias.

In this study, female physicians were considered as the focus group and male physicians as the reference group; the matching variable was defined as the total score obtained by summing all individual items in the BMST and the CMST separately. IBM SPSS Statistics (Version 26) and LISREL 8.80 (Jöreskog & Sörbom, 1993) statistical package programs were used in descriptive statistics and confirmatory factor analysis while DIF was determined with the EZDIF program (Waller, 1998) for the MH method, ΔR^2 results were obtained in SPSS with a special script written by Zumbo (1999) for the LR method.

Descriptive statistics

Table 2 shows the descriptive statistics and reliability values of tests by gender.

Table 2
Descriptive Statistics by Gender

				BMST			CMST		
	N	N%	*Age*	*M*	*SD*	*Cronbach α*	*M*	*SD*	*Cronbach α*
Male	5,841	50.7	27.51	61.99	20.37	.95	66.33	14.18	.89
Female	5,689	49.3	26.68	61.92	18.99	.94	67.12	13.49	.89
Total	11,530	100	27.10	61.95	19.70	.95	66.71	13.85	.89

As indicated in Table 2, 50.7% of the physicians who took the MSE were male, and 49.3% were female. The average age was 26.68 for females and 27.51 for male physicians. The mean BMST and CMST raw scores (total number of correct items) and standard deviations of male and female physicians are very close to each other. Cronbach's alpha coefficient has the following values for the BMST and CMST .94 and .89, respectively. According to Downing and Yudkowsky (2009), the acceptable threshold of Cronbach's alpha reliability coefficient, which is an index of the internal consistency, is .70, and values above .90 are required for high stake exams. The reliability values calculated for both subtests in the study are at an acceptable level. Table 3 shows raw score mean and standard deviation on content areas by gender.

Table 3
Content Area Raw Score Statistics by Gender

	Content Area	Male		Female	
		M	*SD*	*M*	*SD*
BMST					
	Anatomy	6.74	2.89	6.42	2.82
	Histology and Embryology	3.94	1.73	4.03	1.65
	Physiology	4.58	1.56	4.40	1.49
	Medical Biochemistry	10.16	4.68	10.18	4.33
	Medical Microbiology	12.54	4.53	12.80	4.18
	Medical Pathology	10.92	3.77	11.19	3.61
	Medical Pharmacology	10.26	5.03	10.06	4.81
CMST					
	Internal Medicine	25.09	5.34	25.09	4.99
	Pediatrics	14.58	4.02	15.18	3.96
	Surgical	18.81	4.73	18.74	4.57
	Gynecology and Obstetrics	5.83	2.14	6.14	2.07

As detailed in Table 3, the comparison of the content area raw score of gender groups was found similar. However, male physicians' mean scores in anatomy, physiology, medical pathology, medical pharmacology, and surgical content areas are higher than females', whereas female physicians' mean scores in histology and embryology, medical microbiology, pediatrics, and gynecology and obstetrics content areas are higher than males'.

Unidimensionality

The MH is a non-parametric method, and no distributional assumptions are required; however, a unidimensional construct is assumed. Since multidimensionality is an important contributor to false DIF, examination of the unidimensionality assumption is crucial (Shepard, 1982; Teresi, 2006). The Confirmatory Factor Analyses (CFA) were performed to examine the unidimensionality of BMST and CMST. The most commonly used goodness of fit indices are RMSEA (Root Mean Square Error of Approximation), CFI (Comparative Fit Index), NFI (Normed Fit Index), and GFI (Goodness of Fit Index). An RMSEA value of less than .05 to .08 indicates a close fit. CFI, NFI, GFI values close to .90 or .95 reflect a good fit (Hu & Bentler, 1998; Marsh et al., 2004; Schumacker & Lomax, 2010). Although χ^2 test statistics are provided, RMSEA, CFI, NFI, and GFI fit indices are interpreted due to the χ^2 test's sensitivity to sample size to assess model fit. Since χ^2 is a function of a sample, it may reject trivial model-data differences when it is used with a large sample size (Browne & Cudeck, 1993). Table 4 provides CFA fit indices for BMST and CMST items.

Table 4
Fit indices for CFA

Tests	χ^2	RMSEA	CFI	NFI	GFI
BMST	109,736	0.045	0.98	0.98	0.81
CMST	35,376	0.019	0.94	0.93	0.95

As reflected in Table 4, the CFA model for BMST yielded a χ^2=109,736, *df*=6,669, *p*<.0001, RMSEA=0.045, CFI=0.98, NFI=0.98, GFI=0.81 and, for CMST χ^2=35,376, *df*=6,669, *p*<.0001, RMSEA=0.019, CFI=0.94, NFI=0.93, GFI=0.95. Overall model fit seems adequate based on values selected fit indexes. These results indicated that, the unidimensional measurement model have adequate model-data fit for both tests.

Results

For the first research problem of the study, it was examined whether the items in the BMST and CMST contain gender-related DIF. Table 5 shows displaying DIF items in the BMST and the CMST according to MH and LR analyses results. Items that are displaying negligible or A level DIF for both methods were not provided in the table. In other words, the items given in the table were those determined B level DIF at least one of the methods.

Table 5
MH and LR Analyze Results

Item No	Favor	Content Area	MH α	MH χ^2	MH p	MH Δ-MH	MH SE	MH DIF Level	LR ΔR^2	LR DIF Level
BMST										
17	Female	Hist. and Embryology	0.648	86.014	0.000	1.019	0.110	B	0.010	A
23	Male	Physiology	1.670	173.895	0.000	-1.206	0.092	B	0.020	A
100	Male	Medical Pharmacology	1.564	66.978	0.000	-1.051	0.129	B	0.011	A
102	Male	Medical Pharmacology	1.854	188.563	0.000	-1.451	0.106	B	0.025	A
CMST										
7	Female	Internal Medicine	0.596	94.411	0.000	1.216	0.126	B	0.014	A
70	Male	Pediatrics	1.564	28.917	0.000	-1.051	0.195	B	0.008	A
79	Male	Surgical	1.628	51.888	0.000	-1.145	0.159	B	0.011	A
98	Male	Surgical	1.865	117.840	0.000	-1.465	0.136	B	0.019	A
115	Female	Gyn. and Obstetrics	0.632	125.768	0.000	1.080	0.096	B	0.011	A
117	Female	Gyn. and Obstetrics	0.615	130.695	0.000	1.143	0.100	B	0.013	A
118	Female	Gyn. and Obstetrics	0.611	130.636	0.000	1.158	0.099	B	0.014	A

As indicated in Table 5, MH gender-related DIF analyses demonstrated that 11 items were identified as showing DIF out of 234 items. Six of the items functioned in favor of male physicians, whereas five items functioned in favor of female physicians. Note that positive values of Δ-MH favor the focal group (females), while negative values for the reference group (males). In the BMST, four B-level DIF items were observed. Histology and embryology item was in favor of female physicians; one physiology and two medical pharmacology items were in favor of male physicians. In the CMST, seven B-level DIF items were observed. Among these items, one internal medicine and three gynecology and obstetrics items were in favor of female physicians, one pediatrics and two surgical items were in favor of male physicians. The last two columns of Table 5 pertain to the LR method. All of the eleven items determined exhibiting B level DIF by the MH method were classified as exhibiting negligible or A level DIF by the LR method. None of the items were detected by the LR procedure as having nonuniform DIF. Even though among the 234 items analyzed with the LR method, have the relatively high ΔR^2 values are the same as those detected in the MH method, the results indicated that overall there was low agreement between the MH and the LR in detecting DIF.

Tables 6 and 7 show the items exhibiting DIF in BMST and CMST, respectively.

Table 6

BMST Items Exhibiting DIF

Favors	Items	Options
Female	17. A 16-year-old patient admitted with the complaint of amenorrhea is diagnosed with androgen insensitivity syndrome. Which of the following is <u>not</u> seen in this patient?	A) 46, XY chromosome formula B) Female type external genitalia *C) Well-developed uterus and tubes* D) Normally developing breast tissue E) Presence of testicles in the inguinal canal and labial region
Male	23. I. 3% NaCl II. 5% dextrose III.Ringer lactate In which of the above solutions does the mean erythrocyte volume decrease in erythrocytes?	*A) Only I* B) Only II C) Only III D) I and II E) I and III
Male	100. A drug administered intravenously at a rate of 2 mg/minutes has a mean volume of distribution: 80 L, clearance: 500 mL/minute, and an elimination half-life of 2 hours. What is the expected steady-state plasma concentration of this drug in mg/L?	A) 250 B) 12 C) 5,7 *D) 4* E) 0,02
Male	102. I. They show only central nervous system localization. II.Dopamine retransporter protein belongs to the SLC family. III.They can carry a transmitter in the opposite direction with a mechanism independent of sodium ions. Which of the above statements regarding SLC family proteins, which are neurotransmitter reuptake transporters, are correct?	A) Only I B) Only II C) I and II D) I and III *E) II and III*

*Correct answers are shown in italic.

Table 7

CMST Items Exhibiting DIF

Favors	Items	Options
Female	7. Which of the following is <u>not</u> one of the main lifestyle changes that should be recommended to a hypertensive patient?	A) Salt restriction B) Bodyweight control C) Regular physical exercise D) A diet rich in vegetables, fruits, and fiber *E) Fluid restriction*
Male	70. Which of the following is <u>not</u> a CAG triplet nucleotide repeat disease?	A) Huntington's disease *B) Myotonic dystrophy* C) Spinal and bulbar muscular atrophy D) Spinocerebellar ataxia type 1 E) Machado-Joseph disease
Male	79. Which of the following statements about traumatic injuries of the diaphragm is <u>false</u>?	A) The vast majority of blunt injuries are on the left side. B) Since blunt injuries are due to high-energy trauma, additional organ damage and mortality are high. C) Diaphragmatic injuries that develop after blunt trauma are difficult to diagnose. *D) The absence of abdominal organs in the thorax on the posteroanterior chest radiograph excludes diaphragmatic injury.* E) Diaphragmatic injuries can be diagnosed and treated with video-assisted thoracoscopy or laparoscopy.
Male	98. According to the Brisbane 2000 liver terminology based on Couinaud's segmental anatomical classification, which segments of the liver should be included in the resection in right posterior sectionectomy?	A) 5 and 8 B) 2 and 3 C) 1, 2, and 3 *D) 6 and 7* E) 5, 6, 7, and 8
Female	115. A twenty-five-year-old female patient is admitted with complaints of dysmenorrhea and chronic pelvic pain. The patient states that she got married two years ago, could not get pregnant even though she did not use a contraceptive method, and felt pain during sexual intercourse. On pelvic ultrasonography, a cystic lesion of approximately 5 cm in diameter and containing internal echogenicity is detected in the ovarian lodge. Which of the following is the most likely diagnosis for this patient?	*A) Endometriosis* B) Adenomyosis C) Tuboovarian abscess D) Hydrosalpinx E) Hematometra
Female	117. A 15-year-old girl, who learned that breast development and pubic hair growth preceded her peers, is admitted with the complaint of vaginal bleeding. On pelvic examination, the presence of a unilateral adnexal mass is suspected. A cystic mass is detected on ultrasonography and the mass is surgically removed. Which of the following is the most likely histopathological diagnosis of this mass?	A) Sertoli cell tumor *B) Juvenile granulosa cell tumor* C) Leydig cell tumor D) Theoma E) Fibroma
Female	118. Which of the following ovarian reserve tests can be used independently of the menstrual phase?	A) FSH B) Estradiol C) Inhibin B D) Ovarian volume *E) Antimullerian hormone*

Note. Correct answers are shown in italic.

Following DIF analyses, for the second and third research problems, five experts reviewed eleven items to evaluate the possible causes of DIF and to decide if any biased items. The experts did not inform about which items were in favor of which gender. Experts accurately guessed that histology and embryology, and gynecology and obstetrics items might have worked in favor of female physicians due to content area. However, they pointed out items require knowledge, and being a woman does not help to answer these items correctly. One of the experts indicates that "Female physicians may have complaints about gynecological diseases, but since the situations given in the options are not situations where healthy individuals will be encountered frequently, a female physician cannot establish a diagnosis and answer the items correctly." On the other hand, the other experts expressed that familiarity of a gender group with the content of the item may be the reason for possible DIF sources. For physiology, medical pharmacology, internal medicine, pediatrics, and surgical items, experts stated that no situation would require the items to exhibit DIF or even to provide an advantage to a certain gender in any way. Experts conclude that some items are expected to differentiate according to the gender of the physicians, but this cannot be considered as bias. Among eleven DIF items, none of the items was determined as biased by experts.

Discussion and Conclusion

This paper presents DIF analyses of high-stake national postgraduate medical examination. According to MH procedure, there were eleven items identified as having B level DIF. Six items functioned in favor of male physicians, whereas five items functioned in favor of female physicians. Considerable DIF has been identified in the MSE, but according to the expert review, none of the items was determined as biased. A similar pattern of results was obtained in the MRCP UK examination (Hope et al., 2018). The number of items showing DIF in favor of female and male physicians is close to each other indicates that items containing DIF do not provide an advantage in favor of one gender when the total scores are taken into account. The cumulative effect of items exhibiting DIF against one subgroup cancels with other items that exhibit DIF against the comparison group and hence results in there is a cancellation of item-level DIF at the MSE test. In other terms, when calculating the composite scores, DIF cancellation occurs (Teresi, 2006; Wyse, 2013). Considering that the MSE items are used without any pre-testing process, the present research was an important contribution of evidence to the MSE score validity.

When the items were examined in detail, it was observed that the medical specialty preferences of the physicians and the content area of the displaying DIF items were overlapped. It is noteworthy that three out of four items in CMST favor of female physicians belongs to the gynecology and obstetrics subtest and the other item belongs to the internal medicine subtest, which might be suggestive of a confounding effect of physicians' specialization preferences in the MSE performance. According to Bakan Kalaycıoğlu (2020), generally, male physicians prefer surgery and female physicians prefer internal and basic medical sciences, but the exception to this situation is that female physicians in Turkey predominantly prefer gynecology and obstetrics, which is one of the surgical specialties. It can be said that the same situation applies to male physicians as well. Two out of three items in CMST favor of male physicians belongs to the surgical content area. According to the 2017 MSE spring data, 65% of the physicians placed in the gynecology and obstetrics specialty were female, while 89% of the physicians placed in the general surgery specialty were male. It seems likely that physicians are more successful in the content area (medical specialty) they are interested in and plan to prefer specialty training. A histology and embryology item in the BMST favors female physicians and 69% of the physicians placed in the histology and embryology specialty were also female. This result is consistent with Swanson et al.'s (2002), research findings explaining gender-related DIF in the clinical sciences component of the United States Medical Licensing Examination (USMLE-Step 2). According to Swanson et al. (2002), since female examinees disproportionately enter specialty training in obstetrics and gynecology and male examinees enter training in surgery, the largest positive coefficient (favoring women) is for obstetrics and gynecology, and the largest negative coefficient (favoring men) is for surgery which indicating that medical discipline explains the variance of DIF.

However, it is not possible to match the medical pharmacology and pediatrics subtests of the exhibiting DIF items with the medical specialty preferences of the physicians since these items favor male physicians, but these specialties are mostly preferred by female physicians (77% for medical pharmacology and 63% for pediatrics).

Despite the present research's contribution to providing validity evidence of the MSE scores, it also has several limitations. DIF can have multiple explanations and occur for several reasons, including the medical discipline of the item, item count, gender of patients described in test items, medical specialty choice (Swanson et al., 2002), residency training (Clauser et al., 1996b), sociological process of gender differences (Zumbo & Gelin, 2005), item format (Kelly & Dennick, 2009), item difficulty, the phrase or word used in the item, and familiarity of a group with the content of the item (Allaouf et al., 1999). These inevitably have an effect on physicians' familiarity with some content areas, but in this study, only gender-related medical specialty preference of the physicians was discussed. Investigations can be made by considering different variables such as evaluating the content areas of the items showing DIF together with the educational background of physicians (Clauser et al., 1996b) will provide a good starting point for further research in the field of medical education. Furthermore, DIF analyses on undergraduate-level medical data sets are required to evaluate this study's results. In this paper, DIF analyses were conducted on the BMST and the CMST, but each test consists of items from different content areas. DIF analyses may conduct on for each content area. MH and LR techniques were used to determine DIF, using other DIF methods such as IRTLR, MIMIC model may yield different results.

In the overall testing process, ongoing investigations of test validity are crucial. When constructing the test, items should not favor any test taker. This study was conducted after the 2017 MSE to determine displaying DIF items. Examinations must be routinely analyzed before the exam to ensure that they should not unfairly hinder the success of any subgroups. Thus, potentially biased items can be revised or removed. At least, by equalizing the number of items exhibiting DIF in favor of subgroups, the overall impact of DIF can be minimized.

Declarations

Acknowledgments: I would like to thank the Assessment, Selection and Placement Center (ÖSYM) for providing data.

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: Secondary data were used in this study. Therefore, ethical approval is not required.

References

Akcan, R., & Atalay Kabasakal, K. (2019). An investigation of item bias of English test: The case of 2016 year undergraduate placement exam in Turkey. *International Journal of Assessment Tools in Education, 6*(1), 48-62. https://doi.org/10.21449/ijate.508581

Allaouf, A., Hambleton, R., & Sireci, S. (1999). Identifying the causes of translation DIF on verbal items. *Journal of Educational Measurement, 36*(3), 185-198. https://www.jstor.org/stable/1435153

American Educational Research Association. (2018). *Standards for educational and psychological testing.* American Educational Research Association.

Assessment, Selection and Placement Center [Ölçme Seçme ve Yerleştirme Merkezi, ÖSYM]. (2017). *2017 Tıpta Uzmanlık Eğitimi Giriş Sınavı başvuru kılavuzu*. Retrieved from: https://dokuman.osym.gov.tr/pdfdokuman/2017/TUSILKBAHAR/BASVURUKILAVUZU26042017.pdf

Bakan Kalaycıoğlu, D. (2020). Changes in physicians' specalization preferences from 1987 to 2017. *Tıp Eğitimi Dünyası*, *19*(59), 157-170. https://doi.org/10.25282/ted.696179

Bakan Kalaycıoğlu, D., & Berberoğlu, G. (2011). Differential item functioning analysis of the science and mathematics items in the university entrance examinations in Turkey. *Journal of Psychoeducational Assessment*, *29*(5), 467-478. https://doi.org/10.1177%2F0734282910391623

Berrío, Á. I., Gomez-Benito, J., & Arias-Patiño, E. M. (2020). Developments and trends in research on methods of detecting differential item functioning. *Educational Research Review, 31*, 100340. https://doi.org/10.1016/j.edurev.2020.100340

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Sage.

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items*. Sage Publications.

Clauser, B. E., Nungester, R. J., Mazor, K., & Ripkey, D. (1996a). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement, 33*(2), 202-214. https://doi.org/10.1111/j.1745-3984.1996.tb00489.x

Clauser, B. E., Nungester, R. J., & Swaminathan, H. (1996b). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measurement, 33*(4), 453-464. https://doi.org/10.1111/j.1745-3984.1996.tb00501.x

Crane, P. K., Belle, G. van, & Larson, E. B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine, 23*(2), 241-256. https://doi.org/10.1002/sim.1713

Çelik, M., & Özer Özkan, Y. (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

Çepni, Z., & Kelecioğlu, H. (2021). Detecting differential item functioning using SIBTEST, MH, LR and IRT methods. *Journal of Measurement and Evaluation in Education and Psychology*, *12*(3), 267-285. https://doi.org/10.21031/epod.988879

Diaz, E., Brooks, G., & Johanson, G. (2021). Detecting differential item functioning: Item Response Theory methods versus the Mantel-Haenszel procedure. *International Journal of Assessment Tools in Education, 8*(2), 376-393. https://doi.org/10.21449/ijate.730141

Dorans, N. J., & Holland, P. W. (1992). *DIF detection and description: Mantel-Haenszel and standardization* (Research Report 92-10). Educational Testing Service.

Downing, S. M. (2002). Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education, 7*(3), 235-241. https://doi.org/10.1023/A:1021112514626

Downing, S. M., & Yudkowsky, R. (2009). Introduction to assessment in the health professions. In *Assessment in health professions education* (pp. 21-40). Routledge.

Edelen, M. O., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: Application to the Mini-Mental State Examination. *Medical Care, 44*(11), 134-142. https://doi.org/10.1097/01.mlr.0000245251.83359.8c

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*(4), 278-295. https://doi.org/10.1177/0146621605275728

Finch, W. H., & French, B. F. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and Psychological Measurement, 67*(4), 565-582. https://doi.org/10.1177/0013164406296975

Gomez-Benito, J., & Navas-Ara, M. J. (2000). A comparison of χ^2, RFA and IRT based procedures in the detection of DIF. *Quality and Quantity, 34*(1), 17-31. https://doi.org/10.1023/A:1004703709442

Grover, R. K., & Ercikan, K. (2017). For which boys and which girls are reading assessment items biased against? Detection of differential item functioning in heterogeneous gender populations. *Applied Measurement in Education, 30*(3), 178-195. https://doi.org/10.1080/08957347.2017.1316276

Guilera, G., Gómez-Benito, J., Hidalgo, M. D., & Sánchez-Meca, J. (2013). Type I error and statistical power of the Mantel-Haenszel procedure for detecting DIF: A meta-analysis. *Psychological Methods, 18*(4), 553-571. https://psycnet.apa.org/doi/10.1037/a0034306

Güler, N., & Penfield, R. D. (2009). A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. *Journal of Educational Measurement, 46*(3), 314-329. https://doi.org/10.1111/j.1745-3984.2009.00083.x

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care, 44*(11), 182-188. https://doi.org/10.1097/01.mlr.0000245443.86671.c4

Hidalgo, M. D., & Lopez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*(6), 903-915. https://doi.org/10.1177%2F0013164403261769

Holland, P. W., & Thayer, D. T. (1986, April 16-20). *Differential item performance and the Mantel-Haenszel procedure* [Paper presentation]. 67th Annual Meeting of the American Educational Research Association, San Francisco, CA.

Hope, D., Adamson, K., McManus, I. C., Chis, L., & Elder, A. (2018). Using differential item functioning to evaluate potential bias in a high stakes postgraduate knowledge based assessment. *BMC Medical Education*, *18*, 64. https://doi.org/10.1186/s12909-018-1143-0

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424-453. https://doi.org/10.1037/1082-989X.3.4.424

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329-349. https://doi.org/10.1207/S15324818AME1404_2

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Scientific Software International Inc.

Kelly, S., & Dennick, R. (2009). Evidence of gender bias in true-false-abstain medical examinations. *BMC Medical Education*, *9*(1), 1-7. https://doi.org/10.1186/1472-6920-9-32

Khorramdel, L., Pokropek, A., Joo, S. H., Kirsch, I., & Halderman, L. (2020). Examining gender DIF and gender differences in the PISA 2018 reading literacy scale: A partial invariance approach. *Psychological Test and Assessment Modeling*, *62*(2), 179-231.

Kıbrıslıoğlu Uysal, N., & Atalay Kabasakal, K. (2017). The effect of background variables on gender related differential item functioning. *Journal of Measurement and Evaluation in Education and Psychology*, *8*(4), 373-390. https://doi.org/10.21031/epod.333451

MacIntosh, R., & Hashim, S. (2003). Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis. *Applied Psychological Measurement*, *27*(5), 372-379. https://doi.org/10.1177/0146621603256021

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

Muthen, B. O. (1988). Some uses of structural equation modeling validity studies: Extending IRT to external variables. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 213-238). Lawrence Erlbaum.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*(3), 257-274. https://doi.org/10.1177/014662169602000306

Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika*, *6*(2), 150-166.

Schumacker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling* (3rd ed.). Taylor and Francis Group.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 9-30). John Hopkins University Press.

Sunderland, M., Mewton, L., Slade, T., & Baillie, A. J. (2010). Investigating differential symptom profiles in major depressive episode with and without generalized anxiety disorder: True co-morbidity or symptom similarity? *Psychological Medicine*, *40*(7), 1113-1123. https://doi.org/10.1017/S0033291709991590

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370. https://www.jstor.org/stable/1434855

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, *27*(1), 53-75. https://doi.org/10.3102/10769986027001053

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. *Medical Care*, *44*(11), S152-S170. https://doi.org/10.1097/01.mlr.0000245142.74628.ab

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Ed.), *Differential item functioning* (pp. 67-114). Lawrence Erlbaum Associates.

Uğurlu, S., & Atar, B. (2020). Performances of MIMIC and logistic regression procedures in detecting DIF. *Journal of Measurement and Evaluation in Education and Psychology*, *11*(1), 1-12. https://doi.org/10.21031/epod.531509

Wainer, H., & Sireci, S. G. (2005). *Encyclopedia of social measurement*. ScienceDirect.

Waller, N. G. (1998). EZDIF: Detection of uniform and nonuniform differential item functioning with the Mantel-Haenszel and logistic regression procedures. *Applied Psychological Measurement*, *22*(4), 391-391. https://doi.org/10.1177/014662169802200409

Wyse, A. E. (2013). DIF cancellation in the Rasch model. *Journal of Applied Measurement*, *14*(2), 118-128.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Lawrence Erlbaum Associates.

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. Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. *Journal of Educational Research & Policy Studies*, *5*(1), 1-23.