International Journal of Assessment Tools in Education
2019, Vol. 6, No. 1, 48–62

DOI: 10.21449/ijate.508581

Published at http://www.ijate.net http://dergipark.gov.tr/ijate *Research Article*

# An Investigation of Item Bias of English Test: The Case of 2016 Year Undergraduate Placement Exam in Turkey

**Rabia Akcan** [1,*], **Kübra Atalay Kabasakal** [2]

[1] Ministry of National Education, 03700, Afyonkarahisar, Turkey
[2] Hacettepe University, Education Faculty, Measurement and Evaluation Department, 06800, Ankara, Turkey

**ARTICLE HISTORY**

*Received: 25 September 2018*
*Revised: 20 December 2018*
*Accepted: 02 January 2019*

**KEYWORDS**

Undergraduate Placement Exam,
differential item functioning,
differential bundle functioning,
item bias,
MIMIC

**Abstract:** The purpose of this study is to determine whether English test items of Undergraduate Placement Exam (UPE) in 2016 contain differential item functioning (DIF) and differential bundle functioning (DBF) in terms of gender and school type and examine the possible sources of bias of DIF items. Mantel Haenszel (MH), Simultaneous Item Bias Test (SIBTEST) and Multiple Indicator and Multiple Causes (MIMIC) methods were used for DIF analyses. DBF analyses were conducted by MIMIC and SIBTEST methods. Expert opinions were consulted to determine the sources of bias. Data set of the study consisted of responses of 59818 students to 2016 UPE English test. As a result of the analyses carried out on 60 items, it was seen that one item in translation subtest contained DIF favoring male students. In school type based analyses, it was concluded that there were nine DIF items in vocabulary and grammar knowledge subtest, six DIF items in reading comprehension subtest and four DIF items in translation subtest. Experts stated that one item containing DIF by gender was unbiased, and evidence of bias was found in thirteen of nineteen items that contained DIF by school type. According to DBF analyses, it was found that some item bundles contained DBF with respect to gender and school type. As a result of research, it was discovered that there were differences with regard to the number of DIF items identified by three methods and the level of DIF that the items contained; however, methods were consistent in detecting uniform DIF.

## 1. INTRODUCTION

Large scale tests are commonly used throughout the world with the aim of selection and placement of the students. To make fair and right decisions based on the test results, and select students who have the ability and interest in accordance with the departments, the ability to be measured in the test must be evaluated accurately. Hence, it is significant to have well-qualified items for the tests. In a test, probability of answering an item correctly must not be influenced

CONTACT: Rabia Akcan ✉ eltrabia42@hotmail.com Ministry of National Education, 03700, Çay/Afyonkarahisar, Turkey

**ISSN-e: 2148-7456 /© IJATE 2019**

by variables such as examinees' socio economic status, gender or school type they studied. Otherwise, test becomes biased and might not reflect examinees' cognitive abilities.

Bias is described as systematic errors in measurement process (Osterlind, 1983). The concept of bias is directly associated with fairness, and it is the condition in which group characteristics that are not related to the construct to be measured affect test results. Thus, test bias distorts results by allowing examinees' characteristics influence measurement of main construct. Consequently, test measures an irrelevant construct in addition to the intended one (Mcnamara & Roever, 2006). When a group of examinees has a higher probability of responding an item correctly than another group due to some characteristics of the item or inconvenient test conditions, it is called item bias (Zumbo, 1999). Item bias is a possible threat to validity (Clauser & Mazor, 1998). Therefore, doing research on this matter is of importance.

Language teaching has become increasingly important throughout the world. Countries develop language tests for measuring language skills of the students, and decisions are made based on test results. Because these tests shape the future of the students as well as the countries, preparing equal and valid tests is higly significant. Since English is an international language and frequently used in science and technology, in this study English test in 2016 Undergraduate Placement Exam (UPE) held in Turkey was examined in terms of item bias. A test item can be said to be biased when it is in favor of one group and to the disadvantage of another group. These items show differential item functioning (DIF). DIF occurs when testtakers from different groups have different probability of success on an item after they are matched on the ability to be measured. DIF is a necessary condition but containing DIF is not sufficient for item bias (Clauser & Mazor, 1998). DIF can be present in two forms as uniform and non-uniform. When a group of examinees has higher likelihood of answering an item correctly than another group across all ability levels, uniform DIF occurs (Finch, 2005). On the other hand, non uniform DIF is present if the difference of the likelihood of answering an item correctly between the two groups is inconsistent across all ability levels (Camilli & Shepard, 1994).

Although the focus is generally on the single item DIF analysis, there are many tests consisting of small item bundles. An item bundle is described as a set of items selected according to an organizing principle and these items do not have to be adjacent and they are not necessarily related to a text or passage. When DIF analysis is conducted on item bundles, it is called differential bundle functioning (DBF) (Douglas, Roussos, & Stout, 1996). In the literature there are some studies on DIF in language tests however; more research is needed to improve test quality.

Lin and Wu (2003) investigated DIF and DBF with respect to gender in English Profiency Test used in China. Simultaneous Item Bias Test (SIBTEST) method was used for the analyses. Research results revealed that two items contained large DIF, and eleven items contained moderate DIF. Four of these items were listening items favoring females and three of them were grammar and vocabulary items favoring males. Two cloze test items and three reading items also favored males, and only one reading item was in favor of females. According to DBF results, one listening item bundle favored females systematically, and the other item bundles favored males slightly.

Abbott (2007) carried out DIF and DBF analyses of reading passages separated according to bottom-up and top-down strategies. SIBTEST method was used for the analyses. Hypothesis of the research was based on the claim that Chinese students are more successful in bottom-up strategies and Arabic students are more successful in top-down strategies. In analyses, items were separated into two categories in line with these two strategies. Research results showed that there were significant systematic differences between the two groups in using these strategies.

Kan (2007) conducted DIF analysis of items used in Hacettepe University foreign language proficiency examination. DIF analyses were carried out in terms of gender and the departments by using Mantel Haenszel (MH) method. It was reported that one item showed DIF favoring female students. Twelve items contained DIF in terms of departments separated into three categories as social sciences, physical sciences and health sciences.

Karakaya and Kutlu (2012) investigated item bias of Turkish subtests in Level Determination Exam. DIF analyses were conducted in terms of gender and school type using MH and Logistic Regression methods. Expert opinions revealed that only one item (item 19) in 8th grade Turkish subtest was biased in favor of males. Experts stated that item 19 included some expressions associated with feeding fish in an aquarium. Since male students are more interested in aquariums and feeding fish, item 19 was identified as biased.

Although there are many DIF detection methods described in literature, very few of them are used in practice (Clauser & Mazor, 1998). These methods can be broadly categorized into two as Classical Test Theory (CTT) and Item Response Theory (IRT) based methods. Camilli and Shepard (1994) stated that Confirmatory Factor Analysis (CFA) methods can also be used in DIF detection. In this study, MH, SIBTEST and Multiple Indicators Multiple Causes (MIMIC) methods were used for DIF detection.

### 1.1. Mantel-Haenszel

Mantel Haenszel (MH) statistic was proposed by Holland and Thayer (1988) and it has been commonly used for DIF detection since then. In this method, two groups are matched on the ability, and the probability of success on the item is compared between groups. Total test scores are generally used for matching (Clauser & Mazor, 1998). Afterwards, reference and focal groups are matched on total test scores, and a 2x2xS contingency table is created. S represents the different number of total test score. At all ability levels, data for each item can be organized as in Table 1 (Roussos & Stout, 1996).

**Table 1.** MH Method Data Organization.

| Group | 1 =Correct | 0=Incorrect | Total |
|---|---|---|---|
| Reference | $A_j$ | $B_j$ | $N_{Rj}$ |
| Focal | $C_j$ | $D_j$ | $N_{Oj}$ |
| Total | $M_{1j}$ | $M_{0j}$ | $T_j$ |

Using these tables that are formed at all ability levels, likelihood ratio (α) is estimated and this ratio is shown by equation 1 (Clauser & Mazor, 1998).

$$\alpha = \frac{\sum_j A_j D_j / T_j}{\sum_j B_j C_j / T_j} \quad (1)$$

To facilitate interpretation, log of α is taken and resulting value is multiplied by -2.35. Thus $\Delta_{MH}$ is produced. Positive values of $\Delta_{MH}$ show DIF against reference group and negative values of $\Delta_{MH}$ show DIF against the focal group (Clauser & Mazor, 1998). Zieky (1993) classified $\Delta_{MH}$ statistic as the following: $|\Delta_{MH}| < 1$ indicates negligible DIF, $1 \leq |\Delta_{MH}| \leq 1.5$ indicates moderate DIF and $|\Delta_{MH}| \geq 1.5$ indicates large DIF.

### 1.2. SIBTEST

SIBTEST was developed by Shealy and Stout (1993) and is based on the standardization procedure. In this method, test items are seperated into two groups as studied subtest and matching subtest. Corresponding matching subtest scores for the reference and focal groups are

estimated for each matching subtest, and these scores are modified using regression correction. Lastly, the ratio of answering the studied item correctly for reference and focal group is estimated. By using the weighed sum of the difference between these ratios, $\beta$ parameter is found (Roussos & Stout, 1996).

SIBTEST hypothesis is given by:

$$H_0: \beta = 0 \qquad H_1: \beta \neq 0 \tag{2}$$

And the size of DIF is expressed as:

$$\beta = \int [P(\theta, R) - P(\theta, F)] f_F(\theta) d\theta \tag{3}$$

where P (θ, R), probability of correct response for examinees from reference group; P (θ, F), probability of correct response for examinees from focal group; $f_F$ (θ), density function in focal group; d is the width of the scaling interval. With SIBTEST method, items or item bundles in the secondary dimension can be detected, and DIF analysis can be carried out. $\beta$ parameter is used to identify the size of DIF for items or item bundles (Gierl & Khaliq, 2001).

Roussos and Stout (1996) proposed guidelines for $\beta$ parameter to classify the size of DIF that have three levels : negligible DIF ($|\beta| < 0{,}059$), moderate DIF ($0{,}059 \leq |\beta| \leq 0{,}088$) and large DIF ($|\beta| > 0{,}088$). Positive values of $\beta$ show DIF against focal group and negative values of $\beta$ show DIF against the reference group. However, no guidelines have been proposed for classifiying the size of DBF (Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001).

In this research, both uniform and nonuniform DIF have been identified using SIBTEST method. Li and Stout (1996) proposed Crossing-SIBTEST (CSIBTEST) statistic which they see as a better alternative to SIBTEST statistic for identifying nonuniform DIF. This statistic was modified by Chalmers (2017), and it was stated that modified version of CSIBTEST statistic can be used in place of the orijinal CSIBTEST statistic.While data can be analysed with samples consisted of at most 7000 individuals for reference and focal groups in SIBTEST programme, there is no limit in R software. Therefore; DIF analysis was performed with “mirt” package (Chalmers, 2018) in R which gives an opportunity to estimate SIBTEST and CSIBTEST statistics simultaneously.

### 1.3. MIMIC

MIMIC is a model of CFA and can be used to detect DIF. MIMIC models estimate direct and indirect effects for a grouping variable. Latent trait is regressed onto grouping variable by indirect effect to show whether there is group mean differences on the latent trait. By direct effect, item responses are regressed onto grouping variable to find out whether response probabilities differ across groups (Finch, 2005).

MIMIC model in DIF context is expressed as:

$$y^*i = \lambda_i \eta + \beta_i z_k + \varepsilon_i \tag{4}$$

where $y^*i$, latent response variable; $\lambda_i$, factor loading for variable i; η, latent trait; $\beta_i$ , slope relating the group variable with the response; $\varepsilon_i$, random error; $z_k$, a dummy variable showing group membership (Finch, 2005).

Finch (2012) expanded MIMIC model, and used MIMIC as an alternative to SIBTEST in identifying DBF. Results of the research revealed that MIMIC model was as effective as SIBTEST in detecting DBF. In analyses with MIMIC method, positive values of beta show DIF against the group coded as 0, negative values of beta show DIF against the group coded as 1.

In this study focal group was coded as 0, and reference group was coded as 1. No criterion has been proposed to determine the size of DIF and DBF with MIMIC method.

## 2. METHOD

This study is a descriptive research as it investigates DIF and DBF of English test items in UPE, and it is also a qualitative research because it examines the possible sources of bias in DIF items. English test includes 80 items; however, 20 testlet items were excluded from the analysis. Therefore; 60 dichotomous scored items were analysed in terms of gender and school type.

### 2.1. Population and Sample

Population of the study consists of 88284 examinees who took English test in 2016 year UPE. Data set included 87 school types, and 74 of them were not analysed because a part of them was less than 1% of data set and the others were shut down by Ministry of National Education. Rest of the schools was gathered under four school types as they have similar educational objectives. Those four schools are vocational high school (VHS), Anatolian high school (AHS), religious vocational high school (RVHS) and private high school (PHS). Before factor analysis is conducted, data set should be checked whether it is appropriate for the analysis. To accomplish this, it was determined whether the data set included missing values and outliers. It was seen that missing values were below 5% of the data set, and zero imputation was used for the missing values. Data set was also examined in terms of univariate and multivariate outliers, and it was found that there were 1853 multivariate outliers in the data. Analyses were carried out using data from 59818 examinees after these outliers were removed. Distribution of data according to gender and school type is reported in Table 2.

**Table 2.** Gender and School Type Distribution.

| Group | N | % |
|---|---|---|
| Gender | | |
| Female | 36101 | 60.4 |
| Male | 23717 | 39.6 |
| School Type | | |
| Vocational | 10140 | 17.0 |
| Anatolian | 21618 | 36.1 |
| Religious V. | 11194 | 18.7 |
| Private | 16866 | 28.2 |
| Total | 59818 | 100 |

### 2.2. Instrument

English test in UPE consists of three parts as vocabulary and grammar knowledge (15 items), reading comprehension (48 items) and translation (12 items). All items in the test are multiple choice items. Vocabulary and grammar knowledge part includes items that measure basic vocabulary and grammar knowledge of the students. Reading comprehension part contains seven different item types. These are paragraph completion, cloze test, reading paragraphs, dialogue completion, sentence completion, irrelevant sentence and situational dialogue. Translation part consists of English-Turkish translation items and Turkish-English translation items. Exam takes two hours. Students are placed in departments according to their results.

### 2.3. Data Set

Data set used in this research was obtained from Research and Development Unit of Student Selection and Placement Center.

### 2.4. Data Analysis

Test items on the data set were scored as 1 for correct response and 0 for wrong or blank response. To examine the structure of the data, factor analysis was made using "lavaan" package (Rosseel, 2017) in R. It was made based on tetrachoric correlation matrix, and parallel analysis was used to decide the number of factors. To accomplish this "polycor" package (Fox, 2016) and "nFactors" package (Raiche & Magis, 2015) were used. Factor analysis indicated that there were 3 dimensions which were vocabulary and grammar knowledge (items 1-15), reading comprehension (items 21-28, 44-63 and 76-80) and translation (items 64-75) dimensions. Descriptive statistics and item statistics according to groups and DIF analyses were carried out based on these dimensions. DIF analyses were performed by using SIBTEST, MH and MIMIC methods. Descriptive statististics and item statistics were estimated using "CTT" package (Willse, 2018), and DIF analyses with SIBTEST were conducted using "mirt" package (Chalmers, 2018), MH analyses were performed using "difR" package (Magis, Beland, & Raiche , 2016) in R. DIF analyses with MIMIC were carried out using Mplus (Muthen & Muthen, 1998) via "MplusAutomation" package (Hallquist & Wiley, 2018) in R. To determine the possible sources of bias in DIF items expert opinions were consulted. SIBTEST and MIMIC methods were used for DBF anaylses as well.

## 3. FINDINGS

### 3.1. DIF Results and Expert Opinions

In this study, DIF and DBF anaylses of English test of UPE in 2016 was conducted in terms of gender and school type. Analyses were carried out based on three dimensions regarded as subtests. Six comparisons were made in each subtest with regard to school type. Items that show moderate or large DIF with SIBTEST and MH methods were considered as DIF items if they show DIF with MIMIC method at the same time. Ten experts were consulted to determine the possible sources of bias. Four experts work as English language teachers in the Ministry of National Education. Two experts have a degree of doctoral philosophy in department of English teaching, and four experts have a degree of doctoral philosophy in educational measurement and evaluation. DIF analyses results according to gender are reported in Table 3.

**Table 3.** DIF Items by Gender in Each Subtest

| Subtests/Gender | Female | Male |
|---|---|---|
| Vocabulary and grammar knowledge | - | - |
| Reading comprehension | - | - |
| Translation | - | 68 |

*DIF items that contain nonuniform DIF with SIBTEST method

As shown in Table 3, there are no common DIF items with three methods in vocabulary and grammar knowledge and reading comprehension subtests. There is one DIF item in favor of males in translation subtest. Experts stated that there is no evidence of bias in item 68 favoring males. Results of DIF analyses with regard to school type are given in Table 4.

Table 4 shows that there are no common items with three methods in reading comprehension and translation subtests in VHS-AHS comparison. However, in vocabulary and grammar knowledge subtest items 3, 5 and 8 indicated DIF in favor of VHS. It was found that items 7 and 12 contained DIF in favor of AHS with MH and MIMIC methods. These items showed nonuniform DIF with SIBTEST method. Five experts pointed out that as item 3 included some expressions related to information technology this may have given an advantage to the students graduated from information technology (IT) departments of VHS. Because those students are familiar with the expressions and this could be a possible source of bias in the item. Whereas

seven experts stated that there is no evidence of bias in items 5 and 8, three experts identified different sources of bias such as materials used in classes, familiarity with the expressions used in items and knowing problem solving techniques that can help students answer the items easily.

**Table 4.** DIF Items by School Type in Each Subtest

| Subtests/School Type | Vocational | Anatolian |
|---|---|---|
| Vocabulary and grammar knowledge | 3, 5, 8 | 7*,12* |
| Reading comprehension | - | - |
| Translation | - | - |
| **Subtests/ School Type** | **Vocational** | **Religious Voc.** |
| Vocabulary and grammar knowledge | - | 13 |
| Reading comprehension | 57, 58 | - |
| Translation | - | - |
| **Subtests/ School Type** | **Vocational** | **Private** |
| Vocabulary and grammar knowledge | 3, 8, 13* | 12* |
| Reading comprehension | - | - |
| Translation | 71* | - |
| **Subtests/ School Type** | **Anatolian** | **Religious Voc.** |
| Vocabulary and grammar knowledge | 7*,10*,12*,14* | 3 ,4*,5,8,13 |
| Reading comprehension | 57, 58 | - |
| Translation | - | - |
| **Subtests/ School Type** | **Anatolian** | **Private** |
| Vocabulary and grammar knowledge | 7* | - |
| Reading comprehension | - | 26, 47 |
| Translation | 72 | 68 |
| **Subtests/ School Type** | **Religious Voc.** | **Private** |
| Vocabulary and grammar knowledge | 3, 8, 13 | 10*, 12* |
| Reading comprehension | 63 | 45*, 57, 58 |
| Translation | 72* | 73* |

*DIF items that contain nonuniform DIF with SIBTEST method

When Table 4 was examined, it was seen that AHS students were advantageous in items 7 and 12 in comparisons with VHS and RVHS. For item 7, six experts asserted that students graduated from AHS may have been frequently exposed to that type of items and they might be familiar with grammatical structure used in that item. This could be the reason for the difference between AHS and the other school types. For item 12, one of the experts stated that science-related terms used in the item might have helped AHS students understand the item easily. On the other hand, five experts agreed on the idea that AHS students are familiar with the grammatical structures like “unless”, used in item 12, and thus this may have given them an advantage. Four experts found no evidence of bias in item 12.

In VHS- RVHS comparison, item 13 in vocabulary and grammar knowledge subtest contained DIF favoring RVHS. Items 57 and 58 in reading comprehension subtest showed DIF in favor of VHS. However, there is no common item with three methods in translation subtest. According to experts, item 13 is unbiased. Nine experts found no evidence of bias in item 57 and one expert claimed that students graduated from cookery department of VHS may have been familiar with the terms used in the item, and that might be a source of bias. Three experts stated that item 58 is situational dialogue item, and RVHS students might be unfamiliar with the situation given in the item due to socioeconomic and cultural differences. No evidence of bias was identifed in item 58 by the other seven experts.

In VHS-PHS comparison, items 3, 8 and 13 in vocabulary and grammar knowledge subtest showed DIF favoring VHS and item 12 showed DIF favoring PHS. While there is no common item containing DIF with three methods in reading comprehension subtest, one item (71) contained DIF favoring VHS in translation subtest. It was determined that items 13 and 71 showed DIF in favor of VHS and item 12 contained DIF in favor of PHS with MH and MIMIC methods. These items showed nonuniform DIF with SIBTEST method.

As mentioned before five experts stated that item 3 may have favored VHS students because it includes some expressions that IT students can understand more easily. Item 8 was found unbiased by seven experts. Yet two experts assserted that VHS students may have answered this item by just choosing simple option. For item 12, two experts pointed out those science-related words like “body cells” might be the source of bias and three experts found no source of bias. Five experts explained that PHS students might have been exposed to and be familiar with that type of items as in AHS students’ case, so this could be the reason of the difference between VHS-PHS. All experts had a common view that items 13 and 71 were not biased.

In AHS-RVHS comparison, four items were in favor of AHS and five items were in favor of RVHS in vocabulary and grammar knowledge subtest. Two items favored AHS in reading comprehension subtest, and there were no common items showing DIF with three methods in translation subtest. While items 7, 10, 12 and 14 had DIF favoring AHS and item 4 showed DIF favoring RVHS with MH and MIMIC methods, they contained nonuniform DIF with SIBTEST method.

As in items 7 and 12 mentioned in AHS-VHS comparison, six experts stated that item 14 is a grammar item and it includes the frequently used expression “not only, but also”, thus AHS students may have practiced, and they became familiar with that type of items which might be a factor that made AHS students more successful on the item. As for item 10, four experts explained that AHS students are more interested in science, so some expressions used in item 10 such as “brain”, “scientific evidence” and science related content of the item might have given an advantage to AHS students. Whereas three experts considered familiarity with item type as a source of bias, three experts found no source of bias in the item.

Six experts stated that item 57 was not biased. On the other hand, four experts said that item 57 is a situational dialogue item and RVHS students may not be familiar with the sample situation in the item as it includes the words “vegetarian”, “beefsteak” etc. Socioeconomic and cultual differences were proposed as the cause of bias. Similarly, item 58 was determined to be biased by three experts as it contains a situational dialogue that is not familiar with RHVS students. For item 3, only one expert asserted that item might be biased because it requires students to remember the information and RVHS students are mostly educated based on rote learning. Item 5 was found to be unbiased by seven experts, and one expert stated that this item also requires memory like item 3. Socioeconomic differences were defined as the cause of bias by two experts. For item 8, four experts pointed out that words used in the item such as “temple” and “dome” might create the difference between the schools because religious terms may have made the item easier to understand for RHVS students. Two experts explained that they might have responded the item just choosing the simple disractor as well. No evidence of bias was found by experts in items 4 and 13 between AHS-RVHS comparison.

In AHS-PHS comparison, item 7 showed DIF favoring AHS in vocabulary and grammar knowledge subtest and items 26 and 47 contained DIF favoring PHS in reading comprehension subtest. Items 68 and 72 indicated DIF favoring PHS and AHS, respectively. Eight experts found no evidence of bias in item 26; however, one expert emphasized the importance of language teaching techniques in PHS. Because language is taught for everyday use and students get a chance to practice it, PHS students might be advantageous in this sentence completion item. Morever, one expert stated that item contains expressions related to science such as

"muscular pump", "blood flooding" and heart", so this could be a reason for the difference between schools. For item 47, a dialogue completion item, four experts explained that languge education based on practice and everyday use in PHS may have given an advantage to PHS. Scientific content of the item and eating habits were proposed as sources of bias in that item by two experts. Item 72 was found to be unbiased by experts. Six experts found no evidence of bias in item 68 and four experts pointed out that students in PHS with a higher socioeconomic level might get the item more easily due to their social life and family structure because it includes the words "French and British antique", "antiques bazaar" and "antiques lovers".

In RVHS-PHS comparison, items 3, 8 and 13 showed DIF favoring RVHS and items 10 and 12 showed DIF favoring PHS in vocabulary and grammar knowledge subtest. Items 45, 57 and 58 contained DIF favoring PHS in reading comprehension subtest. Item 72 had DIF favoring RVHS and item 73 had DIF favoring PHS. But, items 10, 12, 45, 72 and 73 indicated nonuniform DIF with SIBTEST method.

When Table 4 was examined, it was seen that items 3, 8 and 13 were in favor of RVHS in RVHS-PHS comparison as they were in AHS-RVHS comparison. Therefore, experts stated that the sources of bias mentioned earlier in AHS-RVHS comparison are also valid in RVHS-PHS comparison. Likewise, AHS and PHS students were more advantageous in items 57 and 58 than RVHS students. Hence, for these items experts showed the same sources of bias given in AHS-RVHS comparison. For item 10, five experts showed the scientific content of the item and three experts showed the familiarity with the item type as a source of bias because PHS students are more likely to encounter that type of items and they tend to learn science. Similary, type of the item and scientific expressions used in item 12 were indicated as a source of bias by experts. It was clearly seen that AHS and PHS students were advantageous in item 12.

For item 45, seven experts pointed out that expressions used in the item such as "technology", "futureFest", "demos" and "innovation" might have given an advantage to PHS students because they are more likely to attend festivals and have an idea about them thanks to their socioeconomic level. Two experts stated that practice and educational activities performed in PHS might affect the results as item 45 is a dialogue completion item. Item 73 was considered as unbiased by six experts. Two experts explained that activities about different countries and cultures may have been done more in PHS than RVHS, and the other two experts showed practice and education based on every day use as the cause of difference between RVHS-PHS. Experts reached a concensus on that items 63 and 72 did not have bias.

### 3.2. DBF Results

Item bundles to be analysed can be chosen using different methods such as content analysis, table of specifications or psychological analysis (Gierl, Bisanz, Bisanz, Boughton & Khaliq, 2001). Nevertheless, UPE Engish test consists of item bundles including different types of items. These item bundles are prepared to measure insructional objectives and cognitive abilities of the students. There are two item bundles in vocabulary and grammar knowledge and translation subtests and six item bundles in reading comprehension subtest. Item bundles and item numbers are given in Table 5.

DBF analyses were carried out using SIBTEST and MIMIC methods. Woods and Grimms (2011) reported that MIMIC model was used to detect nonuniform DIF in their research and it worked better than the other model but type I error of this model was highly inflated. Although MIMIC can be used to identify nonuniform DIF or DBF, in this research only uniform DIF and DBF were detected by MIMIC owing to inflated type I error.

**Table 5.** Item Bundles and Numbers of Items

| Item Bundles | Item Numbers |
|---|---|
| Vocabulary and Grammar Knowledge | |
| Vocabulary | 1, 2, 3, 4, 5 |
| Grammar | 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 |
| Reading Comprehension | |
| Sentence Completion | 21, 22, 23, 24, 25, 26, 27, 28 |
| Dialogue Completion | 44, 45, 46, 47, 48 |
| Paraphrasing | 49, 50, 51, 52, 53 |
| Situational Dialogue | 54, 55, 56, 57, 58 |
| Paragraph Completion | 59, 60, 61, 62, 63 |
| Irrelevant Sentence | 76, 77, 78, 79, 80 |
| Translation | |
| English-Turkish Translation | 64, 65, 66, 67, 68, 69 |
| Turkish- English Translation | 70, 71, 72, 73, 74, 75 |

Furthermore, whether matching subtest contains DIF or not is an important issue in DBF analysis. Finch (2005) stated that DIF items found in matching subtest may threaten the accuracy of statistical methods in identifying DBF. Results of DIF analyses by school type demonsrated that there were quite a number of items including moderate or large DIF in vocabulary and grammar knowledge and translation subtests. Hence, DBF analyses by school type were only performed in reading comprehension subtest. DBF results by gender are given in Table 6.

**Table 6.** DBF Results by Gender

| Methods/Item Bundles | SIBTEST | MIMIC |
|---|---|---|
| Vocabulary | M | M |
| Grammar | F | - |
| Sentence Completion | M | M |
| Dialogue Completion | NU | - |
| Paraphrasing | NU | - |
| Situational Dialogue | F | F |
| Paragraph Completion | M | M |
| Irrelevant Sentence | F | F |
| English-Turkish Translation | M | F |
| Turkish- English Translation | F | F |

*F: Female, M: Male, NU: Non-uniform DIF

As shown in Table 6, seven item bundles showed DBF with both methods. However, methods are inconsistent in English-Turkish translation bundle. DBF results by school type are given in Table 7.

When Table 7 was examined, it was found that all bundles in AHS-VHS and RVHS-PHS comparisons indicated DBF with both methods. Five item bundles showed DBF in VHS-RVHS and AHS-RVHS comparisons with both methods. In VHS-RVHS comparison three item bundles had DBF, and in AHS-PHS comparison there were four common item bundles showing DBF with both methods. Nevertheless, DBF results of two methods do not completely comply with each other.

**Table 7.** DBF Results by School Type

| Methods/Item Bundles (VHS-AHS) | SIBTEST | MIMIC |
|---|---|---|
| Sentence Completion | NU | AHS |
| Dialogue Completion | NU | AHS |
| Paraphrasing | VHS | AHS |
| Situational Dialogue | AHS | AHS |
| Paragraph Completion | NU | AHS |
| Irrelevant Sentence | NU | AHS |
| Methods/Item Bundles (VHS-RVHS) | SIBTEST | MIMIC |
| Sentence Completion | UN | VHS |
| Dialogue Completion | UN | - |
| Paraphrasing | - | - |
| Situational Dialogue | VHS | VHS |
| Paragraph Completion | RVHS | RVHS |
| Irrelevant Sentence | UN | - |
| Methods/Item Bundles (VHS-RVHS) | SIBTEST | MIMIC |
| Sentence Completion | NU | PHS |
| Dialogue Completion | NU | PHS |
| Paraphrasing | VHS | - |
| Situational Dialogue | NU | PHS |
| Paragraph Completion | VHS | PHS |
| Irrelevant Sentence | NU | PHS |
| Methods/Item Bundles (AHS-RVHS) | SIBTEST | MIMIC |
| Sentence Completion | NU | AHS |
| Dialogue Completion | NU | AHS |
| Paraphrasing | RVHS | - |
| Situational Dialogue | NU | AHS |
| Paragraph Completion | RVHS | AHS |
| Irrelevant Sentence | NU | AHS |
| Methods/Item Bundles (AHS-PHS) | SIBTEST | MIMIC |
| Sentence Completion | NU | PHS |
| Dialogue Completion | PHS | PHS |
| Paraphrasing | AHS | AHS |
| Situational Dialogue | NU | AHS |
| Paragraph Completion | NU | - |
| Irrelevant Sentence | AHS | - |
| Methods/Item Bundles (RVHS-PHS) | SIBTEST | MIMIC |
| Sentence Completion | NU | PHS |
| Dialogue Completion | NU | PHS |
| Paraphrasing | RVHS | PHS |
| Situational Dialogue | PHS | PHS |
| Paragraph Completion | RVHS | PHS |
| Irrelevant Sentence | NU | PHS |

* VHS: Vocational High School, AHS: Anatolian High School, RVHS: Religious Vocational High School, PHS: Private High School, NU: Non-uniform DIF

## 4. DISCUSSION and CONCLUSION

In this study it was aimed to determine whether items of English test of UPE in 2016 show DIF and DBF in terms of gender and school type and examine the possible sources of bias of DIF items. MH, SIBTEST and MIMIC methods, which are based on CTT, IRT and CFA respectively, were used for analyses. It was reported in literature that detection methods are influenced by some factors such as sample size, proportion of DIF and ability difference among groups (Finch, 2005; Finch & French, 2007; Narayanan & Swaminathan, 1994). For this reason,

using different DIF detection methods increases reliability of research results. There are also some researches that suggest using more than one method to get more reliable results (Akın Arıkan, Uğurlu, & Atar, 2016; Gök, Kelecioğlu, & Doğan, 2010).

As a result of the research, it was discovered that there were differences with regard to the number of DIF items identified by three methods and the level of DIF that the items contained; however, methods were consistent in detecting uniform DIF. Some research also showed that MH and SIBTEST results comply with each other (Akın Arıkan, Uğurlu, & Atar, 2016; Narayanan & Swaminathan, 1994; Roussos & Stout, 1996).

It should be noted that there may be some advantages and disadvantages when DIF methods used in the research are examined in respect to the length of subtests. As subtests consist of 33, 15 and 12 items, they can be regarded as short tests. In their simulation study, Atalay Kabasakal, Arsan, Gök and Kelecioğlu (2014) reported that MH method had lower type I error in short tests compared with long tests and type I error with SIBTEST method increased when the length of tests decreased. From this point of view, in this research test length might have a positive impact on analyses with MH method and negative impact on analyses with SIBTEST method. Finch (2005) also reported that 20 items had an inflated type I error with MIMIC method with three parameter logistic data. However, it was discovered that 50 items had a lower type I error with three parameter logistic data. Therefore, DIF analyses with MIMIC method might be influenced negatively due to test length. Besides, Finch (2005) stated that as the size of focal group increased, power of SIBTEST and MH methods also increased. In this respect, focal group sizes might have positive impact on DIF analyses. Atalay Kabasakal, Arsan, Gök and Kelecioğlu (2014) found out that when the sizes of focal and reference groups were not equal, type I error was lower with MH method. Further, between the groups with different standard deviations, SIBTEST method had a lower type I error. In this research, the size of focal and reference groups was not equal, and there were differences between standard deviations, which may have contributed to DIF analyses.

DIF results showed that one item in translation subtest contained DIF in favor of male students. There were nine DIF items in vocabulary and grammar knowledge subtest, six DIF items in reading comprehension subtest and four DIF items in translation subtest in terms of school type. The reason why the number of DIF items by school type was higher than the number of DIF items by gender might be the serious gap between schools. Berberoğlu and Kalender (2005) investigated student achievement in Student Selection Examination (SSE) and The Programme of International Student Assessment (PISA) across years, school types and regions. It was found that student achievement changed dramatically according to school types because there is a notable difference in learning between school types. It is also supported by studies that there is a big gap between school types in Turkey (Arga, 2017; Yalçın, 2011; Yiğit, 2010). Research findings reveal the necessity to investigate the factors that cause differences between the school types and to take measures to reduce this difference.

Another finding of the study is that SIBTEST and MIMIC methods were more consistent in DBF analyses by gender compared with DBF analyses by school type. Finch (2012) noted that if group means are different on latent trait MIMIC method worked better than SIBTEST method. Therefore, the differences between methods in DBF analyses according to school type may have been caused by mean differences. Morever, another reason for the inconsistency in DBF analyses in some item bundles might be the testing only uniform DBF with MIMIC method.

In addition, experts stated that one item showing DIF in terms of gender in translation subtest was not biased and evidence of bias was found in thirteen of nineteen items that contained DIF in terms of school type. Expert opinions also revealed that four of the seven items favoring AHS were grammar items which require knowledge. According to experts, being familiar with that

type of questions may become an advantage for AHS students. Socioeconomic status, scientific terms and education based on practice and speaking were suggested as the sources of bias for items favoring PHS. Bakan Kalaycıoğlu (2008) also reported that grammar items based on knowledge were in favor of AHS and items based on reading which do not require knowledge were in favor of PHS. Evidence of bias was found in items favoring VHS or RHVS due to expressions which students from these schools might be more familiar.

In this research, DIF and DBF anaylses of English test in 2016 UPE were carried out with respect to gender and school type. Expert opinions were consulted to identify possible source of bias in items showing DIF. Student Selection and Placement Center carries out different language test every year, DIF anaylses for these tests can be performed and a pattern for language tests may be formed in terms of bias sources.

Testlets which are frequently used as reading comprehension items in languge tests might be examined in terms of DIF. Influence of different booklets on DIF can be studied as well.

**Acknowledgements**

This research was produced from master's thesis titled "An Investigation of Item Bias of English Test in 2016 Year Undergraduate Placement Exam" by Rabia Akcan in supervising of Kübra Atalay Kabasakal, and prepared in the Educational Measurement and Evalaution Program, Hacettepe University,Turkey.

**ORCID**

Rabia Akcan https://orcid.org/0000-0003-3025-774X

Kübra Atalay Kabasakal https://orcid.org/0000-0002-3580-5568

## 5. REFERENCES

Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing, 24*(1), 7-36. DOI: 10.1177/0265532207071510

Akın Arıkan, Ç., Uğurlu, S., & Atar, B. (2016). A DIF and bias study by using MIMIC, SIBTEST, Logistic Regression and Mantel-Haenszel methods. *Hacettepe University Journal of Education, 31*(1), 34-52. DOI:10.16986/HUJE.2015014226

Arga, B. (2017). Gender and student achievement in Turkey: School types and regional differences based on PISA 2012 data (Master's Thesis). İhsan Doğramacı Bilkent University, Ankara.

Atalay Kabasakal, K., Arsan, N., Gök, B., & Kelecioğlu, H. (2014). Comparing performances (Type I error and Power) of IRT Likelihood Ratio SIBTEST and Mantel-Haenszel methods in the determination of differential item functioning. *Educational Sciences: Theory & Practice, 14*(6), 2186-2193. DOI: 10.12738/estp.2014.6.2165

Bakan Kalaycıoğlu, D. (2008). Öğrenci Seçme Sınavı'nın madde yanlılığı açısından incelenmesi [Item bias analysis of the University Entrance Examination]. (Doctoral Dissertation). Hacettepe University, Ankara.

Berberoğlu, G., & Kalender, İ. (2005). Öğrenci başarısının yıllara, okul türlerine, bölgelere göre incelenmesi: ÖSS ve PISA analizi [Investigation of student achievement across years, school types and regions: The SSE and PISA anaylses]. *Eğitim Bilimleri ve Uygulama, 4*(7), 21-35.

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

Chalmers, R. P. (2017). Improving the crossing-SIBTEST statistic for detecting non-uniform DIF. *Psychometrika*. DOI: 10.1007/s11336-017-9583-8

Chalmers, R. P. (2018). *mirt,version 1.27.1: Multidimensional item response theory.* Retrieved from https://cran.r-project.org/web/packages/mirt/index.html

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement Issues and Practice, 17*(1), 31-44.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*(4), 465-484.

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. DOI: 10.1177/0146621605275728

Finch, H. (2012). The MIMIC model as a tool for differential bundle functioning detection. *Applied Psychological Measurement, 36*(1), 40-59. DOI: 10.1177/0146621611432863

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

Fox, J. (2016). *polycor,version 0.7-9: Polychoric and polyserial correlations.* Retrieved from https://cran.r-project.org/web/packages/polycor/index.html

Gierl, M. J., Bisanz, J., Bisanz, G. L., Boughton, K. A., & Khaliq, S. N. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement, 20*(2), 26-36.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A Confirmatory analysis. *Journal of Educational Measurement, 38*(2), 164-187.

Gök, B., Kelecioğlu, H., & Doğan, N. (2010). Değişen madde fonksiyonunu belirlemede Mantel–Haenszel ve Lojistik Regresyon tekniklerinin karşılaştırılması [The comparison of Mantel Haenszel and Logistic Regression tecniques in determining the differential item functioning]. *Eğitim ve Bilim, 35*(156).

Hallquist, M., & Wiley, J. (2018). *MplusAutomation,version 0.7-2: An R package for facilitating large-scale latent variable analyses in Mplus.* Retrieved from https://cran.r-project.org/web/packages/MplusAutomation/index.html

Holland , P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H. I. Braun, *Test Validity* (pp. 129-145). Hillsdale NJ: Erlbaum.

Kan, A. (2007). Test yansızlığı: H.Ü. Yabancı dil muafiyet sınavının cinsiyete ve bölümlere göre DMF analizi [Test fairness: DIF analysis across gender and department of H.U foreign language proficiency examination]. *Eurasian Journal of Educational Research*(29), 45-58.

Karakaya, İ. & Kutlu, Ö. (2012). Seviye belirleme sınavındaki Türkçe alt testlerinin madde yanlılığının incelenmesi [An investigation of item bias in Turkish sub tests in Level Determination Exam]. *Eğitim ve Bilim 37*(*165*).

Li, H.-H., & Stout, W. (1996). A new procedure for detection of crosssing DIF. *Psychometrika, 61*(4), 647-677.

Lin , J., & Wu, F. (2003). Differential performance by gender in foreign language testing. *Paper presented at the Annual Meeting of the National Councilon Measurement in Education.*

Magis, D., Beland, S., & Raiche , G. (2016). *difR, version 4.7:Collection of methods to detect dichotomous differential item functioning(DIF).* Retrieved from https://cran.r-project.org/web/packages/difR/index.html

Mcnamara, T., & Roever, C. (2006). Psychometric approaches to fairness: Bias and DIF. *Language Learning, 56*(S2), 81-128.

Muthen, L. K., & Muthen, B. O. (1998). Mplus user's guide. Los Angeles.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-haenszel and Simultaneous item bias procedures for detecting differential item functioning. *Applied*

*Psychological Measurement, 18*(4), 315-328.

Osterlind, S. J. (1983). *Test item bias.* Sage Publications, Inc.

Raiche, G., & Magis, D. (2015). *nFactors, version 2.3.3: Parallel analysis and non graphical solutions to the Cattell.* Retrieved from https://cran.r- project.org/web/packages/nFactors/index.html

Rosseel, Y. (2017). *lavaan,version 0.5-23.1097: Latent variable analysis.* Retrieved from https://cran.r-project.org/web/packages/lavaan/index.html

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*(2), 215-230.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/ DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159-194.

Willse, J. T. (2018). *CTT,version 2.3.2: Classical test theory functions.* Retrieved from https://cran.r-project.org/web/packages/CTT/index.html

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with Multiple Indicator Multiple Cause models. *Applied Psychological Measurement, 35*(5), 339-361. DOI: 10.1177/0146621611405984

Yalçın, S. (2011). Türk öğrencilerin PISA başarı düzeylerinin veri zarflama analizi ile yıllara göre karşılaştırılması[The comparison of Turkish students' PISA achievement levels in relation to years via data envelopment analysis]. (Master's Thesis). Ankara University, Ankara.

Yiğit, S. (2010). PISA matematik alt test sorularına verilen cevapların bazı faktörlere göre incelenmesi (Kocaeli-Kartepe örneği) [The analysis of the answers to PISA maths subtest questions according to certain factors (Kocaeli-Kartepe case)]. (Master's Thesis). Sakarya University, Sakarya.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. P. W. Holland, & H. Wainer içinde, *Differential Item Functioning* (s. 337-347). Hillsdale NJ:Erlbaum.

Zumbo, B. D. (1999). *A Handbook on the theory and methods of differencial item functioning (DIF):Logistic Regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Ottowa on Directorate of Human Resources Research and Evaluation,Deparment of National Defense.