KURAM VE UYGULAMADA EĞİTİM BİLİMLERİ EDUCATIONAL SCIENCES: THEORY & PRACTICE

**Received:** October 29, 2015
**Revision received:** December 18, 2015
**Accepted:** January 6, 2016
**OnlineFirst:** February 28, 2016


**www.estp.com.tr**
**DOI** 10.12738/estp.2016.1.0329 • February 2016 • 16(1) • 319-330

***Research Article***

# Examining Differential Item Functions of Different Item Ordered Test Forms According to Item Difficulty Levels

Ömay Çokluk[1]
*Ankara University*

Emrah Gül[2]
*Hakkari University*

Çilem Doğan-Gül[3]
*Ankara University*

**Abstract**

The study aims to examine whether differential item function is displayed in three different test forms that have item orders of random and sequential versions (easy-to-hard and hard-to-easy), based on Classical Test Theory (CTT) and Item Response Theory (IRT) methods and bearing item difficulty levels in mind. In the correlational research, the data from a total of 578 seventh graders were gathered using an Atomic Structures Achievement Test. R programming language and "difR" package were employed for all the analyses. As a result of the analyses, it was concluded that a comparison of IRT- and CTT-based methods indicate a greater number of items with distinctively significant differential item functioning. Different item ordering leads students at the same ability levels to display different performances on the same items. As a result, it is found that item order differentiates the probability of correct response to the items for those at the same ability levels. A test form of sequential easy-to-hard questions brings more advantages than that of a hard-to-easy sequence or a random version. The findings show that it is essential to arrange tests that are employed to make decisions about people in consideration with psychometric principles.

**Keywords**

Achievement test • Test form • Item order • Item difficulty • Classical test theory • Item response theory • Differential item function • R programming language

1 **Correspondence to:** Ömay Çokluk (PhD), Department of Measurement and Evaluation, Faculty of Educational Sciences, Ankara University, Ankara Turkey. Email: cokluk@education.ankara.edu.tr

2 Department of Measurement and Evaluation, Hakkari University, Hakkâri Turkey. Email: emrahgul@hakkari.edu.tr

3 Department of Measurement and Evaluation, Ankara University, Ankara Turkey. Email: cdogan@ankara.edu.tr

**Citation:** Çokluk, Ö., Gül, E., & Doğan-Gül, Ç. (2016). Examining differential item functions of different item ordered test forms according to item difficulty levels. *Educational Sciences: Theory & Practice, 16*, 319-330. http://dx.doi.org/10.12738/estp.2016.1.0329

Psychological and educational tests are frequently used to explore individual academic performances, educational needs, and curriculum assessment. Results that are obtained from these tests form the basis for critical decisions to get to know individuals, to employ or place them in institutions or schools, and to select, guide and assess people. As a result, it is essential to prove empirically that test scores have high validity and reliability. What is more, ongoing decisions taken by individual or organizational test developers, practitioners, and interpreters according to test scores depend on developing and implementing eligible methods to examine test development and psychometric qualifications (Camilli & Shepard, 1994; Holland & Wainer, 1993).

While carrying out large-scale assessments in Turkey, such as Transition to Higher Education Examination (YGS) and Public Personnel Selection Examination (KPSS), test forms are produced with different item orders for each examination, and these forms are presented as "personally identifiable booklets." This is considered to have a high potential in giving adverse outcomes for examinees although the primary reason for such an application is to prevent cheating in examinations. The approach shown in the item sequencing process for a test form has a sequence of easy-to-hard questions. In other words, starting a test with easy items or increasing the difficulty level of items through the test is a general principle in measurement and evaluation. Disregarding the principle might lead to outcomes such as increased anxiety of examinees, loss of self-confidence, and disturbance of mental integrity. Taking test forms of different item orders, for instance, may cause examinees to have different anxiety levels: an easy-to-hard test might cause lower anxiety levels than a hard-to-easy test. Another potential problem is the disturbance of content integrity that is caused by different item orders and obstructs mental process of examinees, shortens time of concentration, and hampers focusing on tests. This leads to lower motivation and self-esteem disturbance, and thus, adversely influences test performance (Ankara University, 2011).

The application of "personally identifiable booklets" in nation-scale examinations in Turkey may produce negative outcomes, particularly for psychometric qualifications of tests and for test takers, such as "breaking of the equivalence principle in examinations," "negatively influenced psychology of examinees," and "low performance." Therefore, such examinations can cause biased measurement results both in favor and disfavor of examinees. As the item orders are different due to variance in forms, this alone could cause students to consider items "harder" or "easier" (Balch, 1989; Impara & Foster, 2006; Laffitte, 1984; Pettijohn & Sacco, 2007). Some reviews of test item order studies have concluded that item order does not influence student test performance (Barcikovski & Olsen, 1975; Carlson & Ostrosky, 1992; Gerov, 1980; Klosner & Gellman, 1973; Tippest & Benson, 1989). However, when research on item order in test forms is reviewed, it is clear that most studies have been based on Classical Test Theory (CTT). Some studies concluded that test item order influences test scores, item parameters, and completion time (Balch, 1989; Picou & Milhomme, 1997).

Two issues still important in measurement and evaluation are a comparison between scores from different test forms that have been used to measure the same quality and the exploration of similar functioning of items/tests in different sub-groups. Research on differential item function (DIF) is needed to study and handle the issues further. DIF means the probability differentiation of a correct response to a given item by people in different groups because of a characteristic irrelevant to the measured construct (e.g., gender, ethnicity, and so on) in a comparison between ability levels to be measured by an item (Zumbo, 1999). In other words, the probability of a correct response to an item without DIF by those at the same ability level is the same, and people respond to the item regardless of their own characteristics or group membership (Millsap & Everson, 1993). Probability differentiation is called "item effect" if it is caused by individual ability levels and is necessary to show differences between the ability levels of people. However, if probability differentiation of correct response is observed to be the result of any factors irrelevant to the measured construct such as gender and ethnicity rather than abilities, this results in advantages for certain groups and disadvantages for others in a given item because it means other characteristics are involved in the measurement process.

In the literature, it is obvious that methods for DIF detection are classified as methods based on two main measurement theories: CTT and Item Response Theory (IRT). The CTT- and IRT-based methods that are employed for the research are briefly introduced below.

In the CTT-based "Transformed Item Difficulty (TID)" method, an item with DIF means the item difficult varies for test takers who get the item in groups. First, p values of both the reference group and focal group are calculated, and Z values are obtained by subtracting p values from 1 (1 − p). The obtained result is transformed into the delta scale, with a mean of 13 and a standard deviation of 4. The difference between the delta values of the reference group and the focal group provides data on the DIF level of an item (Camilli & Shepard, 1994; Osterlind, 1983; Santelices & Wilson, 2012).

Another CTT-based method is the "Mantel–Haenszel (MH)" method, which is based on statistics and the basis of equivalent test takers in sub-groups according to total test scores. The test performances of the two equivalent groups are compared according to odds ratios (Angoff, 1993; Camilli & Shepard, 1994; Osterlind, 1983). In "Logistic Regression (LR)," another CTT-based method, DIF analyses are performed using item scores as the dependent variables and group scores and total test scores as the independent variables.

Two IRT-based methods are used in this study: "Lord's Chi-square ()" and "Raju's Area." In Lord's Chi-square calculation, first the item parameters of sub-groups are estimated and covariances are computed; then, Lord's Chi-square statistics are obtained using the scaled parameters and covariance values (Camilli & Shepard, 1994). DIF is determined by comparing the observed and the expected values (Osterlind, 1983).

In the other IRT-based method, "Raju's Area," item characteristic curves are examined. Here, the underlying logic is that items of the same parameter values need to have similar item characteristic curves. The difference between the item characteristic curves drawn for both sub-groups is calculated using a square measure. A difference between item characteristic curves naturally indicates the detection of DIF (Camilli & Shepard, 1994).

As mentioned above, tests form the empirical basis of critical decisions, and it is crucial that a test is unbiased, without providing any sub-groups (e.g., group of girls, or a group that takes tests with different item orders, and so on) with advantages or disadvantages. A biased test means it has systematic errors or, in other words, its validity is seriously affected (Camilli & Shepard, 1994; Holland & Wainer, 1993). As a result, it is essential to increase the number of studies on the effect of item order and to produce empirical proof. When the literature is reviewed, it is obvious that there have been CTT-based examinations in a restricted number of recent studies on item orders in test forms and further research based on IRT methods is needed. In this study, it is considered that incorporating the methods based on the two theories will be beneficial because of IRT's sample free estimation (Hambletton & Swaminathan, 1989) and ability to scale people at the same scale level, and similar advantages could contribute to an increased amount of empirical proof in the DIF issue.

In the light of the above mentioned arguments, the intention of this research is to examine and compare whether the items in three different test forms [sequential easy-to-hard (EH), hard-to-easy (HE) and random (R)] display DIF, with different CTT and IRT methods and bearing item difficulty levels in mind.

## Method

A correlational research method was used in this study, which attempted to examine whether items in test forms of different sequential or random versions according to item difficulty displayed DIF, based on CTT and IRT methods.

### Research Group

The research data[4] were gathered from a total of 578 seventh graders chosen from seven schools in Ankara, the capital city of Turkey, in the 2013–2014 academic year. In the application process of the three forms (EH, HE and R), we attempted to balance the numbers in each group. Distribution of the students in the study group according to schools and types of test forms is presented in Table 1.

4 The research data were gathered by Research Assistant Çilem Doğan-Gül, under the advisory of Associate Professor Ömay Çokluk, for the master thesis entitled "A Comparison of Academic Achievement Scores of Students with High and Low Anxiety Levels in Tests with Different Item Orders in Consideration with Item Difficulty", completed at the Department of Measurement and Evaluation, Faculty of Educational Sciences, Ankara University, 2014.

Table 1
*Distribution of Students in Study Group According to Schools and Different Test Forms*

| Schools | Forms | | | Total |
|---|---|---|---|---|
| | Form EH | Form HE | Form R | |
| 1 School | 39 | 39 | 38 | 116 |
| 2 School | 19 | 19 | 19 | 57 |
| 3 School | 45 | 43 | 48 | 136 |
| 4 School | 23 | 20 | 21 | 64 |
| 5 School | 4 | 6 | 7 | 17 |
| 6 School | 53 | 50 | 52 | 155 |
| 7 School | 12 | 9 | 12 | 33 |
| Total | 195 | 186 | 197 | 578 |

## Instrument

The "Atomic Structures Achievement Test," developed by Tağ (2012) for science and technology courses, is used for data gathering. The test aims to explore student academic achievement in the subject of atomic structures, included in the science and technology curriculum. The test consists of 20 multiple choice items, each with four alternatives. The KR-20 reliability of test scores is 0.73. The reason why the test was applied in this study is that it has been developed in accordance with "test development" principles and has items of three difficulty levels: easy, moderate and hard. Item difficulty indices are presented in Table 2.

Table 2
*Item Difficulty Indices in Atomic Structures Achievement Test*

| Item | Difficulty Level | $p_j$ |
|---|---|---|
| 1 | Easy | 0.60 |
| 3 | Easy | 0.64 |
| 7 | Easy | 0.80 |
| 8 | Easy | 0.64 |
| 10 | Easy | 0.89 |
| 12 | Easy | 0.73 |
| 19 | Easy | 0.69 |
| 11 | Hard | 0.15 |
| 15 | Hard | 0.26 |
| 14 | Hard | 0.31 |
| 18 | Hard | 0.31 |
| 4 | Hard | 0.33 |
| 17 | Hard | 0.36 |
| 2 | Hard | 0.39 |
| 20 | Moderate | 0.40 |
| 5 | Moderate | 0.45 |
| 13 | Moderate | 0.46 |
| 16 | Moderate | 0.53 |
| 9 | Moderate | 0.54 |
| 6 | Moderate | 0.56 |

When Table 2 is examined, it is clear that the difficulty indices of the test items range from 0.15 to 0.89. Item Difficulty Index (IDI) shows the percentage of correct response in a given group and ranges from 0 to 1. When IDI is closer to 0, it indicates that an item is hard, whereas a value close to 1 means the related item is easy. Within the scope of the study, the following ranges are the reference points for the classification of the items according to difficulty level: a) 0.00 to 0.39 = Hard, b) 0.40 to 0.59 = Moderate and c) 0.60 to 1.00 = Easy (Kubiszyen & Borich, 2013).

In the study, three test forms of different item orders are formed and applied, bearing item difficulty indices in mind. The item order in the first form is sequential low difficulty (easy) to high difficulty (hard). This is called the EH Form. The item order in the second form is sequential high difficulty (hard) to low difficulty (easy). This is called the HE Form. The item order in the third form is totally random, and it is called the Random (R) Form. After the rearrangement process, the forms are randomly given to the 578 students in the study group, as follows: 195 EH, 186 HE and 197 R.

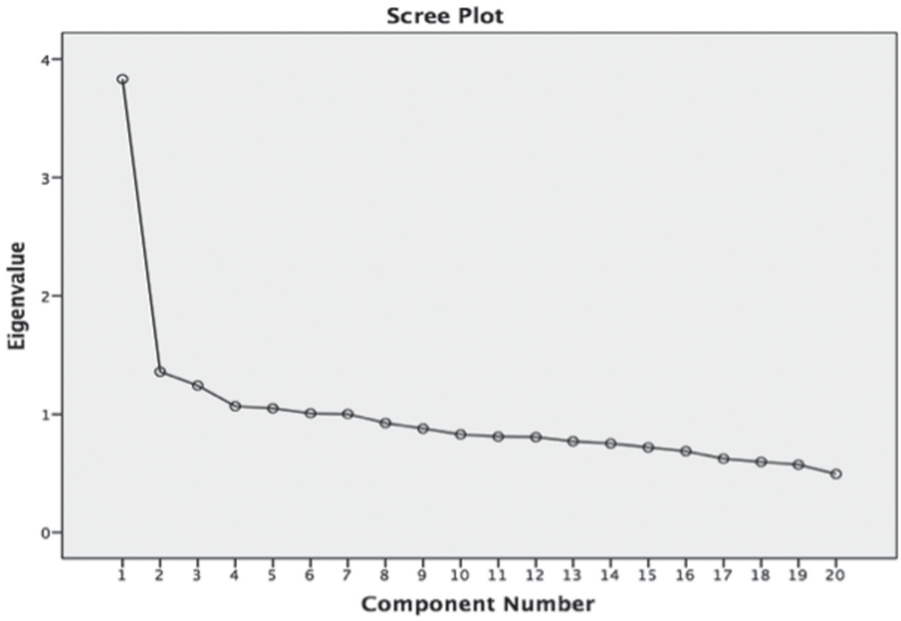
## Procedure

This part discusses the examinations of IRT assumptions for the IRT-based methods that are used for DIF detection, before the presenting data analysis information.

**Testing assumptions.** We explored whether the data construct would meet the assumptions of IRT for the analyses based on the theory. First, the type of model to be employed in the estimation of item parameters was decided. As a result of the model-data fit analyses, when the number of parameters was 20 (under 1PLM), the −2 Loglikelihood value was found to be 6559.1952. 1PLM is a model that merely takes item difficulties into account. Twenty parameters were produced in which solely item hardies were considered for each item as the achievement test consisted of 20 items. Under 2PLM, the −2Loglikelihood value decreased to 6328.6479. The decrease was significant for 20 degrees of freedom in the Chi-square critical value. As 2PLM is a model that takes differentiation into account, as well as item discrimination, a total of 40 parameters (20 difficulty and 20 discrimination parameters) were produced for 20 items. Under 3PLM, the −2 Loglikelihood value was 6300.0453; however, the decrease in the value was not significant. Besides 2PLM, 3PLM is a model that takes guessing parameters into account. Yet, when we shifted from 2PLM to 3PLM, the insignificant decrease in the −2 Loglikelihood value led to a decision to apply 2PLM in the estimation of item parameters.

In the study, Exploratory Factor Analysis (EFA) was used to test whether the Atomic Structures Achievement Test had a unidimensional structure to examine whether unidimensionality, another assumption of IRT, was met. As data were obtained dichotomously from the test (1–0), factor analysis studies were conducted based on a tetrachoric correlation matrix. The Scree plot, which was examined to

decide the number of factors in EFA, is presented in Graphic 1. On examination, Graphic 1 clearly reveals that there is a dominant factor. This case shows that unidimensionality, an assumption of IRT, has been met. In the literature, it is accepted that both assumptions are proven because meeting unidimensionality indicates that local independence is also met (Hambletton & Swaminathan, 1989).




*Graphic 1.* Scree plot.

As a result, the following IRT- and CTT-based methods are used in the study for DIF detection: TID, MH, and Logistic Regression (CTT-based) and Lord's Chi-square and Raju's Area (IRT-based). In the analyses based on IRT, estimates were made according to 2PLM. R programming language and "difR" package were employed in all the analyses. difR, which was developed using R programming language, is a package that can detect DIF with both CTT and IRT methods (Magis, Beland, & Raiche, 2015).

Only DIF detection in items is not satisfactory: DIF levels must also be determined. The classification recommended by Educational Testing Service is widely recognized and employed in the field to objectively interpret DIF levels. The following are generally defined DIF levels although there could be certain changes when specific restrictions of methods are considered (Zwick, 2012):

A: Acceptable DIF

B: Moderate DIF

C: High DIF

## Findings

The present study examined whether the items in three different test forms [sequential EH, HE and random (R) according to item difficulty] display DIF using CTT-based methods, Lord's Chi-square and Raju's Area, and IRT-based methods, Logistic Regression, MH, and TID.

Table 3 displays the findings of whether the test item orders in the forms of sequential EH and HE versions lead to the detection of DIF in the analyses with the CTT and IRT based methods.

Table 3
*DIF Results of Items in the Test Forms Easy-to-Hard and Hard-to-Easy Versions, based on CTT and IRT Methods*

| | CTT-Based Methods | | | | | | IRT-Based Methods | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **MH** | DIF | **TID** | DIF | **LR** | DIF | **Raju's Area** | DIF | **Lord's Chi-square** | DIF |
| Items | | Level | | Level | | Level | | Level | | Level |
| 1 | 0.975 | A | 0.121 | A | 0.892 | A | **0.010** | **C** | **0.028** | **B** |
| **2** | **0.033** | **B** | **0.844** | **B** | 0.059 | A | **0.001** | **C** | **0.028** | **B** |
| 3 | 0.557 | A | 0.406 | A | 0.504 | A | **0.011** | **B** | **0.048** | **B** |
| 4 | 0.877 | A | 0.216 | A | 0.530 | A | **0.000** | **C** | **0.003** | **C** |
| 5 | 0.632 | A | 0.169 | A | 0.865 | A | **0.009** | **C** | **0.033** | **B** |
| 6 | 0.428 | A | 0.251 | A | 0.729 | A | 0.051 | A | 0.192 | A |
| 7 | 0.235 | A | 0.466 | A | 0.462 | A | 0.038 | B | 0.195 | A |
| 8 | 0.462 | A | 0.042 | A | 0.961 | A | **0.003** | **C** | **0.010** | **C** |
| 9 | 0.498 | A | 0.148 | A | 0.331 | A | 0.334 | A | 0.548 | A |
| 10 | 0.687 | A | 0.073 | A | 0.013 | B | **0.002** | **C** | **0.001** | **C** |
| 11 | 0.183 | A | 0.343 | A | 0.503 | A | **0.007** | **C** | **0.020** | **B** |
| 12 | 0.652 | A | 0.106 | A | 0.495 | A | 0.852 | A | 0.698 | A |
| 13 | 0.277 | A | 0.304 | A | 0.142 | B | 0.444 | A | 0.196 | A |
| 14 | 0.467 | A | 0.199 | A | 0.363 | A | **0.004** | **C** | **0.004** | **C** |
| 15 | 0.862 | A | 0.032 | A | 0.974 | A | **0.013** | **B** | 0.059 | A |
| 16 | 0.939 | A | 0.060 | A | 0.839 | A | **0.025** | **B** | 0.132 | A |
| 17 | 0.520 | A | 0.442 | A | 0.495 | A | 0.051 | A | 0.269 | A |
| **18** | **0.057** | **B** | **0.864** | **B** | **0.032** | **B** | 0.135 | A | 0.229 | A |
| 19 | 0.261 | A | 0.774 | **B** | **0.028** | **B** | 0.373 | A | 0.114 | A |
| 20 | 0.627 | A | 0.399 | A | 0.638 | A | **0.047** | **B** | 0.153 | A |
| The number of items with DIF | | 2 | | 2 | | 3 | | 13 | | 9 |

An examination of Table 3 shows that there are two items (Items 2 and 18) with significant DIF (Level B and C) in at least two methods based on CTT. In the IRT-based methods, that number increases to nine (Items 1, 2, 3, 4, 5, 8, 10, 11, and 14) DIF. Item 2 displays DIF in the examinations based on the two theories. Additionally, it is observed that the group with EH Forms is more advantaged, particularly in the first five test items, than those with HE Forms. This also shows that the group with the HE test forms is disadvantaged. Table 4 displays the findings of whether the test

item orders in the forms of the sequential EH and random (R) versions lead to the detection of DIF in the analyses with the CTT- and IRT-based methods.

Table 4
*DIF Results of Items in the Test Forms Easy-to-Hard and Random Versions, based on CTT and IRT methods*

| | CTT-Based Methods | | | | | | IRT-Based Methods | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Items | **MH** | DIF | **TID** | DIF | **LR** | DIF | **Raju's Area** | DIF | **Lord's Chi-square** | DIF |
| | | Level | | Level | | Level | | Level | | Level |
| 1 | 0.097 | B | 0.434 | A | 0.170 | A | 0.143 | A | 0.360 | A |
| 2 | 0.123 | A | 0.498 | A | 0.097 | B | 0.011 | B | 0.113 | A |
| 3 | 0.030 | B | 0.764 | B | 0.130 | A | **0.045** | **B** | **0.063** | **B** |
| 4 | 0.521 | A | 0.251 | A | 0.571 | A | **0.010** | **B** | **0.051** | **B** |
| 5 | 0.634 | A | 0.040 | A | 0.568 | A | **0.035** | **B** | **0.086** | **B** |
| **6** | **−0.004** | **C** | **−1.046** | **C** | **0.001** | **C** | **0.016** | **B** | **0.055** | **B** |
| 7 | 0.352 | A | 0.212 | A | 0.470 | A | 0.112 | A | 0.240 | A |
| 8 | 0.917 | A | 0.139 | A | 0.980 | A | **0.050** | **B** | 0.152 | A |
| **9** | **0.005** | **C** | **−0.819** | **B** | **0.019** | **B** | **0.035** | **B** | **0.030** | **B** |
| 10 | 0.771 | A | 0.027 | A | 0.033 | B | **0.003** | **C** | **0.004** | **C** |
| 11 | 0.240 | A | 0.348 | A | 0.213 | A | 0.201 | A | 0.440 | A |
| 12 | 0.754 | A | 0.109 | A | 0.778 | A | 0.949 | A | 0.917 | A |
| 13 | 0.925 | A | 0.026 | A | 0.887 | A | 0.477 | A | 0.780 | A |
| 14 | 0.812 | A | 0.070 | A | 0.893 | A | **0.086** | **C** | **0.248** | **A** |
| 15 | 0.519 | A | 0.309 | A | 0.274 | A | **0.032** | **B** | **0.031** | **C** |
| 16 | 0.444 | A | 0.301 | A | 0.725 | A | 0.126 | A | 0.331 | A |
| 17 | 0.933 | A | 0.092 | A | 0.693 | A | 0.913 | A | 0.992 | A |
| 18 | 0.142 | A | **0.551** | **B** | **0.009** | **C** | 0.374 | A | 0.480 | A |
| 19 | 0.799 | A | 0.589 | B | 0.543 | A | 0.833 | A | 0.655 | A |
| 20 | 0.587 | A | 0.116 | A | 0.643 | A | 0.645 | A | 0.908 | A |
| The number of items with DIF | | 4 | | 6 | | 5 | | 10 | | 7 |

Table 4 shows that there are three items (6, 9, and 10) with significant DIF (Level B and C) in at least two methods based on CTT. In the IRT-based methods, this number increases to eight (Items 2, 3, 4, 5, 6, 9, 10, 14, and 15) DMF. As it is seen, items 6, 9, and 10 display DIF in the detections based on the two theories. Additionally, it is observed that the group with EH Forms is more advantaged, particularly in the first five test item, than those with R Forms. This also shows that the group with the random test forms is disadvantaged.

Table 5 displays the results whether the test item orders in the forms of the sequential HE and random (R) versions lead to the detection of DIF in the analyses with the CTT- and IRT-based methods.

Table 5
*DIF Results of Items in the Test Forms of the Hard-to-Easy and Random Versions, based on CTT and IRT methods*

| Items | CTT-Based Methods | | | | | | IRT-Based Methods | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **MH** | DIF | **TID** | DIF | **LR** | DIF | **Raju's Area** | DIF | **Lord's Chi-square** | DIF |
| | | Level | | Level | | Level | | Level | | Level |
| 1 | 0.075 | A | 0.336 | A | 0.089 | A | **−0.036** | **B** | 0.084 | A |
| 2 | 0.781 | A | 0.257 | A | 0.927 | A | 0.153 | A | 0.343 | A |
| 3 | 0.442 | A | 0.361 | A | 0.747 | A | 0.228 | A | 0.387 | A |
| 4 | 0.999 | A | 0.037 | A | 0.944 | A | 0.057 | A | 0.145 | A |
| 5 | 0.529 | A | 0.136 | A | 0.576 | A | 0.503 | A | 0.738 | A |
| **6** | **0.032** | **B** | **0.819** | **B** | **0.012** | **B** | **−0.029** | **B** | **0.003** | **C** |
| 7 | 0.711 | A | 0.203 | A | 0.485 | A | 0.950 | A | 0.997 | A |
| 8 | 0.928 | A | 0.155 | A | 0.950 | A | 0.108 | A | 0.258 | A |
| 9 | 0.041 | B | 0.658 | B | 0.072 | B | 0.133 | A | 0.314 | A |
| 10 | 0.599 | A | 0.077 | A | 0.832 | A | −0.184 | A | 0.379 | A |
| **11** | **0.995** | A | 0.018 | A | 0.307 | A | **0.009** | **C** | **0.011** | **B** |
| 12 | 0.309 | A | 0.212 | A | 0.253 | A | 0.689 | A | 0.533 | A |
| 13 | 0.313 | A | 0.261 | A | 0.034 | B | 0.394 | A | 0.147 | A |
| 14 | 0.526 | A | 0.083 | A | 0.583 | A | −0.048 | B | 0.075 | A |
| 15 | 0.622 | A | 0.303 | A | 0.297 | A | 0.529 | A | 0.789 | A |
| 16 | 0.461 | A | 0.258 | A | 0.676 | A | 0.262 | A | 0.483 | A |
| 17 | 0.479 | A | 0.299 | A | 0.531 | A | **−0.032** | **B** | **0.046** | **B** |
| 18 | 0.500 | A | 0.230 | A | 0.198 | A | **−0.019** | **B** | **0.035** | **B** |
| 19 | 0.363 | A | 0.130 | A | 0.245 | A | 0.598 | A | 0.443 | A |
| 20 | 0.919 | A | 0.239 | A | 0.989 | A | 0.036 | B | 0.154 | A |
| The number of items with DIF | | 2 | | 2 | | 3 | | 6 | | 4 |

Table 5 shows that there is a single item (6) with high DIF (Level B and C) in at least two methods according to the CTT-based methods and the number of items increases to four (Items 6, 11, 17, and 18) according to the IRT-based methods. It is clear that Item 6 is found to display DIF in the analyses based on the two theories. Additionally, the finding show that those with the test forms of the sequential HE version are more disadvantaged than those who are given the random test form.

## Discussion

The study examines whether the items in three tests with different item orders display DIF, based on CTT and IRT methods. The obtained results have shown that CTT and IRT give different results, but results of both CTT and IRT, produced with different methods on their own merits, are mostly consistent. Two criteria are used to decide whether an item has DIF: a) being characterized as "DIF detected" in analyses with at least two methods as mentioned in the literature and b) B or C level DIF.As a result of the analyses in the study, it is concluded that the IRT-based methods indicate far more items with DIF than

the CTT-based methods. When the probability of Type I error and the calculation logic of methods based on different theories are considered, it could be interpreted that the IRT-based methods give more sensitive results. Different item orders have led the participating students at the same ability level to display different performances on the same test items. It is concluded that the group with the sequential EH test is more advantaged, particularly in the first items of the test, than the group with the sequential HE version. The findings also show that the group that starts the tests with hard items is disadvantaged. In other words, starting from easy test items ensures higher scores and better individual performances. Conversely, the fact that the examinees with the random test form are more disadvantaged than those with the sequential version of the EH test forms has shown that the best test performance is when the item order is EH, in agreement with the previous finding. The final result obtained by the study is that the participant group with the random form is found to be more advantaged than the group with the sequential HE test forms. This leads to the conclusion that groups with the sequential HE test forms are disadvantaged in any condition, when compared to groups with other test forms.

In this study, it is concluded that the arrangement of the items in the Atomic Structures Achievement Test, as three different forms according to item difficulty, has caused changes in the perceptions of certain items and thus, differentiation in correct response because examinees at the same ability level vary in their performances when they take different test forms. In other words, as test item orders are different because forms vary, this could cause examinees to consider items "harder" or "easier" (Balch, 1989; Impara & Foster, 2006; Laffitte, 1984; Pettijohn & Sacco, 2007).

Some studies on the same issue have obtained results that reveal item discrimination indices vary although the present study has obtained findings that show the hardiness and/or perceived hardiness of test forms of different item orders vary (Brenner, 1964; Carlson & Ostrosky, 1992; Doğan-Gül, 2014).

As a result, different item orders affect the probability of correct responses to a given item by those at the same ability level, as well as item parameters. It is considered to be advantageous for examinees to have an EH test, and even a random version is more advantageous than a HE test. This reveals the necessity for test arrangements in accordance with psychometric principles in order to make decisions about people. Therefore, it is now more likely that different test forms called "personally identifiable booklets" in national-scale examinations in Turkey will affect test performance and cause an increase in test anxiety level.

It is essential that educational and psychological tests should not be influenced by any qualities except individual abilities and they should remain unbiased without advantageous or disadvantageous outcomes for any groups. This case reveals the necessity for careful consideration of basic principles of measurement in any kind of test practices.

## Reference


Angoff, W. H. (Ed.). (1993). *Perspectives on differential item functioning methodology*. *Differential item functioning*. New Jersey, NJ: Lawrence Erlbaum Associates Publishers.

Ankara Üniversitesi Eğitim Bilimleri Fakültesi. (2011). *Ankara Üniversitesi Eğitim Bilimleri Fakültesi'nin YGS hakkında görüşü* [*The view of Faculty of Education, Ankara University about the Higher Education Entrance Examination*]. Retrieved from http://rigel2.cc.ankara.edu.tr/dyr.php?id=11705

Balch, W. R. (1989). Item order affects performance on multiple-choice exams. *Teaching of Psychology*, *16*(2), 75–77.

Brenner, M. H. (1964). Test hardy, reliability, and discriminations of item hardy order. *Journal of Applied Psychology, 48*(2), 98–100.

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

Carlson, J. L., & Ostrosky, A. L. (1992). Item sequence and student performance on multiple-choice exams: Further evidence. *Journal of Economic Education*, *2*(3), 232–235.

Doğan-Gül, Ç. (2014). *Madde güçlüklerine göre farklı sıralanan testlerde düşük ve yüksek kaygılı öğrencilerin akademik başarıları puanlarının karşılaştırılması* [*A comparison of academic achievement scores of students with high and low anxiety levels in different sequence tests according to item difficulty*] (Master's thesis, Ankara University, Ankara, Turkey). Retrieved from http://tez2.yok.gov.tr/

Hambletton, R. K., & Swaminathan, H. (1989). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff Publishing.

Holland, P. W., & Wainer, H. (Ed.). (1993). *Differential item functioning*. New Jersey, NJ: Lawrence Erlbaum Associates Publishers.

Impara, J., & Foster, D. (2006) *Strategies to minimize test fraud.* In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 91–114). Mahwah, NJ: Lawrence Erlbaum Associates.

Kubiszyen, T., & Borich, G. (2013). *Educational testing and measurement: Classroom application and practice*. Hoboken, NJ: John Wiley & Sons, INC.

Laffittee, R. G. (1984). Effects of item order on achievement test scores and students' perceptions of test hardy. *Teaching of Psychology*, *77*(4), 212–214.

Magis, D., Beland, S., & Raiche, G. (2015). *Collection of methods to detect dichotomous Differential Item Functioning (DIF). Package* 'difR'. Retrieved from https://cran.r-project.org/web/packages/difR/difR.pdf

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297–334.

Osterlind, S. (1983). *Test item bias*. Newbury Park, CA: Sage.

Pettijohn, T. F., & Sacco, M. F. (2007). Multiple-choice exam question order influences on student performance, completion time, and perceptions. *Journal of Instructional Psychology*, *34*(3), 142–149.

Picou, A., & Milhomme, A. J. (1997). The effect of random and sequential versions on student test performance. *Financial Practice & Education*, *7*, 85–90.

Santelices, M. V., & Wilson, M. (2012). On the relationship between differential item functioning and item hardy: An issue of methods? Item response theory approach to differential item functioning. *Educational and Psychological Measurement 72*(1), 5–36.

Tağ, S. M. (2012). *Atomun yapısı konusunu öğrenmede klasik yöntemler ile bilgisayar destekli öğretimin öğrenci başarısına etkileri* [The effects of classical methods and computer assisted instruction on student achievement in atomic structure learning] (Master's thesis, Fırat University, Elazığ, Turkey). Retrieved from http://tez2.yok.gov.tr/

Zumbo, B. D. (1999). *A handbook on the theory and methods of Differential Item Functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (Ordinal) item scores*. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R. J. (2012). *A review of ETS differential item functioning assessment procedures: Flagging principles, minimum sample size requirements, and criterion refinement* (Research Report). Educational Testing Service.