Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN:1309-6575 *Bahar* 2020 *Spring* 2020 *Cilt: 11- Sayı: 1* *Volume: 11- Issue: 1*

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309 – 6575

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (EPOD) yılda dört kez yayınlanan hakemli ulusal bir dergidir. Yayımlanan yazıların tüm sorumluğu ilgili yazarlara aittir.

Journal of Measurement and Evaluation in Education and Psychology (EPOD) is a national refereed journal that is published four times a year. The responsibility lies with the authors of papers.

İletişim
e-posta: epodderdergi@gmail.com
Web: https://dergipark.org.tr/tr/pub/epod

Contact
e-mail: epodderdergi@gmail.com
Web: http://dergipark.org.tr/tr/pub/epod

Dizinleme / Abstracting & Indexing

Emerging Sources Citation Index (ESCI), DOAJ (Directory of Open Access Journals), SCOPUS, TÜBİTAK TR DIZIN Sosyal ve Beşeri Bilimler Veri Tabanı (ULAKBİM), Tei (Türk Eğitim İndeksi)

Hakem Kurulu / Referee Board

Ahmet Salih ŞİMŞEK (Kırşehir Ahi Evran Üni.)
Ahmet TURHAN (American Institute Research)
Akif AVCU (Marmara Üni.)
Alperen YANDI (Abant İzzet Baysal Üni.)
Asiye ŞENGÜL AVŞAR (Recep Tayyip Erdoğan Üni.)
Ayfer SAYIN (Gazi Üni.)
Ayşegül ALTUN (Ondokuz Mayıs Üni.)
Arif ÖZER (Hacettepe Üni.)
Aylin ALBAYRAK SARI (Hacettepe Üni.)
Bahar ŞAHİN SARKIN (İstanbul Okan Üni.)
Belgin DEMİRUS (MEB)
Bengü BÖRKAN (Boğaziçi Üni.)
Betül ALATLI (Gaziosmanpaşa Üni.)
Betül TEKEREK (Kahramanmaraş Sütçü İmam Üni.)
Beyza AKSU DÜNYA (Bartın Üni.)
Bilge GÖK (Hacettepe Üni.)
Bilge BAŞUSTA UZUN (Mersin Üni.)
Burak AYDIN (Recep Tayyip Erdoğan Üni.)
Burcu ATAR (Hacettepe Üni.)
Burhanettin ÖZDEMİR (Siirt Üni.)
Celal Deha DOĞAN (Ankara Üni.)
Cem Oktay GÜZELLER (Akdeniz Üni.)
Cenk AKAY (Mersin Üni.)
Ceylan GÜNDEĞER (Aksaray Üni.)
Çiğdem REYHANLIOĞLU KEÇEOĞLU (MEB)
Cindy M. WALKER (Duquesne University)
Çiğdem AKIN ARIKAN (Ordu Üni.)
David KAPLAN (University of Wisconsin)
Deniz GÜLLEROĞLU (Ankara Üni.)
Derya ÇAKICI ESER (Kırıkkale Üni)
Derya ÇOBANOĞLU AKTAN (Hacettepe Üni.)
Devrim ALICI (Mersin Üni.)
Devrim ERDEM (Niğde Ömer Halisdemir Üni.)
Didem KEPIR SAVOLY
Didem ÖZDOĞAN (İstanbul Kültür Üni.)
Dilara BAKAN KALAYCIOĞLU (Gazi Üni.)
Dilek GENÇTANRIM (Kırşehir Ahi Evran Üni.)
Durmuş ÖZBAŞI (Çanakkele Onsekiz Mart Üni)
Duygu Gizem ERTOPRAK (Amasya Üni)
Duygu KOÇAK (Alanya Alaaddin Keykubat Üni.)
Ebru DOĞRUÖZ (Çankırı Karatekin Üni.)
Elif Bengi ÜNSAL ÖZBERK (Trakya Üni.)
Emine ÖNEN (Gazi Üni.)
Emrah GÜL (Hakkari Üni.)
Emre ÇETİN (Doğu Akdeniz Üni.)
Emre TOPRAK (Erciyes Üni.)
Eren Can AYBERK (Pamukkale Üni.)
Eren Halil ÖZBERK (Trakya Üni.)
Ergül DEMİR (Ankara Üni.)
Erkan ATALMIS (Kahramanmaras Sütçü İmam Üni.)
Ersoy KARABAY (Kirşehir Ahi Evran Üni.)
Esin TEZBAŞARAN (İstanbul Üni)
Esin YILMAZ KOĞAR (Niğde Ömer Halisdemir Üni.)
Esra Eminoğlu ÖZMERCAN (MEB)
Fatih KEZER (Kocaeli Üni.)
Fatih ORCAN (Karadeniz Teknik Üni)
Fatma BAYRAK (Hacettepe Üni.)
Fazilet TAŞDEMİR (Recep Tayyip Erdoğan Üni.)
Funda NALBANTOĞLU YILMAZ (Nevşehir Üni.)
Gizem UYUMAZ (Giresun Üni.)
Gonca USTA (Cumhuriyet Üni.)
Gökhan AKSU (Adnan Menderes Üni.)
Gözde SIRGANCI (Bozok Üni.)
Gül GÜLER (İstanbul Aydın Üni.)
Gülden KAYA UYANIK (Sakarya Üni.)
Gülşen TAŞDELEN TEKER (Hacettepe Üni.)
Hakan KOĞAR (Akdeniz Üni.)
Hakan SARIÇAM (Dumlupınar Üni.)
Hakan Yavuz ATAR (Gazi Üni.)
Halil İbrahim SARI (Kilis Üni.)
Halil YURDUGÜL (Hacettepe Üni.)
Hatice KUMANDAŞ (Artvin Çoruh Üni.)
Hülya KELECİOĞLU (Hacettepe Üni.)
Hülya YÜREKLI (Yıldız Teknik Üni.)
İbrahim Alper KÖSE (Abant İzzet Baysal Üni.)
İlhan KOYUNCU (Adıyaman Üni.)
İlkay AŞKIN TEKKOL (Kastamonu Üni.)
İlker KALENDER (Bilkent Üni.)
İsmail KARAKAYA (Gazi Üni.)
Kübra ATALAY KABASAKAL (Hacettepe Üni.)
Levent YAKAR (Kahramanmaraş Sütçü İmam Üni.)
Mehmet KAPLAN (MEB)
Melek Gülşah ŞAHİN (Gazi Üni.)
Meltem ACAR GÜVENDİR (Trakya Üni.)
Meltem YURTÇU (İnönü Üni.)
Metin BULUŞ (Adıyaman Üni)
Murat Doğan ŞAHİN (Anadolu Üni.)
Mustafa ASİL (University of Otago)
Mustafa İLHAN (Dicle Üni.)
Nagihan BOZTUNÇ ÖZTÜRK (Hacettepe Üni.)
Nail YILDIRIM (Kahramanmaraş Sütçü İmam Üni.)
Neşe GÜLER (İzmir Demokrasi Üni.)
Neşe ÖZTÜRK GÜBEŞ (Mehmet Akif Ersoy Üni.)
Nuri DOĞAN (Hacettepe Üni.)
Nükhet DEMİRTAŞLI (Emekli Öğretim Üyesi)
Okan BULUT (University of Alberta)
Onur ÖZMEN (TED Üniversitesi)
Ömer KUTLU (Ankara Üni.)
Ömür Kaya KALKAN (Pamukkale Üni.)
Önder SÜNBÜL (Mersin Üni.)
Özge ALTINTAS (Ankara Üni.)
Özge BIKMAZ BİLGEN (Adnan Menderes Üni.)
Özlem ULAŞ (Giresun Üni.)

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Mart 2020, Sayı: 11-1)

Journal of Measurement and Evaluation in Education and Psychology (March 2020, Issue: 11-1)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(1); 1-12

Performances of MIMIC and Logistic Regression Procedures in Detecting DIF *

Seçil UĞURLU ** Burcu ATAR ***

Abstract
In this study, differential item functioning (DIF) detection performances of multiple indicators, multiple causes (MIMIC) and logistic regression (LR) methods for dichotomous data were investigated. Performances of these two methods were compared by calculating the Type I error rates and power for each simulation condition. Conditions covered in the study were: sample size (2000 and 4000 respondents), ability distribution of focal group [N(0, 1) and N(-0.5, 1)], and the percentage of items with DIF (10% and 20%). Ability distributions of the respondents in the reference group [N(0, 1)], ratio of focal group to reference group (1:1), test length (30 items), and variation in difficulty parameters between groups for the items that contain DIF (0.6) were the conditions that were held constant. When the two methods were compared according to their Type I error rates, it was concluded that the change in sample size was more effective for MIMIC method. On the other hand, the change in the percentage of items with DIF was more effective for LR. When the two methods were compared according to their power, the most effective variable for both methods was the sample size.

Key Words: Differential item functioning, MIMIC model, Logistic regression, Uniform DIF, Type I error rate and power.

INTRODUCTION

Test items may be biased since they may contain constructs that are undesired to be measured along with the desired ones. Any item may also be in relation with a second or more factors other than the one which is of interest. Those factors that are irrelevant to the construct being measured may affect the performances of individuals. This issue is known as test bias. While test bias focuses on test scores and is interested in fairness of a test, item bias focuses on the relationship between answering an item correctly and group membership. And hence, item bias is related to a specific item. Differential item functioning (DIF), which is a statistical method used in item bias analysis, has been the subject of a vast majority of recent studies (Zumbo, 1999).

DIF occurs when respondents who are at the same ability level but from different groups have different item response probabilities on a specific item (Crane, Belle & Larson, 2004; Mazor, Kanjee & Clauser, 1995). In other words, the expression of DIF is that an item displays different statistical properties in different groups for individuals who are at the same ability levels (Holland & Wainer, 1993). Many methods have been developed for detecting test items with DIF. Some DIF detection methods used for dichotomously scored items are; chi-square test based on item response theory (Lord, 1980), standardization (Dorans & Kulick, 1986), Mantel-Haenszel (MH) (Holland & Thayer, 1988), item response theory likelihood ratio test (IRT-LRT) (Thissen, Steinberg & Wainer, 1988), logistic regression (LR) (Swaminathan & Rogers, 1990), simultaneous item bias test (SIBTEST) (Shealy & Stout, 1993), and multiple indicators, multiple causes (MIMIC) model (Finch, 2005; Oort, 1998).

Fleishman, Spector, and Altman (2002) mentioned in their study that when there are more than two groups, methods get very complicated for testing DIF in IRT framework. As they mentioned in their study, the MIMIC model has an advantage of including multiple exogenous variables to the analysis

* This study is based on Seçil Uğurlu's master thesis titled "Performance of Multiple Indicators Multiple Causes and Logistic Regression Procedures in Detecting Differential Item Functioning".

** Res. Assist., Hacettepe University, Faculty of Education, Ankara-Turkey, secilarslan@hacettepe.edu.tr, ORCID ID: 0000-0002-3495-7797
*** Assoc. Prof. Ph.D., Hacettepe University, Faculty of Education, Ankara-Turkey, burcua@hacettepe.edu.tr, ORCID ID: 0000-0003-3527-686X

To cite this article:
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. doi: 10.21031/epod.531509
Received: 23.02.2019
Accepted: 22.11.2019

simultaneously. Because of allowing a simultaneous analysis of several groups in a single framework, MIMIC model seems to be very useful (Muthen, 1988). This method has become an interesting research subject when its advantages on DIF researches are considered. MIMIC method is quite new with respect to the other methods mentioned above, and especially regarding dichotomous data, there are few studies in the literature involving MIMIC method (see Finch, 2005). Some recent studies on this method were conducted by Fleishman et al. (2002), Woods (2009), Wang, Shih, and Yang, (2009), Woods, Oltmanns and Turkheimer (2009), and Wang and Shih, (2010). Considering these studies, it is reasonable to investigate that under which circumstances MIMIC method is more effective in DIF detection. The aim of the current study is to compare the performance of MIMIC method with LR method - a commonly used method - in detecting items with DIF and interpret the results of these two methods. The DIF detection methods used in this study was explained in detail in the following sections:

Logistic Regression DIF Detection Method

As specified by Swaminathan and Rogers (1990), in detection of differential item functioning, LR model for the two groups of interest can be expressed as:

$$P(u_{ij}=1|\theta_{ij})= \frac{e^{(\beta_{0j}+\beta_{1j}\theta_{1j})}}{1+e^{(\beta_{0j}+\beta_{1j}\theta_{1j})}}, \; i=1, \ldots, n_j, j = 1, 2. \quad (1)$$

u_{ij}: response of ith individual in jth group to the item,

β_{0j}: intercept parameter for jth group,

β_{1j}: slope parameter for jth group,

θ_{ij}: ability of ith individual in jth group.

In Equation 1, if logistic regression curves are the same for the two groups, i.e., $\beta_{01} = \beta_{02}$ and $\beta_{11} = \beta_{12}$, no DIF is present. However, if $\beta_{11} = \beta_{12}$ and $\beta_{01} \neq \beta_{02}$, since the LR curves are parallel, it can be concluded that uniform DIF exists. If $\beta_{01} = \beta_{02}$ and $\beta_{11} \neq \beta_{12}$, since the curves are not parallel, it can be concluded that nonuniform DIF exists (Swaminathan & Rogers, 1990).

MIMIC DIF Detection Method

MIMIC method, which is newer than LR, is based on confirmatory factor analysis (CFA) (Finch, 2005). As outlined by Finch (2005), in DIF context, MIMIC model is as Equation 2:

$$y_i^* = \lambda_i \eta + \beta_i z_k + \varepsilon_i \quad (2)$$

where y_i^* is the latent response variable for ith item (when $y_i^* > \tau_i$, y_i is equal to 1, otherwise y_i is equal to 0; τ_i is the threshold parameter and is related to item difficulty for ith item), η is latent trait variable that is aimed to be measured by the test, λ_i is the factor loading, ε_i is random error, z_k is grouping variable that indicates the group membership and β_i is the slope that relates z_k with y_i^* (Finch, 2005; Wang et al., 2009).

MIMIC is a method that allows conducting DIF analyses with multiple grouping variables, and the z symbol in Figure 1 is defined as a vector of the aforementioned multiple grouping variables. The z vector may have continuous or categorical values. Thus, it can be said that MIMIC method is more flexible than traditional DIF detection methods (MH, SIBTEST, IRT-LRT, etc.) that use just only one categorical grouping variable (Wang et al., 2009).

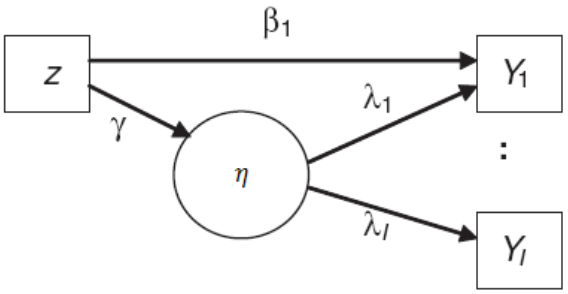


Figure 1. Detecting Differential Item Functioning in Item Y1 with the MIMIC Method. Adapted from "The MIMIC Method with Scale Purification for Detecting Differential Item Functioning" by W. C. Wang, C. L. Shih and C. C. Yang, 2009, *Educational and Psychological Measurement, 69*(5), p. 717. Copyright 2009 by SAGE Publications.

The underlying base method for DIF detection by MIMIC method involves evaluation of both direct and indirect effects for a grouping variable. By investigating the indirect effect of the grouping variable (z) on item responses through the latent trait (η), it is indicated whether the mean of this latent variable differs across the groups or not; thus, computations are carried out for group differences on the latent trait. By investigating the direct effect of the grouping variable (z) on item responses (Yi), i.e. $\beta_1 \neq 0$, it is indicated whether any difference in response probabilities exists across the groups or not. This relation, after checking the differences in the mean of latent trait for groups, is the test of uniform DIF (Finch, 2005).

DIF detection models to be used in bias studies must be appropriate for the test used and for the properties of the groups to which the test is applied. This study used different conditions for dichotomous data to investigate the circumstances under which the MIMIC method produces more accurate results in DIF detection. The conditions used in the current study differ from previous studies in terms of the levels of these three conditions: sample size, ability distribution across groups, and percentage of items with DIF. It is an important question whether the MIMIC method works similarly in cases with different sample sizes (Wang & Shih, 2010). Therefore, different sample sizes in the study were compared. The data used in the study were produced according to the three-parameter logistic model (3PLM), and the test length was taken as 30 items to show similarity with actual applications. In addition, the focus of this study was on the assessment of uniform DIF.

In this study, the MIMIC method was compared to the LR method, which is a relatively more traditional method. This study compared how Type I error rates and power of MIMIC and LR DIF detection methods changed according to sample size, ability distributions of the groups, and percentage of items with DIF. In summary, the goal of this study was to investigate the performances of MIMIC and LR methods under various conditions according to their type I error rates and power when detecting DIF items on dichotomous tests. The research questions were as the following:

1. How do Type I error rates and power of MIMIC and LR DIF detection methods differ according to sample size?
2. How do Type I error rates and power of MIMIC and LR DIF detection methods differ according to ability distributions of the groups?
3. How do Type I error rates and power of MIMIC and LR DIF detection methods differ according to percentage of items with DIF?

METHOD

Simulation Conditions and Data Generation

This study is a DIF detection research using MIMIC and logistic regression methods for dichotomous data based on various simulation conditions. In this simulation study, conditions different from those of previous studies in which the MIMIC model was used were investigated.

The conditions that were kept constant throughout the study

For all conditions, the ability parameters of the individuals in the reference group were generated based on the standard normal distribution, N(0, 1). Furthermore, 30 dichotomously scored (either 0 or 1) responses for each individual were produced. The change in the item difficulty parameters between the groups for the items with DIF was set to a constant value as 0.6 units against the focal group to form medium DIF. The ratio of the focal group to the reference group (1:1) is another condition that was kept constant.

The conditions that were varied throughout the study

One of the conditions that was varied in this study was the sample size. Two levels of large sample size were used: 2000 (R: 1000, F: 1000) and 4000 (R: 2000, F: 2000). Finch (2005) found in his study that MIMIC method produces type I error rates higher than .05 nominal alpha level for a shorter test (i.e., 20 items) responded by a sample of 1000 (R: 500, F: 500) individuals under 3PL model. Based on the findings of Finch (2005), for a test with 30 items under 3PL model considered in this study, larger sample sizes were taken into account. In addition to sample size, ability distribution of the focal group was also a condition that was varied. Two levels of ability distribution of focal group were used: N(0, 1) and N(-0.5, 1). For the first level of the ability distribution of focal group condition, the cases where the distribution of the reference group and the focal group is the same were considered. For the second level of the ability distribution of focal group condition, the cases where the distribution of the focal group is lower than the reference group were considered Another condition that was varied in this study was the percentages of items with DIF. Two levels were used for this condition: 10% (3 items) and 20% (6 items). Items with DIF were kept the same throughout the test. In 10% of items with DIF condition, DIF was formed for items 4, 15, and 27 and in 20% of items with DIF condition, it was formed for items 1, 4, 15, 18, 26, and 27. By crossing the levels of each condition, total of 8 simulation conditions were created.

For each simulation condition, the data were derived for dichotomously scored (0/1) items using a 3PLM via R 3.0.2 program (R Core Team, 2013). The derivation of the data was performed 100 times for each condition. The item parameters used in this study were selected randomly from the item parameters used in Finch's (2005) study. The selected parameters are shown in Table 1.

Data Analysis Procedures and Evaluation Criteria

In the DIF analyses of the data, Mplus 6.12 (Muthén & Muthén, 1998, 2010) program was used for the MIMIC method and SAS 9.1.3 (SAS Institute, 2007) program was used for the logistic regression method. The DIF analyses were conducted using a pairwise approach in which the groups are compared with each other (i.e., focal group compared with reference group) (Sari & Huggins, 2014).

In the study, the effects of sample size, ability distribution of focal group, and the percentage of items with DIF on Type I error rates and power were investigated. The level of significance (α level) was assumed to be .05 in detecting items with DIF. Type I error is defined as a misclassification of an item without DIF as an item with DIF. Under 10% of items with DIF condition, there were 27 non-DIF items whereas under 20% of items with DIF condition, there were 24 non-DIF items. The percentage of non-DIF items that were falsely detected as DIF items was calculated for Type I error rate. The concept of power, on the other hand, is correct classification of an item with DIF as an item with DIF. Under 10% of items with DIF condition, there were 3 DIF items whereas under 20% of items with DIF condition, there were 6 DIF items. The percentage of DIF items that were correctly detected as DIF items was calculated for power. Both Type I error and power are equally important for DIF researches (Vaughn & Wang, 2010). According to Cohen and Cohen (1983) when investigators need to set the power, it is reasonable for them to choose a value in the .70 - .90 range. In the current study, the desired value for power rate was considered as .70 and above.

Table 1. Item Parameter Values Used in Generation of Simulated Data

	Reference Group		
Item	a_i	b_i	c_i
1	1.10	-0.70	.20
2	0.70	-0.60	.20
3	1.40	0.10	.20
4	0.40	0.80	.20
5	1.40	-0.40	.20
6	1.60	-0.10	.16
7	1.20	0.50	.20
8	1.20	1.40	.11
9	1.80	1.40	.12
10	2.00	1.60	.16
11	1.00	1.60	.13
12	1.50	1.70	.09
13	0.70	-0.50	.20
14	1.20	-0.30	.20
15	0.90	0.20	.20
16	0.70	-0.40	.20
17	1.00	0.70	.15
18	1.60	1.10	.12
19	1.10	2.00	.06
20	1.10	2.40	.09
21	1.70	1.30	.17
22	0.90	1.00	.15
23	0.50	-0.60	.20
24	1.30	0.40	.18
25	1.30	1.40	.06
26	1.10	1.20	.05
27	0.90	0.80	.20
28	0.40	-0.40	.20
29	0.80	-0.70	.20
30	1.00	1.10	.13

RESULTS

Type I Error Rate

Type I error rates are calculated for each condition, namely sample size, ability distribution of focal group, and percentage of items with DIF and given in Table 2.

Table 2. Type I Error Rates According to Sample Size, Ability Distribution of Focal Group, and Percentage of Items with DIF

DIF %	Sample Size	Ability Distributions R/F	MIMIC	LR
10	2000	(0,1) / (0,1)	.121	.069
		(0,1) / (-0.5,1)	.120	.068
	4000	(0,1) / (0,1)	.065	.087
		(0,1) / (-0.5,1)	.090	.097
20	2000	(0,1) / (0,1)	.129	.122
		(0,1) / (-0.5,1)	.128	.129
	4000	(0,1) / (0,1)	.076	.244
		(0,1) / (-0.5,1)	.078	.189

Note. DIF % refers to the percentage of items with DIF; LR = Logistic Regression; MIMIC = Multiple Indicators, Multiple Causes Model.

The main finding of this study was that the sample size was an important factor in DIF analyses conducted with MIMIC and LR methods. As the sample size increased from 2000 to 4000, the type I error rates decreased for MIMIC method but increased for the LR method when other conditions of the study were equal. For the MIMIC method, while the lowest rate was calculated under the condition

where the sample size was 4000, percentage of items with DIF was 10%, and the ability distribution of both groups showed a standard normal distribution N(0, 1), the highest rate was calculated under the condition where the sample size was 2000, percentage of items with DIF was 20%, and the ability distribution of both groups showed a standard normal distribution N(0, 1). On the other hand for the LR method, while the lowest rate was calculated under the condition where the sample size was 2000, percentage of items with DIF was 10%, and ability distribution of the focal group was N(-0.5, 1), the highest rate was calculated under the condition where the sample size was 4000, percentage of items with DIF was 20%, and the ability distribution of both groups showed a standard normal distribution N(0, 1).

The second important finding was that the percentage of DIF items was an important factor that effected the type I error rates. As the percentage of DIF items increased from 10% to 20%, type I error rates were very similar in MIMIC method, however, increased in LR method when other conditions of the study were equal. According to the study results, in terms of type I error rates, the percentage of DIF items was more effective factor for the LR method.

The third finding was that the change in the ability distribution of focal group did not have an important effect on type I error rates for both methods.

Power

Table 3 presents the power values for the two DIF detection methods for all conditions included in the study. The acceptable power rate for this study was .70 and above. In general, both methods had power rates above acceptable levels for all conditions.

The power rate of the MIMIC method was quite high for conditions with a sample size of 4000 respondents. The power rate of the LR method, on the other hand, was quite high for conditions wherein the sample size was large and the ability distribution of both groups showed a standard normal distribution N(0, 1). The standard definition of power at a specified level of alpha is not meaningful in cases where Type I error rates are high (Finch, 2005). However, all power results were included in this study for comparison purposes. The power rates were shown in italics for cases where Type I error rate was higher than .10. Considering all conditions, both methods had power high enough and these results reached a higher value when sample size increased.

Table 3. Power Rates According to Sample Size, Ability Distributions, and Percentage of Items with DIF

DIF %	Sample Sizes	Ability Distributions R/F	MIMIC	LR
10	2000	(0,1) (0,1)	*.770*	.800
		(0,1) (-0.5,1)	*.750*	.700
	4000	(0,1) (0,1)	.933	.910
		(0,1) (-0.5,1)	.910	.817
20	2000	(0,1) (0,1)	*.852*	*.827*
		(0,1) (-0.5,1)	*.780*	*.772*
	4000	(0,1) (0,1)	.977	*.935*
		(0,1) (-0.5,1)	.943	*.872*

Note. DIF % refers to the percentage of items with DIF; LR = Logistic Regression; MIMIC = Multiple Indicators, Multiple Causes Model.

The condition in which the power was closest to perfect for the MIMIC method was the one in which the sample size was 4000 respondents, ability distributions of the reference and focal groups showed a standard normal distribution, and percentage of items with DIF was 20%. The power results of the MIMIC method were larger than those of the LR method, except for a single condition. This condition was the one in which the sample comprised 2000 respondents, ability distributions of the reference and focal groups showed a standard normal distribution, and percentage of items with DIF was 10%. The differentiation of the ability distributions for the focal group affected the power of the LR method

more than the power of the MIMIC method for almost all conditions. In addition, the change in the percentages of items with DIF did not substantially change the power of both methods.

DISCUSSION and CONCLUSION

In this study, the performances of MIMIC and LR methods were compared according to their type I error rate and power. It can be concluded in this study that the MIMIC method produced lower Type I error rates than the LR method in conditions where the sample size was larger (4000 respondents); the LR method produced lower Type I error rates than the MIMIC method in conditions where the percentage of items with DIF was lower (10%) with smaller sample size (2000 respondents). In general, the Type I error rates of the MIMIC method were observed to be lower than those of the LR method. However, for both methods, Type I error rates exceeded acceptable alpha level ($\alpha = .05$) in all conditions. Specifically, while the increase in the sample size substantially reduced the Type I error rate of the MIMIC method for all conditions, its effect on the type I error rate of the LR method changed according to the percentage of items with DIF. While the change in the sample size had a very small effect on the Type I error rate of the LR method for 10% DIF items conditions, it caused a substantial increase in the Type I error rate of this method for 20% DIF items conditions. In the study conducted by Finch and French (2007), Type I error rates of the LR and CFA methods in detecting items with nonuniform DIF were not substantially affected by the increase in the sample size. Based on this results, it can be concluded that similar results obtained from current study for the LR method with only the 10% DIF items conditions. As can be understood from this current research, in the conditions where the percentage of items with DIF is high the LR method is more sensitive to the sample size condition. But the MIMIC method is affected by the sample size in the same manner for all conditions. The difference based on CFA between current and Finch and French's (2007) study can be attributed to the type of DIF. In their study they focused on nonuniform DIF and emphasized the question of the usefulness of CFA method for identifying this type of DIF. MIMIC method is also based on CFA and it is capable of detecting uniform DIF as also stated by Woods (2009), and Woods et al. (2009).

On the other hand, in the current study the increase in the percentage of items with DIF did not affect the Type I error rate of the MIMIC method importantly but increased that of the LR method. It can be seen in Finch's (2005) results that for the MIMIC method, in the bigger test length condition the effect of percentage of items with DIF was reduced for both sample size conditions, 600 and 1000 respondents. In the current study for both sample size (2000 and 4000 examinees) the effect of percentage of items with DIF was already quite low but still the type one error rates were not small enough as they were desired. By combining the result of these two studies it can be concluded for the MIMIC method that, big sample sizes or relatively small sample sizes with bigger test lengths are needed to reduce the effect of percentage of items with DIF.

The other result obtained from this study is that, the difference in the ability distribution of the focal group did not substantially affect the Type I error rates of both methods. In conclusion, when these two methods were compared in terms of Type I error rates, the change in the sample sizes was more effective for the MIMIC method while the change in the percentages of items with DIF was more effective for the LR method.

When the results were examined in general, the power of both methods for all conditions was above the acceptable level (.70). For conditions where the sample size was higher, the power results of the MIMIC method were quite high. The power of the LR method, on the other hand, was quite high for conditions where the sample size was large and the ability distribution of both groups showed a standard normal distribution. The power results of the MIMIC method were higher than those of the LR method, except for a single condition. This condition was the one in which the sample comprised 2000 respondents, the ability distributions of the reference and focal groups showed a standard normal distribution, and the percentage of items with DIF was 10%.

The increase in the sample size increased the power for both methods. The fact that the ability distribution of the focal group differed from the ability distribution of the reference group decreased

the power of both methods. The amount of reduction that this change in the ability distribution caused was more for the LR method for almost every condition. The increase in the percentage of items with DIF increased the power of both methods to a small extent. As a result, considering the change in the power, the sample size was the most effective variable for both methods.

Specifically, the change in the sample size was very effective in changing the power of the MIMIC method. The power of the MIMIC method increased as the sample size increased. Finch (2005) concluded in his study that the power results of the MIMIC method for 2PLM were generally as high as the power results of the classical methods or even in some conditions higher than those of the SIBTEST and MH methods. Similar results were obtained in this study for 3PLM, the power results of the MIMIC method were higher than those of the LR method for almost all conditions.

In the study conducted by Finch and French (2007), the power results of the LR and CFA methods in detecting items with nonuniform DIF were below .70 for all conditions. In current study, the power results were over .70 for both methods for all conditions. Finch and French (2007) reported in their study that the power of the LR method increased as the sample size increased. But, according to their results the power of the CFA method decreased or stayed the same while the sample size increased. In current study, as the sample size increased, the power of both LR and MIMIC methods increased. These two studies support each other in terms of the increase in power of the LR method according to the sample size condition. However, the results differed in terms of the change in the power of the MIMIC method, which is a method based on CFA. As mentioned before this difference between two studies can be attributed to the difference of the type of DIF (uniform or nonuniform) used in these studies.

In this study, three main conditions and eight sub-conditions were considered, with two different sample sizes, two different ability distributions for the focal group, and two different percentages of items with DIF. The number of items in the test was kept constant for all conditions. In future studies, the number of items in the test can be increased to see how the results are affected in long tests. As seen in the comparison of recent and previous research, test length may have an important effect on MIMIC method.

It is an important issue how the MIMIC method performs in terms of DIF at different sample sizes. Two different sample sizes, 2000 and 4000 individuals, were used in the study. However, the desired Type I error rates could not be achieved even with a sample size of 4000 individuals. This points out an important issue. And hence, future studies can be conducted on larger sample sizes to investigate the ideal sample size for the MIMIC method.

In the study, the ratio between the reference and focal group sizes was taken as 1:1. However, during the actual examinations, there can be different situations regarding the proportions of sample size of these two groups. Therefore, studies can be done using different ratios. Furthermore, the study was conducted with 3PL model-based data. Similar work can be conducted with 2PL model-based data, and comparisons can be made between these studies.

It is thought that this study will be a reference to the studies on DIF detection through the MIMIC method and that it will make it easy for researchers to decide the appropriate DIF detection method according to sample size and ability distributions in the analysis of the actual test results.

The aim of this study is to provide a reliable source to researchers in selecting DIF detection techniques that are appropriate for the test to be used and the properties of the test group. Thus, with the help of more reliable DIF detection techniques, tests can be made fairer.

Based on the results obtained from this research, it can be suggested to choose the LR method in DIF analysis studies performed on small samples such as the one comprising 2000 respondents and with small amount of DIF items such as 10% of test items; and the MIMIC method in DIF analysis studies performed on samples as large as approximately 4000 respondents and higher. Subsequent to the detection of items with DIF using these methods, it is advisable to refer to expert's opinion to conduct a study to detect bias in these items.

REFERENCES

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*(4), 355-368.

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, 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. doi: 10.1177/0013164406296975

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *Journal of Gerontology: Social Sciences, 57B*(5), 275-284.

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

Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, 32*(2), 131-144.

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

Muthén, L. K. & Muthén, B. O. (1998, 2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 5*(2), 107-124. doi: 10.1080/10705519809540095

R Core Team (2013). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Sari, H. I. & Huggins, A. C. (2014). Differential item functioning detection across two methods of defining group comparisons: Pairwise and composite group comparisons. *Educational and Psychological Measurement, 75*(4), 648-676. doi: 10.1177/0013164414549764

SAS Institute Inc. (2007). *SAS® 9.1.3 qualification tools user's guide*. Cary, NC: SAS Institute Inc.

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.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Vaughn, B. K., & Wang, Q. (2010). DIF trees: Using classification trees to detect differential item functioning. *Educational and Psychological Measurement, 70*(6), 941-952. doi: 10.1177/0013164410379326

Wang, W. C., & Shih, C. L. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement, 34*(3), 166-180. doi: 10.1177/0146621609355279

Wang, W. C., Shih, C. L., & Yang, C. C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement, 69*(5), 713-731. doi: 10.1177/0013164409332228

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research,44*(1), 1-27. doi: 10.1080/00273170802620121

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2009). Illustration of MIMIC-model DIF testing with the schedule for nonadaptive and adaptive personality. *J Psychopathol Behav Assess, 31*, 320-330. doi: 10.1007/s10862-008-9118-9

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.* Ottowa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

MIMIC ve Lojistik Regresyon Yöntemlerinin DMF Belirleme Performansları

Giriş

DMF (Değişen Madde Fonksiyonu), eşit yetenek düzeyinde ancak farklı gruplarda yer alan bireylerin belirli bir maddeye verdikleri cevapların doğru olma olasılığının birbirinden farklı olması durumunda ortaya çıkar (Crane, Belle & Larson, 2004; Mazor, Kanjee & Clauser, 1995). DMF'li maddeleri tespit etmek üzere çok sayıda DMF belirleme yöntemi geliştirilmiştir. Bu çok sayıdaki yöntem arasından MIMIC (Multiple Indicators, Multiple Causes) yöntem oldukça yenidir ve özellikle iki kategorili puanlanan test maddelerinde MIMIC yöntemin kullanıldığı araştırma sayısının eksikliği göze çarpmaktadır (Finch, 2005). Bu nedenle, MIMIC yöntemin DMF belirlemedeki performansının araştırılması gerekli görülmektedir.

Hem sürekli hem de kategorik birden çok sayıda gruplama değişkeni ile kullanılabilen MIMIC yöntemin, sadece tek bir kategorik değişkenle analiz yapmaya izin veren yöntemlere kıyasla daha esnek olduğunu ifade etmek mümkündür (Wang, Shih & Yang, 2009). IRT (Item Response Theory) kapsamında ele alınan DMF testlerinde ikiden fazla grup söz konusu olduğunda yöntemlerin oldukça karmaşıklaştığı görülmekte iken MIMIC yöntemin aynı anda çok sayıda değişkeni analize ekleyebilme avantajı söz konusudur (Fleishman, Spector & Altman, 2002). Birden fazla grubun eşzamanlı olarak tek bir aşamada analizine olanak sağladığı için MIMIC yöntemi oldukça kullanışlı bulunmaktadır (Muthen, 1988). DMF araştırmalarındaki avantajları göz önüne alındığında bu yöntem oldukça ilgi çekici bir araştırma konusu haline gelmektedir.

Yanlılık araştırmalarında kullanılan DMF belirleme yöntemleri kullanılan teste ve testin uygulandığı grubun özelliklerine uygun olmalıdır. Bu amaçla, bu araştırmada MIMIC yöntemin hangi koşullar altında daha doğru sonuçlar verdiği ortaya çıkarılmak istenmiş ve araştırma iki kategorili verilerle çeşitli koşullar kullanılarak yürütülmüştür. Çalışmada etkisi incelenen koşullar örneklem büyüklüğü, DMF'li madde yüzdesi ve gruplar arası yetenek dağılımlarıdır. Ayrıca, bu araştırmada tek biçimli (uniform) DMF'nin belirlenmesi üzerine odaklanılmıştır. Özetle bu araştırmada MIMIC ve LR (Logistic Regression) yöntemleri farklı örneklem büyüklüğü, grupların yetenek dağılımı farklılıkları ve DMF'li madde yüzdesinin değiştiği koşullarda Tip 1 hata ve güçlerine dayalı olarak karşılaştırılmıştır. Buna bağlı olarak araştırmanın problem cümlesine aşağıda yer verilmiştir:

MIMIC ve LR DMF belirleme yöntemlerinin Tip 1 hata ve güçleri örneklem büyüklüğü, grupların yetenek dağılımları ve DMF'li madde yüzdesine göre nasıl değişmektedir?

Yöntem

Bu çalışma iki kategorili puanlanan veriler için yürütülmüş, simülasyona dayalı bir DMF belirleme çalışmasıdır. Çalışmada kullanılan DMF belirleme yöntemleri MIMIC ve LR'dir. Çalışmanın verilerini üretmek üzere R 3.0.2, DMF belirleme analizleri içinse MPlus 6.12 ve SAS 9.3.1 programlarından yararlanılmıştır. Analizler her bir koşula ait veri setleri üzerinde 100 kez tekrarlanmıştır. Ayrıca araştırmanın verileri 3 parametreli lojistik modele (3PLM) uygun olacak şekilde üretilmiştir.

Çalışmada sabit tutulan koşullar şu şekildedir: referans grupta yer alan bireylerin yetenek parametrelerine ait dağılım [N(0,1)], test uzunluğu (30 madde), DMF'li maddeler için gruplara ait güçlük parametreleri farkı (0.6 birim), odak gruptaki bireylerin sayısının referans gruptakilere oranı (1:1). Çalışmanın değişen koşulları ise şu şekildedir: örneklem büyüklüğü (2000, 4000), odak grupta yer alan bireylere ait yetenek dağılımları [N(0,1), N(-0.5, 1)] ve DMF'li madde yüzdesi (%10, %20).

Sonuç ve Tartışma

Özetle bu araştırmada örneklem büyüklüğü, yetenek dağılımı ve DMF'li madde yüzdesinin MIMIC ve LR yöntemlerine ait Tip 1 hata ve güç üzerindeki etkileri incelenmiştir. Genel olarak bakıldığında MIMIC yöntemine ait Tip 1 hatanın LR yöntemininkilere göre daha düşük olduğu göze çarpmıştır. Ancak her iki yöntem için de tüm koşullarda Tip 1 hatalarının kabul edilebilir alfa düzeyinden (α = .05) yüksek çıktığı görülmüştür. Koşullar detaylı olarak incelenecek olursa, örneklem büyüklüğündeki artış tüm koşullar için MIMIC yöntemin Tip 1 hatasını önemli ölçüde düşürmüştür. Ancak LR yöntemin Tip 1 hatasındaki değişim DMF'li madde yüzdesine bağlı olarak değişmiştir. %10 DMF içeren koşullarda Tip 1 hata önemli ölçüde değişiklik göstermezken %20 DMF'li madde koşulunda hata önemli ölçüde artmıştır. Demek oluyor ki LR yöntemi DMF'li madde yüzdesi arttıkça örneklem büyüklüğüne duyarlı hale gelmiştir. Daha önce benzer şekilde LR ve DFA (Doğrulayıcı Faktör Analizi) yöntemleri ile yürütülen Finch ve French'in (2007) çalışma bulguları ise neredeyse her iki yöntem için de bu araştırmanın sonuçlarından farklılık göstermektedir ve bu farklılık MIMIC yöntem için daha belirgin çıkmıştır. Finch ve French'in (2007) bulguları LR ve DFA yöntemlerinin Tip 1 hatalarının örneklem büyüklüğünden önemli derecede etkilenmediklerini işaret etmiştir. MIMIC yöntemi DFA'ya dayalı bir yöntemdir. Bu iki çalışmanın sonuçları arasındaki farklılığın sebebinin bu açıdan düşünüldüğünde DMF türü olabileceği söylenebilir. Çünkü DFA yönteminin tek biçimli olmayan DMF'yi belirlemedeki kullanışlılığından şüphe duyulduğu Finch ve French'in (2007) araştırma sonuçları arasındadır. Ayrıca DFA'ya dayanan MIMIC yönteminin de tek biçimli DMF'yi belirleyebildiği, tek biçimli olmayan DMF'yi belirlemede yetersiz olduğu Woods (2009), Woods, Oltmanns ve Turkheimer'in (2009) araştırmalarında açıkça belirtilmiştir.

Çalışmanın bir başka sonucuna göre, hem 2000 hem de 4000 kişilik örneklem büyüklüklerinde DMF'li madde yüzdesindeki artışın MIMIC yöntemin Tip 1 hatasına etki etmediği ancak LR yöntemininkini arttırdığı görülmüştür. Finch'in (2005) yürüttüğü araştırmada 600 ve 1000 örneklem büyüklüklerinde test uzunluğunun artması ile DMF'li madde yüzdesinin MIMIC yöntem üzerindeki etkisinin azaldığı görülmüştür. Bu iki araştırmanın sonuçları birlikte düşünüldüğünde DMF'li madde yüzdesinin MIMIC yöntem üzerindeki etkisini azaltmak için 2000 ve 4000 gibi daha büyük örneklem büyüklüklerine ya da 600 veya 1000 gibi nispeten daha küçük örneklem büyüklükleri ile birlikte daha büyük test uzunluklarına ihtiyaç duyulmaktadır.

Araştırmanın bir başka sonucu ise odak grubun yetenek dağılımındaki farklılığın her iki yöntemin de Tip 1 hatalarını etkilemediği yönündedir. Özetle, iki yöntem Tip 1 hataları bakımından karşılaştırıldığında MIMIC yöntem için örneklem büyüklüğündeki değişim daha etkili iken, LR yöntem için DMF'li madde yüzdesindeki değişim daha etkili olmuştur.

Araştırma sonuçları yöntemlerin güçleri bakımından incelendiğinde, her iki yöntemin güç değerlerinin tüm koşullar için kabul edilebilir değerin (.70) üzerinde olduğu gözlemlenmiştir. Araştırma sonuçlarına göre her iki yöntem için de güç değerleri açısından, örneklem büyüklüğü en etkili değişken olmuştur. Ayrıca sonuçlar neredeyse tüm koşullarda MIMIC yöntemin güç değerlerinin LR yöntemininkilerden daha yüksek olduğunu işaret etmiştir. Benzer bir sonuca Finch'in (2005) araştırmasında rastlanmıştır. Bu araştırmada da MIMIC yöntemin güç değerlerinin klasik yöntemlerinki kadar yüksek olduğu vurgulanmış ve hatta bazı koşullarda SIBTEST ve MH yöntemlerine göre daha yüksek güç değerlerine sahip olduğu belirtilmiştir.

Bu araştırmada her iki yönteme ait güç değerlerinin tüm koşullar için .70 ve üzeri değerler verdiği tespit edilmiştir. Finch ve French'in (2007) araştırma sonuçlarına göre ise LR ve DFA yöntemlerinin güç değerlerinin neredeyse tüm koşullarda .70 değerinin altında olduğu görülmüştür. Ayrıca, örneklem büyüklüğü arttıkça LR yönteminin güç değerinin arttığı ancak, DFA yönteminin güç değerinin azaldığı ya da aynı kaldığı belirtilmiştir. Bu araştırmanın sonuçlarına göre ise örneklem büyüklüğü arttıkça LR ve MIMIC yöntemlerin güç değerlerinin arttığı gözlenmiştir. Bu bakımdan iki çalışma LR yöntemi sonuçlarına dayalı olarak birbirini destekler nitelikte iken MIMIC ve DFA yöntemleri sonuçları bakımından birbirini desteklememektedir. Daha önce de belirtildiği üzere MIMIC yöntem DFA'ya

dayalı bir yöntemdir ve bu iki araştırma sonucundaki farklılığın sebebinin DMF türüne (tek biçimli ve tek biçimli olmayan) dayandığı söylenebilir.

Bu araştırmada test uzunluğu sabit tutulmuştur. Ancak test uzunluğunun MIMIC yöntem üzerindeki etkisinin daha net ortaya konabilmesi için ileriki araştırmalarda araştırmacılara daha büyük test uzunluklarını kullanarak araştırmalar yürütmeleri önerilebilir. Ayrıca, MIMIC yöntemin farklı örneklem büyüklüklerinde nasıl sonuçlar verdiği önemli bir araştırma sorusudur. Bu araştırmada 2000 ve 4000 olmak üzere iki farklı örneklem büyüklüğü ele alınmıştır. Ancak, 4000 kişilik örneklem büyüklüğünde dahi istenen Tip 1 hata oranına ulaşılamamıştır. Bu nokta önemli bir soruna işaret etmektedir. İleriki araştırmalarda daha yüksek örneklem büyüklükleri kullanılarak MIMIC yöntemin yaklaşık hangi örneklem büyüklüğünde ideal sonuçlar verdiği tartışılmalıdır.

Bu araştırma ile, MIMIC yöntemin kullanılarak DMF'li maddelerin belirlenmeye çalışıldığı araştırmalara bir referans olması amaçlanmıştır. Böylece, kullanılan teste ve testi alan grubun özelliklerine uygun DMF belirleme yöntemlerinin seçiminde araştırmacılara güvenilir bir kaynak sağlanması umulmaktadır. Bununla birlikte, gerçek test sonuçlarının analizinde örneklem büyüklüğü ve yetenek dağılımlarına bağlı olarak uygun DMF belirleme yönteminin seçilmesinde araştırmacılara yardımcı olmak istenmiştir. Daha güvenilir yöntemlerin yardımıyla testler daha adil hale getirilebilir.

Bu araştırmadan elde edilen sonuçlara dayanılarak 2000 gibi küçük örneklem büyüklükleri ve %10 gibi küçük oranda DMF'li madde içeren çalışmalarda LR yönteminin, yaklaşık 4000 ya da daha yüksek örneklem büyüklükleri ile yürütülen çalışmalarda ise MIMIC yöntemin tercih edilmesi önerilebilir. DMF'li maddelerin belirlenmesinin ardından, bu maddelere yönelik yanlılık çalışması yapmak üzere uzman kanısına başvurulması da önerilmektedir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(1); 13-26

Investigating Preservice Middle School Mathematics Teachers' Competencies in Statistics and Probability in Terms of Various Variables *

Okan KUZU ** Muhammet ARICAN ***

Abstract
In this study, probabilities of preservice middle school mathematics teachers' possession of four fundamental cognitive skills required for learning and teaching statistics and probability topics were examined by using the log-linear cognitive diagnostic model, which is one of the cognitive diagnostic models. Moreover, the probabilities of preservice teachers' possession of these skills were investigated according to gender, university ranking, and grade level variables. Hence, it was examined whether there was a significant relationship between the probabilities of having each skill and these variables. A Statistical Reasoning Test, which was developed by Arican and Kuzu in 2019, measured preservice teachers' possession of four critical skills was used in collecting the data. These four skills included representing and interpreting data, drawing inferences about populations based on samples, selecting and using appropriate statistical methods to analyze data, and understanding and applying basic concepts of probability. In the 2016-2017 academic year, the test was applied to 456 preservice teachers selected from four different universities in Turkey, and probabilities of their possession of each attribute were calculated. Later, the relationship between the preservice teachers' test scores and gender was examined by using the Mann-Whitney U test, and the relationship between their test scores and ranking of the attended university and grade level were examined using the Kruskal Wallis-H test. Although probabilities of the preservice teachers' possession of these four skills did not significantly differ according to gender, some significant differences were detected for university ranking and grade level variables.

Key Words: Cognitive diagnostic models, gender, grade level, preservice middle school mathematics teachers, statistics and probability, university ranking.

INTRODUCTION

Statistics, which is defined as a branch of science, consists of techniques and methods related to data collection, analysis, and interpretation of results (Saraçbaşı & Kutsal, 1987). Statistics, which is based on the principles such as determining the relationship between variables, making generalizations according to the results obtained from samples, and making predictions for the future, have become the focus of interest in many countries and have taken place in mathematics education programs of many countries (Ardıç, Yılmaz & Demir, 2012; Makar & Rubin, 2009; Shaughnessy, 2007; Watson, 2006). When the constantly developing and renewing mathematics curricula are examined, statistical competencies such as reading data, representing data, using central tendency and spread measures, making predictions and inferences from data, and calculating probability are given more attention in different class levels than previous years (Ministry of Education-MEB, 2013, 2018).

Statistics is based on calculations of probability and enables mathematical treatment of random events and making inferences from data. Statistics and probability, which interact with real-life problems and

* A part of this study were presented as an oral presentation at the X. International Congress of Educational Research held in Nevşehir between April 27-30, 2018. Moreover, this study was supported by the Kirsehir Ahi Evran University Scientific Research Projects Coordination Unit. Project Number EGT.A3.16.014.
** Assist. Prof., Kirsehir Ahi Evran Üniversity, Faculty of Education, Kirsehir-Turkey, okan.kuzu@ahievran.edu.tr, ORCID ID: 0000-0003-2466-4701
*** Assist. Prof., Kirsehir Ahi Evran Üniversity, Faculty of Education, Kirsehir-Turkey, muhammetarican@gmail.com, ORCID ID: 0000-0002-0496-9148

To cite this article:
Kuzu, O., & Arıcan, M. (2020). Investigating preservice middle school mathematics teachers' competencies in statistics and probability in terms of various variables. *Journal of Measurement and Evaluation in Education and Psychology, 11*(1), 13-26. doi: 10.21031/epod.562586

Received: 09.05.2019
Accepted: 01.09.2019

other disciplines (e.g., economics, physical education, etc.), have been the focus of mathematics education from the past to the present day and have been included in the learning standards of the leading international educational institutions (e.g., National Council of Teachers of Mathematics-NCTM; National Assessment of Educational Progress-NAEP) (Batanero & Díaz, 2010; Franklin et al., 2007; Jones, 2005). Although the topics of statistics and probability have such importance, teachers and students face various difficulties in teaching and learning these topics (Batanero & Díaz, 2012). For example, Gürbüz, Toprak, Yapıcı, and Doğan (2011) found that teachers stated probability as one of the most difficult subjects in Turkish secondary school mathematics curriculum. Moreover, Boyacıoğlu, Erduran, and Alkan (1996) found that while 91% of the students stated probability as one of the most difficult subjects to understand, 84% of the teachers stated probability as one of the most difficult subjects to teach. In addition to these findings, students' difficulties with statistics and probability are also reported in international studies. When the eighth grade mathematics results of the Trends in International Mathematics and Science Study (TIMSS) were investigated, which included numbers, algebra, geometry, and data and chance domains, 22 out of 39 participating countries, including Turkey, obtained average scores in the data and chance domain that were lower than the TIMSS median-score of 500 (Mullis, Martin, Foy & Hooper, 2016). Furthermore, in the data and chance domain Turkey was ranked 12th among 13 European countries with an average score of 466 points. Although among the four domains, Turkish students obtained the highest average score from the data and chance domain, it was the only domain in which the average score of students decreased when compared with the TIMSS 2011 results (Mullis, Martin, Foy & Arora, 2012).

In addition to studies that have been conducted for identifying the difficulties encountered in teaching and learning of statistics and probability, there are also studies aimed at comparing the mathematics achievement of male and female students. The relationship between academic achievement and gender is an issue that has been discussed for many years (Eitle, 2005). When the studies on students' mathematics performances are examined, although there are many studies indicating that boys are more successful than girls (e.g., Felson & Trudeau, 1991; Fryer & Levitt, 2010; Stoet & Geary, 2013), there are also studies emphasizing girls are more successful than boys (e.g., Chambers & Schreiber, 2004; Farooq, Chaudhry, Shafiq & Berhanu, 2011). On the other hand, it is possible to find studies indicating that there is no difference between mathematics achievement of girls and boys (e.g., Chiesi & Primi, 2015; Duckworth & Seligman, 2006; Else-Quest, Hyde & Linn, 2010; Lindberg, Hyde, Petersen & Linn, 2010). When the effect of gender on mathematics performance is analyzed in terms of statistics, boys and girls do not differ in terms of their mathematics ability; however, in comparison to male students, female students have more negative attitudes towards statistics and have less confidence in their abilities (Chiesi & Primi, 2015). Bulut, Yetkin, and Kazak (2002) examined preservice mathematics teachers' (PSTs) achievements on probability and found that male students were more successful in probability than female students. Furthermore, in the same study, Bulut et al. (2002) also examined PSTs' attitudes towards the mathematics course and probability subject and found that girls reflected more positive attitudes towards the mathematics course, but there was no significant difference between the two groups in terms of their attitudes towards probability subject. When the TIMSS 2015 eighth grade mathematics results were examined, female students were more successful in mathematics than male students in seven countries; male students were more successful in six countries, and no significant difference was found between male and female students in 26 countries. In terms of data and chance domain, the mean scores of female and male students were very close to each other (Female: 475; Male: 472). When the data and chance mean scores of Turkish students were analyzed according to their gender, female students obtained slightly better mean score than male students (Female: 470; Male: 464).

When the studies on statistics and probability were examined, it was recognized that these two subjects were among the least investigated subjects in mathematics. On the other hand, the studies conducted on these two subjects generally aimed to understand students' performance, strengths, and weaknesses (Ulutaş & Ubuz, 2008). Some studies (e.g., Batanero & Díaz, 2012; Batanero, Godino & Roa, 2004; Franklin & Mewborn, 2006) emphasized that the difficulties faced by teachers and PSTs on statistics and probability were originated from the inadequately developed statistics and probability curriculum in universities. In addition, teachers who had little opportunity to obtain accurate information about

the principles and concepts of underlying practices of data analysis had difficulty in forming statistical knowledge (Franklin et al., 2007). Overall, relying on the Classical Test Theories (CTT), the studies conducted on statistics and probability (e.g., Olpak, Baltaci & Arican, 2018; Tsakiridou & Vavyla, 2015; Zhang & Maas, 2019) more often used total score-based evaluation systems. In these studies, the students' performances were evaluated in terms of the average scores that they obtained. Assessment approaches that use a single score (e.g., average score) have been criticized for not providing very detailed information on students' performances (Leighton & Gierl, 2007; Nichols, Chipman & Brennan, 2012), and alternatively, cognitive diagnostic models (CDMs) have been developed for obtaining more detailed assessments (Rupp, Templin & Henson, 2010). In CDMs, rather than calculating the total scores, the probability of each student's possession of the desired skill is determined, and diagnostic feedback is provided on their strengths and weaknesses. For instance, in a single score-based assessment system, a student with a score of 59 can be assessed as unsuccessful in a test with an average score of 60, whereas in CDMs, assessments are provided in terms of students' possession of the required skills rather than their scores. Thus, CDMs offer a more effective assessment of students' performances than CTTs.

Cognitive Diagnostic Models

Cognitive diagnostic models, also known as diagnostic classification models (DCMs), are a family of psychometric models that provide diagnostic assessments of participants' expertise on skills, which are referred as attributes, that the test aims to measure by calculating the likelihood that they have these skills based on their responses to the test items. CDMs provide participants with cognitive feedback about the skills to be measured and offer more detailed information about their cognitive strengths and weaknesses. One of the strengths of CDMs is that they provide more reliable estimates than CTTs, even if a small number of test items are used (Templin & Bradshaw, 2013). In recent years, researchers have used CDMs to provide diagnostic assessment on the results that students (e.g., Choi, Lee & Park, 2015; Dogan & Tatsuoka, 2008; Im & Park, 2010; Lee, Park & Taylan, 2011; Sen & Arican, 2015), teachers (e.g., Bradshaw, Izsak, Templin & Jacobson, 2014), and PSTs (e.g., Arican & Kuzu, 2019) obtained from several subjects of mathematics.

CDMs classified into three categories: compensatory models, non-compensatory models, and general models (Ravand & Robitzsch, 2015). Deterministic input, noisy-or-gate model (DINO) (Templin & Henson, 2006), and compensatory reparameterized unified model (C-RUM) (Hartz, 2002) are the examples of compensatory models. Deterministic input, noisy-and-gate model (DINA) (Junker & Sijtsma, 2001) and non-compensatory reparameterized unified model (NC-RUM) (DiBello, Stout & Roussos, 1995; Hartz, 2002) can be given as the examples of non-compensatory models. Finally, the general diagnostic model (GDM) (von Davier, 2005), the log-linear cognitive diagnostic model (LCDM) (Henson, Templin & Willse, 2009), and generalized deterministic input, noisy-and-gate model (G-DINA) (de la Torre, 2011) are the examples of general models that allow both compensatory and non-compensatory relationships.

This study was conducted using LCDM, which is one of the general models. LCDM places participants' responses to items in latent classes and thus helps researchers to determine their attributes (Bradshaw et al., 2014). Depending on the size and direction of the item parameters, LCDM can model attribute effects on each item response in a compensatory or non-compensatory manner, which gives researchers greater flexibility (Bradshaw et al., 2014). Therefore, LCDM was used to analyze the present data because of this flexibility.

The Purpose of the Study

In order to overcome the problems encountered in the learning and teaching of statistics and probability, it has been given importance recently to develop students' statistical skills in the Turkish education system and to equip students with these necessary skills (MEB, 2013, 2018). Moreover, as mentioned above, students' inadequacy in statistics and probability subjects raised questions about

how well preservice mathematics teachers graduated from higher education programs were educated in these subjects. The fact that students, teachers, and PSTs encounter some difficulties in statistics and probability suggests that they may have deficiencies in terms of the skills required in teaching and learning of these subjects. Therefore, providing diagnostic feedback on these deficiencies will contribute to educators to address the difficulties encountered.

Using the four fundamental cognitive skills required for preservice middle school mathematics teachers in statistics and probability, this study examined whether the PSTs' possesions of these skills differ according to their gender, ranking of the attended university, and grade level. Therefore, the following research questions were investigated in this study:

1. Do preservice middle school mathematics teachers' possession of skills differ according to their gender?
2. Do preservice middle school mathematics teachers' possession of skills differ according to the base scores of the universities they study?
3. Do preservice middle school mathematics teachers' possession of skills differ according to their grade level?

METHOD

In this quantitative study, the descriptive survey model was used to determine whether the PSTs' possession of attributes differ according to their gender, ranking of the attended university, and grade levels. The descriptive survey model is a research method that aims to describe a situation, views, interests, and competencies, which happened in the past or still exists, as it is (Karasar, 2005).

Sample

The sample of the study was composed of 456 PSTs (315 females, 108 males; 33 unspecified) studying in four different universities. In 2016, 67 universities had middle school mathematics teacher programs. These universities were ranked from the highest to the lowest by taking into account the average of the university entrance scores of the relevant program in the last five years. Four universities were randomly selected by using a stratified sampling method, which is one of the probability-based sampling techniques. Using the interquartile range, which is a descriptive statistical measure, 1 high from the first 17 universities (66 PSTs), 2 medium between 18 and 50 (224 PSTs), and 1 low from the last 17 universities (166 PSTs) were selected. The universities were located in three different regions of Turkey (1 Western Anatolia, 2 Central Anatolia, and 1 Eastern Anatolia), and the descriptive information on the PSTs was presented in Table 1.

Table 1. The Distribution of the Sample

		Grade				Total
		1	2	3	4	
Gender	Female	75	108	110	22	315
	Male	27	30	32	19	108
	Unspecified	4	12	16	1	33
Total		106	150	158	42	456

Data Collection Instruments

The Statistical Reasoning Test developed by Arican and Kuzu (2019) was used in this study. The test measured four attributes: A1: Representing and interpreting data; A2: Drawing inferences about populations based on samples; A3: Selecting and using appropriate statistical methods to analyze data; and A4: Understanding and applying basic concepts of probability. While determining these four attributes, national (MEB secondary school mathematics curriculum) and international (NCTM and

Common Core State Standards-CCSS) standards were examined. The test consisted of 20 items (15 multiple-choice and five open-ended), and when preparing these items, questions included in the national and international (TIMSS and The Programme for International Student Assessment-PISA) large-scale tests were taken into account. In order to determine which attribute or attributes each item measures, three academicians specialized in mathematics education and two mathematics teachers independently coded the test items in terms of the attributes they measure (*1: if the items measure the attributes*; *0: if the items do not measure the intended attributes*). If at least three experts agreed that the item measures an intended attribute, then it was included in the Q-matrix. The Q-matrix was presented in Table 2.

Table 2. The Q-Matrix

Attribute/Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
A1	1	0	0	1	0	0	1	0	0	1	1	1	1	0	0	1	1	0	0	0	9
A2	1	0	0	0	0	1	1	0	0	1	1	1	1	1	0	0	1	0	0	1	10
A3	0	0	0	0	0	1	0	0	0	1	1	1	0	1	0	1	0	0	1	1	8
A4	1	1	1	1	1	0	1	1	1	0	0	0	0	0	1	0	0	1	0	0	10

Table received from "Diagnosing Preservice Teachers' Understanding of Statistics and Probability: Developing a Test for Cognitive Assessment" by M. Arican and O. Kuzu, 2019, *International Journal of Science and Mathematics Education*, pp. 1-20. All rights reserved to Springer Nature.

In CDMs, the degree to which an item distinguishes between masters and nonmasters of an attribute is calculated by the item-attribute indices. Although there is no critical cut-off score stated for the removal of test items, de la Torre (2008) reported .31 as low. Accordingly, as seen in Table 3, item-attribute indices were low only in Items 6, 15, and 18.

Table 3. Item-Attribute Discrimination Indices

Items	A1	A2	A3	A4
Item 1	.55	.63		.39
Item 2				.69
Item 3				.78
Item 4	.61			.56
Item 5				.86
Item 6		.27	.23	
Item 7	.58	.45		.73
Item 8				.65
Item 9				.73
Item 10	.52	.45	.41	
Item 11	.45	.43	.35	
Item 12	.53	.55	.38	
Item 13	.41	.38		
Item 14		.75	.59	
Item 15				.21
Item 16	.51		.54	
Item 17	.44	.42		
Item 18				.22
Item 19			.68	
Item 20		.50	.63	

Table received from "Diagnosing Preservice Teachers' Understanding of Statistics and Probability: Developing a Test for Cognitive Assessment" by M. Arican and O. Kuzu, 2019, *International Journal of Science and Mathematics Education*, pp. 1-20. All rights reserved to Springer Nature.

The item difficulty index ranges from 0 to 1 and represents the proportion of students who correctly answered an item. In this study, the item difficulty index ranged between .13 and .86 and had an average of .49 (Table 4). The average item difficulty index of a test is recommended to be around .50

(Çepni, et al., 2008). Therefore, there was a good balance among the items in terms of their difficulty indices.

Table 4. Item Difficulty Indices

Items	İndex	Items	İndex
Item 1	.50	Item 11	.75
Item 2	.43	Item 12	.68
Item 3	.46	Item 13	.82
Item 4	.57	Item 14	.52
Item 5	.44	Item 15	.23
Item 6	.86	Item 16	.56
Item 7	.33	Item 17	.53
Item 8	.67	Item 18	.13
Item 9	.31	Item 19	.29
Item 10	.24	Item 20	.42

Data Analysis

Arican and Kuzu (2019) examined cognitive skills that PSTs required to have for teaching statistics and probability topics and determined four fundamental skills (i.e., attributes), and the results are presented in Table 5. When Table 5 is examined, the probability of the PSTs' possession of Attribute 1 was .647, and this value was higher than the probability of having the remaining three attributes. Although the lowest probability was obtained for Attribute 2, in general, the PSTs were less likely to have Attributes 2, 3, and 4. Using the reliability criterion developed by Templin and Bradshaw (2013), Arican and Kuzu (2019) stated that the test measures each attribute with .89, .82, .83, and .90 reliability, respectively. Moreover, with the help of the Mplus program, Arican and Kuzu (2019) eliminated classification problems by removing non-meaningful one-way and two-way interaction effects that did not contribute to the calculation of the PSTs' probabilities for having attributes. In addition, calculating the bivariate model fit information, item pairs indicating misfit were determined which consisted of only 7% of the total item pairs. Therefore, the test items and Q-matrix used were found to be appropriate for calculating the probabilities of desired attributes.

Table 5. The Probabilities of the PSTs' Possessions of Attributes

Attributes	Probability	Sd
A1 Representing and interpreting data	.647	.396
A2 Drawing inferences about populations based on samples	.286	.347
A3 Selecting and using appropriate statistical methods to analyze data	.476	.396
A4 Understanding and applying basic concepts of probability	.427	.410

This study examined whether the probabilities of the PSTs' possession of four attributes (see Table 5) differed according to gender, ranking of the attended university, and grade levels. For this purpose, the PSTs' answers to the test items were coded as 0 (wrong answer), 1 (correct answer), and 9 (incomplete answer). Then, the coded answers were transferred into the Mplus 6.12 program (Muthen & Muthen, 2011) together with the Q-matrix in Table 2, and with the help of LCDM, the individual probabilities of each PST's possession of the attributes were calculated. The PSTs' answers were not transferred directly to the SPSS program, and the total and average scores of them for each attribute were not calculated. The reason for doing this was that the total or average scores that the PSTs obtain from the test items do not give clear information about whether the PSTs have that attribute or not. For instance,

as presented in Table 6, a PST with a high total or average score may be less likely to have that attribute.

Table 6. Distribution of Four PSTs' Scores for Each Attribute

PST/Attribute	A1			A2			A3			A4		
	T	M	P	T	M	P	T	M	P	T	M	P
PST 17	5	.556	.929	5	.500	.064	4	.500	.579	1	.100	.119
PST 51	6	.667	.727	6	.600	.230	4	.500	.743	3	.300	.049
PST 268	5	.556	.004	5	.500	.000	3	.375	.997	3	.300	.000
PST 376	5	.556	.885	7	.700	.050	4	.500	.456	2	.200	.701

Note. T: Total item score; M: Mean item score; P: Probability of attribute possession

As shown in Table 2, the total maximum scores that the PSTs can receive from the items that measure A1, A2, A3, and A4 are 9, 10, 8, and 10, respectively. When Table 6 is examined, PST 17 received a total of 5 points from items measuring A1 (mean: .556); PST 51 received a total of 6 points for this attribute (mean: .667). Although, in terms of CTT, it is thought that PST 51 has more chance for mastering A1, LCDM analysis shows us that PST 17 has a higher probability of having this attribute than PST 51 (.727 < .929). Similarly, PST 268 obtained a total of 3 points for A4 (mean: .300). Although PST 376 received 2 points from A4 (mean: .200), PST 376 has more chance of having A4 than PST 268 (.000 < .701). PST 268's probability of having A4 is .00, and her probability of having A3 is .99. Moreover, although the points obtained by PST 17, PST 51, and PST 376 from the items measuring A3 are the same, they all have different probabilities for having this attribute. PST 51 has more chance for mastering A3 than the remaining PSTs, and PST 376 has less chance of having this attribute. The reason for the difference between the CTT and LCDM results in Table 6 can be explained by the fact that LCDM takes into account the possibility of nonmasters of any attribute answering these items correctly presumably by guessing, and attributes having different effects in obtaining correct answers. A PST who correctly answers an item may not necessarily have all the attributes associated with that item with the same probability. Furthermore, the PSTs' answers to the items measuring a specific attribute, as well as their answers to the items not measuring this attribute affect the calculation of probabilities.

After calculating the probability of each attribute, the responses were transferred into the SPSS program with their information about gender, ranking of the attended university, and grade levels. Next, the data were checked for normality by considering the skewness and kurtosis coefficients, Kolmogorov-Smirnov test, and graphs. It is expected that if the number obtained by dividing the skewness and kurtosis coefficients by their standard errors is between -1.96 and +1.96, the distribution of data does not differ significantly from the normal distribution (Kim, 2013). These values calculated respectively as -5.47 and -5.90 for A1; 7.48 and -3.88 for A2; 0.73 and -7.48 for A3; 2.84 and -7.24 for A4. As a result of conducting the Kolmogorov-Smirnov test, the p-value was found to be less than .05. Moreover, Histogram, Q-Q plot, and Box plot graphs were not satisfying normal distribution assumptions. Hence, it was concluded that the distribution of data was not normal. In addition, homogeneity of variance was examined by the Levene Test. Because the p-value was less than .05, the homogeneity of variance was not satisfied. Therefore, we determined that the data were not satisfying parametric test assumptions and so we used Mann-Whitney U test to investigate the effect of gender on the PSTs' possession of attributes and Kruskal Wallis-H test to investigate the effects of the ranking of the attended university and grade level on their possession of these attributes.

RESULTS

In this section, the findings of the PSTs' competencies in statistics and probability are reported in agreement with the sub-problems of the study.

Examining the PSTs' Competencies in Statistics and Probability According to Gender Variable

Mann-Whitney U test was used to investigate whether the probabilities of the PSTs' possessions of attributes differed according to their gender. The test results are presented in Table 7. According to Table 7, since the p-value for each attribute is greater than .05, the probabilities of the PSTs' possessions of attributes did not statistically differ according to their gender. The distribution of probabilities for each attribute according to the gender was presented in Table 8.

Table 7. Mann-Whitney U Test Results

	Gender	Mean Rank	Sum of Ranks	U	z
A1	Female	210.54	66321.50	16551.50[a]	-.418
	Male	216.25	23354.50		
A2	Female	214.77	67653.50	16136.50[a]	-.798
	Male	203.91	22022.50		
A3	Female	212.01	66783.50	17007.50[a]	-.003
	Male	211.97	22893.50		
A4	Female	209.78	66082.50	16312.50[a]	-.637
	Male	218.46	23594.50		

a. $p > .05$

Table 8. The Distribution of Probabilities According to Gender

Attributes	Gender	Probability
A1 Representing and interpreting data	Female	.663
	Male	.667
A2 Drawing inferences about populations based on samples	Female	.288
	Male	.271
A3 Selecting and using appropriate statistical methods to analyze data	Female	.462
	Male	.461
A4 Understanding and applying basic concepts of probability	Female	.423
	Male	.484

Examining the PSTs' Competencies in Statistics and Probability According to the Ranking of the Attended University

Kruskal Wallis-H test was used to investigate whether the probabilities of the PSTs' possessions of attributes statistically differed according to the ranking of the attended university. Next, the Mann-Whitney U test was applied to determine differences among high, middle, and low-ranking groups. The findings were presented in Table 9.

Table 9. The Probabilities of the PSTs' Possessions of Attributes According to the Ranking of the Attended University

	Group	N	Mean Rank	df	χ^2	Difference
A1	High	66	312.02	2	87.497**	High>Middle
	Middle	224	257.55			High>Low
	Low	166	156.10			Middle>Low
A2	High	66	208.96			
	Middle	224	222.94	2	4.091	-
	Low	166	243.78			
A3	High	66	156.67	2	97.445**	Low>Middle
	Middle	224	191.07			Low>High
	Low	166	307.57			Middle>High
A4	High	66	307.51	2	64.932**	High>Middle
	Middle	224	250.16			High>Low
	Low	166	167.86			Middle>Low

** $p < .01$

When Table 9 was examined, the p-value for A1 was found to be significant, $p < .01$. Table 9 showed that there was a statistically significant difference among all groups, and this difference was in favor of the university with a high base entrance score. The PSTs studying at the university with high base entrance scores were found to be more likely to have A1 than remaining PSTs. Moreover, the findings suggested that the higher the base entrance score of the university, the higher the probability of having A1. In terms of A2, the p-value was calculated as $p > .05$, and so we concluded that the PSTs' possessions of attributes did not statistically differ according to the ranking of the attended university. Although the mean likelihoods of having A2 were similar in each grade level, in general, each mean score was very low. For A3, the p-value was calculated as $p < .01$, and so we decided that there was a statistically significant difference between all groups. This difference was found to be in favor of universities with low base scores. The PSTs studying in a university with a low base score were more likely to have A3 than PSTs studying at a university with medium and high base scores. Furthermore, the PSTs attending a university with a high base score were less likely to have A3 than the PSTs studying at other universities. In addition, p was calculated as $p < .01$ for A4. There was a statistically significant difference among all groups in favor of the university with high base score. Therefore, the PSTs attending to the university with high base score were more likely to have A4 than the PSTs attending at the remaining universities. Thus, the higher the university ranking was, the higher the probability of the PSTs having A4.

When the above findings were considered, the PSTs had the most difficulty in having A2. There was a great chance of the PSTs attending at the university with a high base score for having A1 and A4 in comparison to the PSTs attending universities with medium or low base scores. Although the PSTs attending at the universities with low base scores were stronger in A3, they were weak in A1 and A4. Moreover, the probabilities of the PSTs' possessions of A2 did not differ statistically in terms of the ranking of the attended university, and each mean score was quite low. The distribution of the probabilities according to the ranking of the attended university was presented in Table 10.

Table 10. The Distribution of Probabilities According to Ranking of the Attended University

Attributes	Success Level	Probability
A1 Representing and interpreting data	High	.833
	Middle	.744
	Low	.441
A2 Drawing inferences about populations based on samples	High	.216
	Middle	.273
	Low	.332
A3 Selecting and using appropriate statistical methods to analyze data	High	.278
	Middle	.361
	Low	.710
A4 Understanding and applying basic concepts of probability	High	.653
	Middle	.474
	Low	.275

Examining the PSTs' Competencies in Statistics and Probability According to Grade Levels

Kruskal Wallis-H test was used to determine whether the PSTs' competencies in statistics and probability statistically differed according to their grade levels. The Mann-Whitney U test was used to determine which groups differed, and the findings were presented in Table 11.

When Table 11 is examined, the p-value for A1 was calculated as $p < .01$ which indicated a statistically significant difference among the groups. There was a significant difference between the PSTs attending to the first grade and second grade and between first grade and fourth grade, in favor of the first grade. Similarly, there was a significant difference between the PSTs attending to the third grade and second grade and between the third grade and fourth grade, in favor of the third grade. Moreover, the PSTs attending in the third grade had a higher probability of having A1 than the PSTs in remaining grades.

Table 11. The Probabilities of the PSTs' Possessions of Attributes According to Grade Levels

	Grade	N	Mean Rank	df	χ^2	Difference
A1	1	106	240.41	3	15.235**	1>2
	2	150	204.47			1>4
	3	158	253.63			3>2
	4	42	189.73			3>4
A2	1	106	227.30	3	5.546	--
	2	150	247.62			
	3	158	212.63			
	4	42	222.94			
A3	1	106	214.94	3	10.128*	2>1
	2	150	247.13			2>3
	3	158	210.47			4>1
	4	42	264.01			4>3
A4	1	106	243.97	3	24.357**	1>2
	2	150	196.31			1>4
	3	158	260.39			3>2
	4	42	184.44			3>4

* $p < .05$, ** $p < .01$

On the other hand, the PSTs attending the fourth grade had the lowest probability of having A1. For A2, the p-value was calculated as $p > .05$, and this finding showed that the PSTs' probabilities of having A2 did not significantly differ according to the grade levels. While the PSTs in the second grade had the highest probability of having A2, the third grade PSTs had the lowest probability of having A2. In terms of A3, the p-value was calculated as $p < .05$, and this finding suggested that there was a statistically significant difference among the groups. The difference was found to be significant between the PSTs attending to the second grade and first grade and between the second grade and third grade, in favor of the second grade. By the same token, there was a significant difference between the PSTs attending to the fourth grade and first grade and between the fourth grade and third grade, in favor of the fourth grade. Furthermore, while the fourth grade PSTs were more likely to have A3, the third grade PSTs were less likely to have A3. Regarding A4, the p-value was calculated as $p < .01$ that indicated a statistically significant difference among the groups. There was a significant difference between the PSTs attending to the first grade and second grade and between the first grade and fourth grade, in favor of the first grade. Similarly, there was a significant difference between the PSTs attending in the third grade and second and fourth grades in favor of the third grade. In addition, while the PSTs attending in the third grade were more likely to have A4, fourth grade PSTs were less likely to have A4. Overall, each grade level was found to be quite strong in mastering A1, but all levels were found to be quite weak in mastering A2. Finally, the PSTs attending in the second and fourth grades were quite strong in mastering A3, the PSTs attending in the third grade were strong in mastering A4. The relationship between the PSTs' probabilities of having each attribute and grade levels was presented in Table 12.

Table 12. Distribution of Probability of Having Attributes of PSTs According to Grade Levels

Attributes	Grade	Probability
A1 Representing and interpreting data	1	.702
	2	.574
	3	.716
	4	.528
A2 Drawing inferences about populations based on samples	1	.292
	2	.342
	3	.242
	4	.272
A3 Selecting and using appropriate statistical methods to analyze data	1	.432
	2	.532
	3	.420
	4	.597
A4 Understanding and applying basic concepts of probability	1	.453
	2	.332
	3	.528
	4	.326

DISCUSSION and CONCLUSION

This study investigated whether the PSTs' possession of four fundamental skills in statistics and probability differed according to gender, university entrance base score, and grade level variables by using their responses to the Statistical Reasoning Test developed by Arican and Kuzu (2019). The test measured four key skills, which are referred as attributes: Representing and interpreting data (A1), Drawing inferences about populations based on samples (A2) Selecting and using appropriate statistical methods to analyze data (A3), and Understanding and applying the basic concepts of probability (A4). The PSTs' responses were analyzed in the Mplus program using LCDM, one of the cognitive diagnostic models, and the probabilities of having attributes for each PST were calculated. Subsequently, these probabilities were examined in terms of gender, ranking of the attended university, and grade level variables.

The findings showed that the PSTs' possessions four key attributes in statistics and probability did not significantly differ according to gender. This result supports studies (e.g., Chiesi & Primi, 2015; Duckworth & Seligman, 2006; Else-Quest et al., 2010; Lindberg et al., 2010) indicating that the achievement gap in mathematics between female and male students is decreasing or ending. In terms of statistics and probability, this result is also consistent with the finding that eight grade female and male students obtained very close mean scores in the data and chance domain in TIMSS 2015 study (Mullis et al., 2016). On the other hand, this result differs from studies (e.g., Bulut et al., 2002) that emphasize that males are more successful in probability than females. While the probabilities of male and female PSTs' possession of Attribute 1 and Attribute 3 were very close to each other, male PSTs obtained a higher probability for the possession of Attribute 4, whereas female PSTs obtained higher probability for the possession of Attribute 2. This finding showed that female PSTs were more successful in making predictions and drawing inferences from data than male PSTs. Furthermore, male PSTs were more successful in understanding and applying the basic concepts of probability than female PSTs.

When the PSTs' possession of four attributes in statistics and probability are examined in terms of the attended universities' base entrance score levels (i.e., high, medium, low), there was a significant difference between all groups for A1, A3, and A4, and there was no statistically significant difference for A2. The analysis showed that the PSTs who were attending the university with a higher base entrance score were more successful in A1 and A4 than the other two groups. In their study with first-year students (i.e., freshman), Atuahene and Russell (2016) found that the students' university entrance scores made an extraordinary contribution to their performance in mathematics courses at the university level. Therefore, this result supports our finding that the PSTs attending the university with a higher base entrance score were more successful in A1 and A4 than the PSTs who were attending the remaining universities with lower scores. On the other hand, compared to the other two groups, the PSTs who were attending universities with lower base entrance scores were more successful in A3. In order for the PSTs to use appropriate statistical methods, they have to know rules and formulas learning which require mechanical and rote methods such as memorizing. For this reason, the PSTs attending universities with low base scores may possess this attribute more likely than the other two groups.

This study also examined whether the PSTs' possession of attributes in statistics and probability differed according to their grade levels. The findings showed that the PSTs' possession of attributes differed statistically for A1, A3, and A4, and no significant difference was found for A2. In terms of the probabilities of having A1 and A4, a significant difference was found among the PSTs attending first grade and second and fourth grades in favor of the first grade, and there was a significant difference among the PSTs attending to the third grade and second grade and between the third grade and fourth grade, in favor of the third grade. This result may be due to the fact that first-year PSTs had studied statistics and probability topics during the preparation process of university exams. Similarly, the success of the third year PSTs in having these attributes can be explained by the fact that statistics and probability courses are provided in the third year in mathematics education programs. Therefore, the PSTs' past experiences on statistics and probability made a positive effect on their possession of Attribute 1 and Attribute 4. It is noteworthy that except Attribute 3, the probabilities of fourth grade

PSTs' possessions of attributes were lower than the other grade levels. Although the PSTs are expected to be well prepared for teaching statistics and probability topics in their last year of the program, this finding revealed an opposite condition. Therefore, as stated by Batanero and Díaz (2012), Batanero et al., (2004), and Franklin and Mewborn (2006), the fourth grade PSTs' lack of three fundamental attributes pointed to the shortcomings of higher education programs in terms of teaching statistics and probability topics.

Suggestions

In this study, although the PSTs' probabilities of having four attributes varied according to the ranking of the attended university and grade levels, their probabilities of having Attribute 1 were generally high for these two variables. However, their probabilities of having the remaining three attributes, especially Attribute 2, were quite low. Therefore, this result suggests that teacher education programs should be planned more effectively for teaching statistics and probability topics. For this purpose, real-life activities should be prepared in order to increase the PSTs' cognitive competence and to generate their meaningful learning of statistics and probability topics. These activities should be included in secondary and higher education programs and associated with the standards existed in curricula. In addition, although little known about CDMs in comparison to CTTs, which is one of the limitations of this study, it is important that they provide a different perspective on the field. For this reason, the inclusion of CDMs in mathematics education studies will allow educators providing diagnostic evaluations and solution suggestions for the problems encountered.

REFERENCES

Ardıç, E. Ö., Yılmaz, B., & Demir, E. (2012, June). *İlköğretim 8. sınıf öğrencilerinin merkezi eğilim ve yayılım ölçüleri hakkındaki istatistiksel okuryazarlık düzeylerinin solo taksonomisine göre incelenmesi*. Paper session presented at the meeting of X. Fen Bilimleri ve Matematik Eğitimi Kongresi, Niğde, Türkiye. Retrieved from http://kongre.nigde.edu.tr/xufbmek/dosyalar/tam_metin/pdf/2430-30_05_2012-18_28_56.pdf

Arican, M., & Kuzu, O. (2019). Diagnosing preservice teachers' understanding of statistics and probability: Developing a test for cognitive assessment. *International Journal of Science and Mathematics Education*. 1-20. Advance online publication. doi: 10.1007/s10763-019-09985-0, Online first.

Atuahene, F., & Russell, T. A. (2016). Mathematics readiness of first-year university students. *Journal of Developmental Education, 39*(3), 12-20. Retrieved from www.jstor.org/stable/44987415

Batanero, C., & Díaz, C. (2010). Training teachers to teach statistics: What can we learn from research? *Statistique et Enseignement, 1*(1), 5-20. Retrieved from http://statistique-et-enseignement.fr/article/view/3

Batanero, C., & Díaz, C. (2012). Training school teachers to teach probability: Reflections and challenges. *Chilean Journal of Statistics, 3*(1), 3-13. Retrieved from http://chjs.mat.utfsm.cl/volumes/03/01/Batanero_Diaz(2012).pdf

Batanero, C., Godino, J. D., & Roa, R. (2004). Training teachers to teach probability. *Journal of statistics Education, 12*(1). Retrieved from http://www.amstat.org/publications/jse/v12n1/batanero.html

Boyacıoğlu, H., Erduran, A., & Alkan, H. (1996, September). *Permütasyon, kombinasyon ve olasılık öğretiminde rastlanan güçlüklerin giderilmesi*. Paper session presented at the meeting of II. Ulusal Eğitim Sempozyumu, Marmara University, İstanbul.

Bradshaw, L., Izsak, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice, 33*(1), 2-14. doi: 10.1111/emip.12020

Bulut, S., Yetkin, İ. E., & Kazak, S. (2002). Matematik öğretmen adaylarının olasılık başarısı, olasılık ve matematiğe yönelik tutumlarının cinsiyete göre incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 22*(22), 21-28. Retrieved from https://dergipark.org.tr/en/download/article-file/87889

Çepni, S., Bayrakçeken, S., Yılmaz, A., Yücel, C., Semerci, Ç., Köse, E., …, Gündoğdu, K. (2008). *Ölçme ve değerlendirme*. Ankara: Pegem Akademi.

Chambers, E. A., & Schreiber, J. B. (2004). Girls' academic achievement: Varying associations of extracurricular activities. *Gender and Education, 16*(3), 327-346. doi: 10.1080/09540250042000251470

Chiesi, F., & Primi, C. (2015, February). Gender differences in attitudes toward statistics: Is there a case for a confidence gap? In K. Krainer & N. Vondrova (Eds.), *CERME 9-Ninth congress of the European society*

for research in mathematics education (pp. 622-628). Prague, Czech Republic: Charles University & ERME.

Choi, K. M., Lee, Y. S., & Park, Y. S. (2015). What CDM can tell about what students have learned: An analysis of TIMSS eighth grade mathematics. *Eurasia Journal of Mathematics, Science & Technology Education, 11*(6), 1563-1577. doi: 10.12973/eurasia.2015.1421a

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*(4), 343-362. doi: 10.1111/j.1745-3984.2008.00069.x

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179-199. doi: 10.1007/s11336-011-9207-7

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood–based classification techniques. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361-390). Hillsdale, NJ: Lawrence Erlbaum.

Dogan, E., & Tatsuoka, K. (2008). An international comparison using a diagnostic testing model: Turkish students' profile of mathematical skills on TIMSS-R. *Educational Studies in Mathematics, 68*(3), 263-272. doi: 10.1007/s10649-007-9099-8

Duckworth, A. L., & Seligman, M. E. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology, 98*(1), 198-208. doi: 10.1037/0022-0663.98.1.198

Eitle, T. M. (2005). Do gender and race matter? Explaining the relationship between sports participation and achievement. *Sociological Spectrum, 25*(2), 177-195. doi: 10.1080/02732170590883997

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*(1), 103-127. doi: 10.1037/a0018053

Farooq, M. S., Chaudhry, A. H., Shafiq, M., & Berhanu, G. (2011). Factors affecting students' quality of academic performance: A case of secondary school level. *Journal of Quality and Technology Management, 7*(2), 1-14. Retrieved from http://pu.edu.pk/images/journal/iqtm/PDF-FILES/01-Factor.pdf

Felson, R. B., & Trudeau, L. (1991). Gender differences in mathematics performance. *Social Psychology Quarterly, 54*(2), 113-126. doi: 10.2307/2786930

Franklin, C., & Mewborn, D. (2006). The statistical education of PreK-12 teachers: A shared responsibility. In G. Burrill (Ed.), *NCTM 2006 Yearbook: Thinking and reasoning with data and chance* (pp. 335-344). Reston, VA: NCTM.

Franklin, C., Kader, G., Mewborn, D. S., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-K-12 curriculum framework.* Alexandria, VA: American Statistical Association. Retrieved from https://www.amstat.org/asa/files/pdfs/gaise/gaiseprek-12_full.pdf

Fryer Jr, R. G., & Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics, 2*(2), 210-40. doi: 10.1257/app.2.2.210

Gürbüz, R., Toprak, Z., Yapıcı, H., & Doğan, S. (2011). Subjects perceived as difficult in secondary mathematics curriculum and their reasons. *Gaziantep University Journal of Social Sciences, 10*(4), 1311-1323. Retrieved from https://dergipark.org.tr/en/download/article-file/223364

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practice* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191-210. doi: 10.1007/s11336-008-9089-5

Im, S., & Park, H. J. (2010). A comparison of US and Korean students' mathematics skills using a cognitive diagnostic testing method: Linkage to instruction. *Educational Research and Evaluation, 16*(3), 287-301. doi: 10.1080/13803611.2010.523294

Jones, G. A. (2005). *Exploring probability in school: Challenges for teaching and learning*. New York, NY: Springer.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric Madde response theory. *Applied Psychological Measurement, 25*(3), 258-272. doi: 10.1177/01466210122032064

Karasar, N. (2005). *Bilimsel araştırma yöntemleri.* Ankara: Nobel Yayınevi

Kim, H. Y. (2013). Statistical notes for clinical researchers: Assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics, 38*(1), 52-54. doi: 10.5395/rde.2013.38.1.52

Lee, Y. S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing, 11*(2), 144–177. doi: 10.1080/15305058.2010.534571

Leighton, J. P., & Gierl, M. J. (2007). Why cognitive diagnostic assessment? In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education* (pp. 3-18). Cambridge: Cambridge University Press.

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136*(6), 1123-1135. doi: 10.1037/a0021276

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal, 8*(1), 82-105. Retrieved from https://www.stat.auckland.ac.nz/~iase/serj/SERJ8(1).pdf#page=85

Milli Eğitim Bakanlığı. (2013). *Ortaokul matematik dersi öğretim programı*. Ankara: Milli Eğitim Bakanlığı Talim ve Terbiye Kurulu Başkanlığı.

Milli Eğitim Bakanlığı. (2018). *Matematik dersi öğretim programı.* Ankara: Milli Eğitim Bakanlığı Talim ve Terbiye Kurulu Başkanlığı.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Amsterdam: International Association for the Evaluation of Educational Achievement. Retrieved from https://pdfs.semanticscholar.org/9802/a1fabea7578ffd251e50bec4ac13831fbca0.pdf

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). TIMSS 2015 *international results in mathematics.* Boston College: TIMSS & PIRLS International Study. Retrieved from http://timssandpirls.bc.edu/timss2015/international-results/

Muthen, L. K., & Muthen, B. O. (2011). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthen & Muthen.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (2012). *Cognitively diagnostic assessment*. Hillsdale, NJ: Erlbaum.

Olpak, Y. Z., Baltaci, S., & Arican, M. (2018). Investigating the effects of peer instruction on preservice mathematics teachers' achievements in statistics and probability. *Education and Information Technologies*, *23*(6), 2323-2340. doi: 10.1007/s10639-018-9717-3

Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic modeling using R. *Practical Assessment, Research & Evaluation, 20*(11), 1-12. doi: 10.7275/5g6f-ak15

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic assessment: Theory, methods, and applications.* New York, NY: Guilford Press.

Saraçbaşı, T., & Kutsal, A. (1987). *Betimsel istatistik.* Ankara: Hacettepe Üniversitesi.

Sen, S., & Arican, M. (2015). A diagnostic comparison of Turkish and Korean students' mathematics performances on the TIMSS 2011 assessment. *Journal of Measurement and Evaluation in Education and Psychology, 6*(2), 238-253. doi: 10.21031/epod.65266

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 957-1009). Reston, VA: The National Council of Teachers of Mathematics.

Stoet, G., & Geary, D. C. (2013). Sex differences in mathematics and reading achievement are inversely related: Within-and across-nation assessment of 10 years of PISA data. *PloS One*, *8*(3), 1-253. doi: 10.1371/journal.pone.0057988

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30*(2), 251-275. doi: 10.1007/s00357-013-9129-4

Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287-305.

Tsakiridou, H., & Vavyla, E. (2015). Probability concepts in primary school. *American Journal of Educational Research, 3*(4), 535-540. doi: 10.12691/education-3-4-21

Ulutaş, F., & Ubuz, B. (2008). Matematik eğitiminde araştırmalar ve eğilimler: 2000 ile 2006 yılları arası. *İlköğretim Online, 7*(3), 614-625. Retrieved from http://ilkogretim-online.org.tr/index.php/io/article/view/1751/1587

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Report No. RR-05-16). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.2005.tb01993.x

Watson, J. M. (2006). *Statistical literacy at school: Growth and goals.* Mahwah, NJ: Lawrence Erlbaum.

Zhang, X., & Maas, Z. (2019). Using R as a simulation tool in teaching introductory statistics. *International Electronic Journal of Mathematics Education, 14*(3), 599-610. doi: 10.29333/iejme/5773

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(1); 27-42

Item Parameter Estimation for Dichotomous Items Based on Item Response Theory: Comparison of BILOG-MG, Mplus and R (ltm)*

Şeyma UYAR [**] Neşe ÖZTÜRK GÜBEŞ [***]

Abstract
The aim of this study is twofold. The first one is to investigate the effect of sample size and test length on the estimation of item parameters and their standard errors for the two parameter item response theory (IRT). Another is to provide information about the performance of Mplus, BILOG-MG and R (ltm) programs in terms of parameter estimation under the conditions which were mentioned above. The simulated data were used in this study. The examinee responses were generated by using the open-source program R. After obtaining the data sets, the parameters were estimated in BILOG-MG, Mplus and R (ltm). The accuracy of the item parameters and ability estimates were evaluated under six conditions that differed in the numbers of items and examinees. After looking at the resulting bias and root mean square error (RMSE) values, it can be concluded that Mplus is an unbiased program when compared to BILOG-MG and R (ltm). BILOG-MG can estimate parameters and standard errors close to the true values, when compared to Mplus and R (ltm).

Key Words: IRT, parameter estimation, Mplus, BILOG-MG, ltm

INTRODUCTION

In recent years, especially in the fields of education and psychology, item response theory (IRT) has been popular (Foley, 2010). Provision of the opportunity of modelling the relationship between examinees' ability and their response to an item, makes IRT models more preferable than classical test theory models (CTT) (de Ayala, 2009; Hambleton, Swaminathan, & Rogers, 1991; Yen & Fitzpatrick, 2006). CTT focuses on the number of correct answers given by the examinee in the test. In other words, two examinees with the same number of correct answers get the same score in terms of the measured property, regardless of whether the item is difficult or easy (Proctor, Teo, Hou & Hsieh, 2005). Moreover, the major advantage of CTT is that it is easy to meet the assumptions in real test data (Fan, 1998; Hambleton & Jones, 1993). On the other hand, IRT requires stronger assumptions than CTT (Crocker & Algina, 1986). IRT is based on the probability of an examinee's ability to perform on any item according to his or her ability. IRT models are functions of items, characterized by item parameters, and the ability of the examinees. As its name implies, IRT models test the behavior at the item level. IRT models can be unidimensional or multidimensional. In this study, we considered only unidimensional IRT models. There are three item parameters used in unidimensional IRT models. These are difficulty, *b*; discrimination, *a*; and pseudo-guessing, *c* parameters (Hambleton, Swaminathan & Rogers, 1991; Van Der Linden & Hambleton, 1997).

Unidimensional IRT models vary in the number of item parameters that are used. The one parameter logistic (1PL) model assumed that all items have an equal discrimination index and the probability of guessing an item correctly is zero. In the three parameter logistic (3PL) model all three item parameters vary across items. And in the two parameter logistic (2PL) model only the item difficulty and discrimination indices vary across items (Lord, 1980). The item response function for the two parameter logistic (2PL) model is defined as follows:

* This study was presented as oral presentation at 6th International Congress on Measurement and Evaluation in Education and Psychology in KOSOVO.

** Dr, Mehmet Akif Ersoy University Faculty of Education, Burdur-Turkey, syuksel@mehmetakif.edu.tr, ORCID ID: 0000-0002-8315-2637
*** Dr, Mehmet Akif Ersoy University Faculty of Education, Burdur-Turkey, nozturk@mehmetakif.edu.tr, ORCID ID: 0000-0003-0179- 1986

To cite this article:
Uyar, Ş. & Öztürk-Gübeş, N. (2020). The Importance of Sample Weights and Plausible Values in Large-Scale Assessments. *Journal of Measurement and Evaluation in Education and Psychology, 11*(1), 27-42. doi: 10.21031/epod.591415
Received: 12.07.2019
Accepted: 06.01.2020

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1+e^{Da_i(\theta - b_i)}} \quad \text{(i=1, 2, …, n)} \qquad (1)$$

where $P_i(\theta)$ is the probability that a randomly selected examinee with ability θ answers item i correctly. The parameter b_i is referred to as index to item difficulty or threshold parameter and describes the point on the ability scale at which an examinee has a 50 percent probability of answering item i correctly. The discrimination parameter a_i is propotional to the slope of $P_i(\theta)$ at point $\theta = b_i$. The constant D is a scaling factor that places the scale of the latent ability approximately on the standard normal metric when set to 1.7 (Hambleton & Swaminathan, 1985).

One of the advantages of IRT is that item parameters can be estimated independent of the group and ability parameters can be estimated independent of the item (Hambleton, Swaminathan & Rogers, 1991). For this reason, IRT provides an appealing conceptual framework for test development (Hambleton, 1989) and IRT-based item and ability estimations are frequently mentioned in test development studies. The aim of test development studies is to present the models which can estimate the most accurate and stable item and ability parameters. The estimation of parameters is important because the examinees' reported score based on these parameters can affect any decision about examinees. For this reason, researchers aim to reveal the most accurate model to estimate the parameters in various conditions (Rahman & Chajewski, 2014).

In the literature, the effect of sample size and test length on parameter estimation is frequently investigated in IRT based test development studies. In these studies (Lim & Drasgow, 1990; Lord, 1968; Öztürk-Gübeş, Paek & Yao, 2018; Patsula & Gessroli, 1995; Şahin & Anıl, 2017; Yen, 1987; Yoes, 1995) although the minimum number of sample size and the exact length of the test cannot be certainly specified (Foley, 2010), the optimal number of sample size and test length which should be reached under various conditions can be revealed. The common point of these studies is that the number of sample size and test length should be particularly large in complex models and IRT models require large sample size to make accurate parameter estimations (Hambleton, 1989; Hulin, Lissak & Drasgow, 1982).

Lord (1968) stated that, at least 50 items and 1000 sample sizes were required to estimate the discriminant parameter (a parameter) accurately for the 3PL model. Swaminathan and Gifford (1983) investigated the effect of sample size, test length, and the ability distribution on the estimation of item and ability parameters using the 3-PL model. Their results showed that the condition in which sample size was 1000 and test length was 20 produced more accurate estimates of the difficulty and guessing parameters, and fairly good estimates of the item discrimination parameters than the conditions in which sample size was 50 and test lengths were 10 or 15 and sample size was 200 and test lengths were 10 and 15. Hulin et al. (1982) suggested that at least 500 samples and 30 items were needed for the 2PL model. They also suggested that the number of sample size should be 1000 and the number of items should be 60 for the 3PL model or when sample size was 2000, test length should be 30. Also, for the 2PL model, Lim & Drasgow (1990) suggested 750 as the sample size for 20 items; Şahin and Anıl (2017) suggested 500 as the sample size for 20 items and Gübeş, Paek and Yao (2018) pointed out that when the sample size was 500 or greater, estimation methods produced same and appropriate results with the test lengths of 11 (small) , 22 (medium) or 44 (large).

In many test applications, it is not always possible to increase the sample size or test length. Therefore, in recent times researchers focus on the use of the most accurate model and computer program according to the sample size or test length. Baker (1987) stated that the parameter estimation and the computer program that is used constitute an inseparable whole. And the characteristics of the obtained parameters will be affected by the underlying mathematics of the program. For this reason, many computer programs are available at various times depending on the possibilities offered by technology. BILOG-MG (Zimowski et al., 2003) has been widely used for parameter estimation in dichotomous items and has a long history (Baker, 1990; Lim & Drasgow, 1990; Swaminathan & Gifford, 1983). Recently, IRT analyses have been conducted using the libraries (e.g. package ltm, irtoys) in the open source program R (Rizopoulos, 2006, 2013; Bulut & Zopluoğlu, 2013; Pan, 2012). Mplus (Muthén & Muthén, 1998-2012) is another program that is preferred in analyzing latent models. Although there are a lot of programs for parameter estimation, they are questionable in terms of making accurate estimates.

Therefore, simulation studies can be effective to evaluate the accuracy of estimations. Such studies allow researchers to compare the estimation results with the true values in various test conditions (Şahin & Colvin, 2015).

Yen (1987), compared the performance of BILOG and LOGIST in terms of parameter estimates and item characteristic functions for the three-parameter logistic model. They used 1000 sample size with 10, 20 and 40 test lengths. They indicated that BILOG always produced more accurate estimates of item parameters especially in short tests. But they pointed out that two programs performed equally for the 20 and 40 item tests. Mislevy & Stocking (1989) recommended using BILOG in short tests and/or small examinee samples, while LOGIST might be preferred in longer tests.

Şahin and Colvin (2015) investigated the accuracy of the item and ability parameters which were obtained from "ltm" R package. They compared item and ability estimates with the true parameters when test lengths were 20 and 40 and sample sizes were 250, 1000 and 2000. They considered bias, mean absolute deviation (MAD), and root mean square error (RMSE) for the evaluation of accuracy of "ltm package" in terms of parameter estimation. According to their findings, it can be concluded that accurate estimates with the 1PL, 2PL, and 3PL can be provided by using ltm. Especially to estimate b parameters, ltm produced more accurate results. Their findings showed that while ltm estimated difficulty and ability parameters accurately there were some problems in guessing parameter (c) estimates. Results obtained from all the conditions showed that the accuracy of parameter estimation with ltm increased in all the three models as the number of examinees increased.

Rahman and Chahewski (2014) investigated the calibration results of 2PL and 3PL IRT models with 100 items and 1000 examinees in BILOG-MG, PARSCALE, IRTPPRO, flexMIRT, and R (ltm). They mentioned that ltm is the only software with a negative bias for the discrimination and guessing parameters while estimating the 3PL model. Their findings indicated that BILOG and PARSCALE underestimate item difficulties and latent traits, whereas IRTPRO and flexMIRT mostly overestimate them for 2PL models. And, R package ltm also showed negligible bias for item difficulty in 2 PL models. The package ltm is unable to perform with the other software programs in 3 PL models, but its recovery is precise for the latent trait using the 2PL model. Although there is some research about comparing performance of computer programs in IRT model parameter estimates, it is still necessary to conduct more research to compare the performance of different programs in parameter estimating.

The aim of this study is to investigate the effect of sample size and test length on the estimation of item parameters and their standard errors in 2PL models. Another aim of this study is to compare the performance of Mplus, BILOG-MG and R (ltm) in terms of parameter estimation in different sample sizes and test lengths. This study will contribute to the discussions about sufficient sample size or test length when studies are conducted based on IRT. On the other hand, the researchers will be able to get information about which of the programs they need to access in accordance with the available data or the parameters to be estimated. This research is original as it includes standart error comparison of parameters. The data which was simulated based on the parameters of a real test was used in the current study.

The basic problem investigated in the current study was "How do the parameters and their standard error estimates change in the BILOG-MG, Mplus and R (ltm) programs when the test length and sample size change?

METHOD

This research is a simulation based study examined the performance of different programs in terms of parameter estimation under specific conditions.

Data Generation

The simulated data were used in this study. To mimic a real test situation, examinee responses were generated based on TIMSS 2015 mathematic test item parameters. The mean and standard deviation of item parameters which were used in data generation were given in Table 1.

Table 1. Item Parameters Means and Standard Deviations Obtained from TIMSS 2015 Application

	Test length = 30				**Test length = 60**			
	a	**se (a)**	**b**	**se (b)**	**a**	**se (a)**	**b**	**se (b)**
Mean	1.22	0.09	0.70	0.05	1.24	0.09	0.66	0.05
Std. dv.	0.35	0.04	0.54	0.04	0.37	0.05	0.54	0.03

Std. dv: Standart deviation

Furthermore, the ability parameters are drawn from a standard normal distribution which has mean zero and standard deviation one, N~(0,1). For the response of the *i*th item and *n*th examinee; firstly, item response function was calculated based on 2PL model (see equation 1) then uniform random numbers were sampled from (0, 1). If the uniform random number was equal or less than the probability of correctly answering item, item was scored as 1 (correct). Otherwise, item *i* was scored as 0 (incorrect).

In data simulation, test length and sample size were varied: sample sizes were 500, 1000 and 2000; test lengths were 30 and 60. In the current study, 3 sample sizes and 2 test lengths conditions yielded to generate six different data conditions. For each condition, 50 data sets were generated, which resulted in 300 generated response sets. Six simulation conditions are given in Table 2.

Table 2. Simulation Conditions

Condition	Sample Size	Number of Items
1	500	30
2	1000	30
3	2000	30
4	500	60
5	1000	60
6	2000	60

Data Analysis

In the first step of the data analysis, item parameters were estimated by using the Maximum Likelihood Estimation (MLE) method according to 2PL model for each condition of test length and sample size. Parameters were estimated in BILOG-MG, Mplus and R (ltm). In all the programs, default settings were used.

Mplus is a statistical modeling program which has a flexible modeling capacity. Mplus allows researchers to do factor analysis, mixture modeling and structural equation modeling. In Mplus, categorical and continuous data that have single-level or multi-level structure can be analyzed. In addition, Mplus has extensive facilities for Monte Carlo simulation studies. Normally, non-normally distributed, missing or clustering data can be generated by using Mplus (Muthén & Muthén, 1998, 2002, 2012).

BILOG-MG is a software program that is designed for analysis, scoring and maintenance of measurement instruments within the framework of IRT. The program is appropriate for the binary items scored right, wrong, omitted- or non-presented. The program is concerned with estimating the parameters of an item and the position of examinees on the underlying latent trait (Zimowski et al., 2003).

Latent trait models which is shortly abbreviated as "*ltm*" is an open-source R software package. ltm can do analysis of univariate and multivariate dichotomous and polytomous data using latent trait models under the IRT. The package includes IRT models of Rasch, 2PL, 3PL, graded response and generalized partial credit (Rizopoulos, 2006). In the current study, analyses based on latent trait models were run under another R package, *irtoys* . The irtoys is a package which combined some useful IRT programs. These programs are ICL, BILOG-MG and ltm. In the installing process of irtoys the ltm package is also automatically loaded (Partchev, 2017).

In the second step of the data analysis, the accuracy of item parameters was investigated by computing discrepancy between the estimate and true value of the parameter. In order to evaluate the recovery of item parameters and their standard errors, bias and root mean square error (RMSE) were calculated. Bias is defined as the average difference between true and estimated parameters. It is a measure of any systematic error in estimation. To obtain the average bias value, bias was calculated for each replication of each condition, and then an average bias for each condition was calculated. Bias can take both positive and negative values. When the bias value is zero and close to zero, it can be decided that the parameter estimation is unbiased. RMSE is a measure of precision that, like standard deviation, provides information about the average magnitude of parameter variation around the true parameter. RMSE always yields positive values and the minimum value of RMSE is zero. If the RMSE value obtained in the relevant condition is close to zero, it is decided that the estimation stability is high. As the RMSE value moves away from zero it is interpreted as low estimation stability. For a given parameter, bias and RMSE indexes were calculated as in equations 2 and 3:

$$Bias = \left(\frac{1}{R}\right)\sum_{r=1}^{R}\hat{\varphi}_r - \varphi \qquad (2)$$

$$RMSE = \sqrt{\sum_{r=1}^{R}(\hat{\varphi}_r - \varphi)^2/R} \qquad (3)$$

where φ is the parameter of interest and r is the replication number index (r = 1, 2, ... , R). In the item parameter recovery investigation, each of the data generating parameters is φ. These indices were averaged across all items to compute summary indices for a given condition.

RESULTS

The averages of RMSE and bias value for the estimated parameters in Mplus, BILOG-MG and R (with ltm) programs across the 50 runs are given in Table 3.

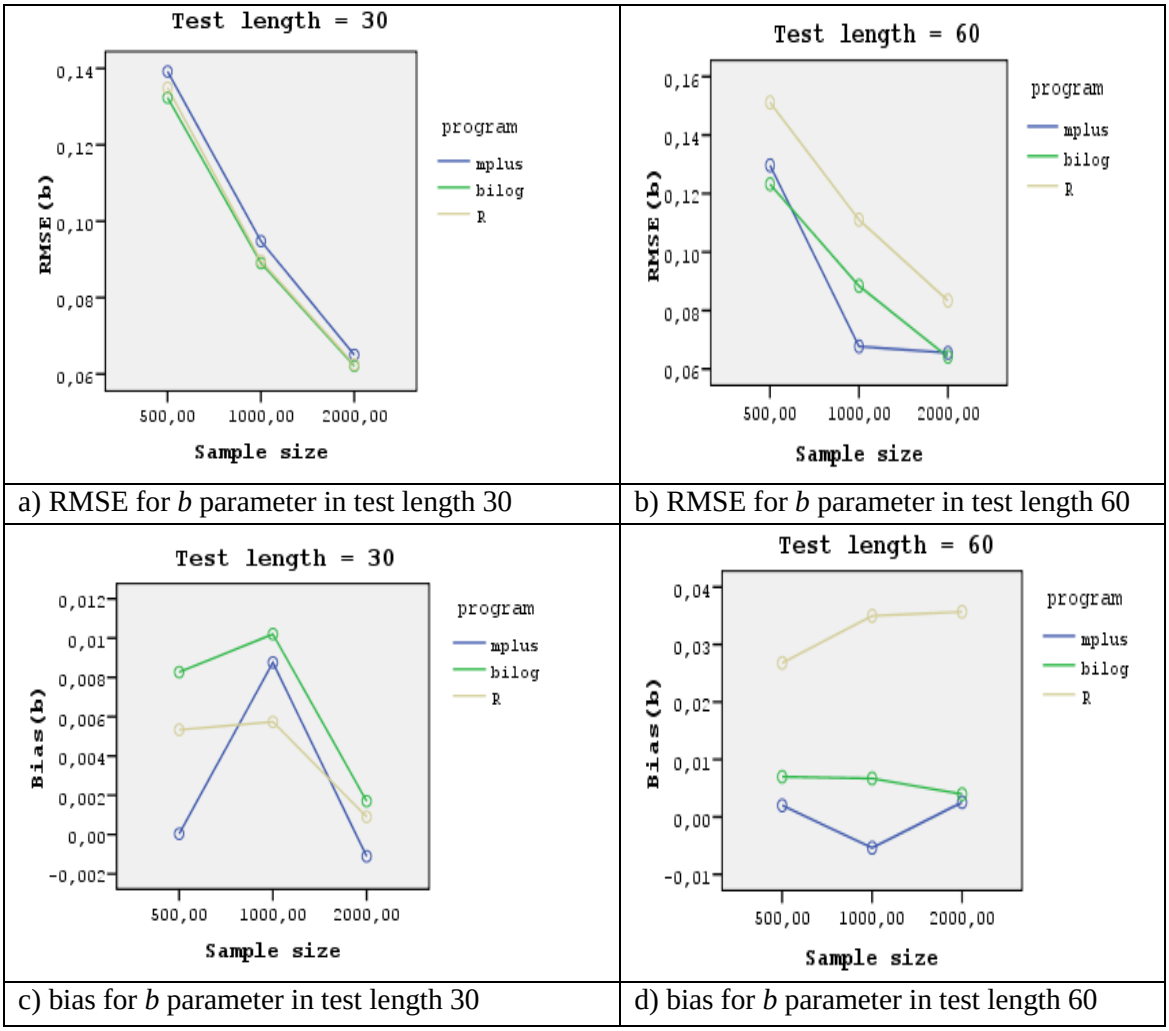
Table 3. RMSE and Bias Averages for Item Parameters and Standard Errors

	RMSE				**Bias**			
	b	se (b)	a	se (a)	b	Se (b)	a	se (a)
Mplus	0,092	0,054	0,112	0,051	0,001	0,046	0,004	0,030
BILOG-MG	0,093	0,046	0,111	0,042	0,006	0,036	-0,012	0,018
R (ltm)	0,109	0,056	0,121	0,044	0,023	0,047	-0,037	0,019

As seen in Table 1, while the *b* parameter estimates of Mplus have the smallest average of RMSE values (0.092), R (ltm) estimates have the largest average (0.109). On the other hand, the standard error of *b* parameter estimates of BILOG-MG program has the smallest RMSE average (0.046), and again R (ltm) estimates have the largest RMSE average (0.056). The slope (*a*) parameter estimates of BILOG-MG have the smallest RMSE average (0.111) and R (ltm) estimates have the largest value (0.121). Similarly, BILOG-MG program has the smallest RMSE average (0.042) for the standard error of a parameter but Mplus estimates have the largest values (0.051).

Considering the bias values in Table 3, it can be said that the Mplus program has the smallest bias values for *a* (0.004) and *b* parameters (0.001); BILOG-MG has the smallest bias values for the se(*b)* (0.036) and *se(a)* (0.018) parameters. While, the R (ltm) has the largest mean of bias values for the *b* (0.023), *se(b)* (0.047) and *a* (-0.037) parameters; Mplus program has the largest bias values for the *se(a)* (0.030) parameter.

For each of the six conditions, the average of RMSE and bias values for the "*b*" parameter over 50 replications are plotted in Figure 1.



a) RMSE for *b* parameter in test length 30 | b) RMSE for *b* parameter in test length 60

c) bias for *b* parameter in test length 30 | d) bias for *b* parameter in test length 60

Figure 1. The Graphics for *b* Parameter Based on RMSE and Bias

As seen in Figure 1a and Figure 1b where test lengths were 30 and 60, as sample size increased the RMSE values of *b* parameter estimates decreased in all programs. When test length was 30 and sample size was 500, while BILOG-MG *b* parameter estimates had the smallest RMSE values, Mplus had the largest ones. When sample size was 1000, the *b* parameter estimates of R and BILOG-MG programs had similar and smaller RMSE values than Mplus. When the sample size was 2000, although BILOG-MG and R (ltm) had similar and smaller RMSE values than Mplus, Mplus got very close RMSE values to other two programs (see Figure 1a).

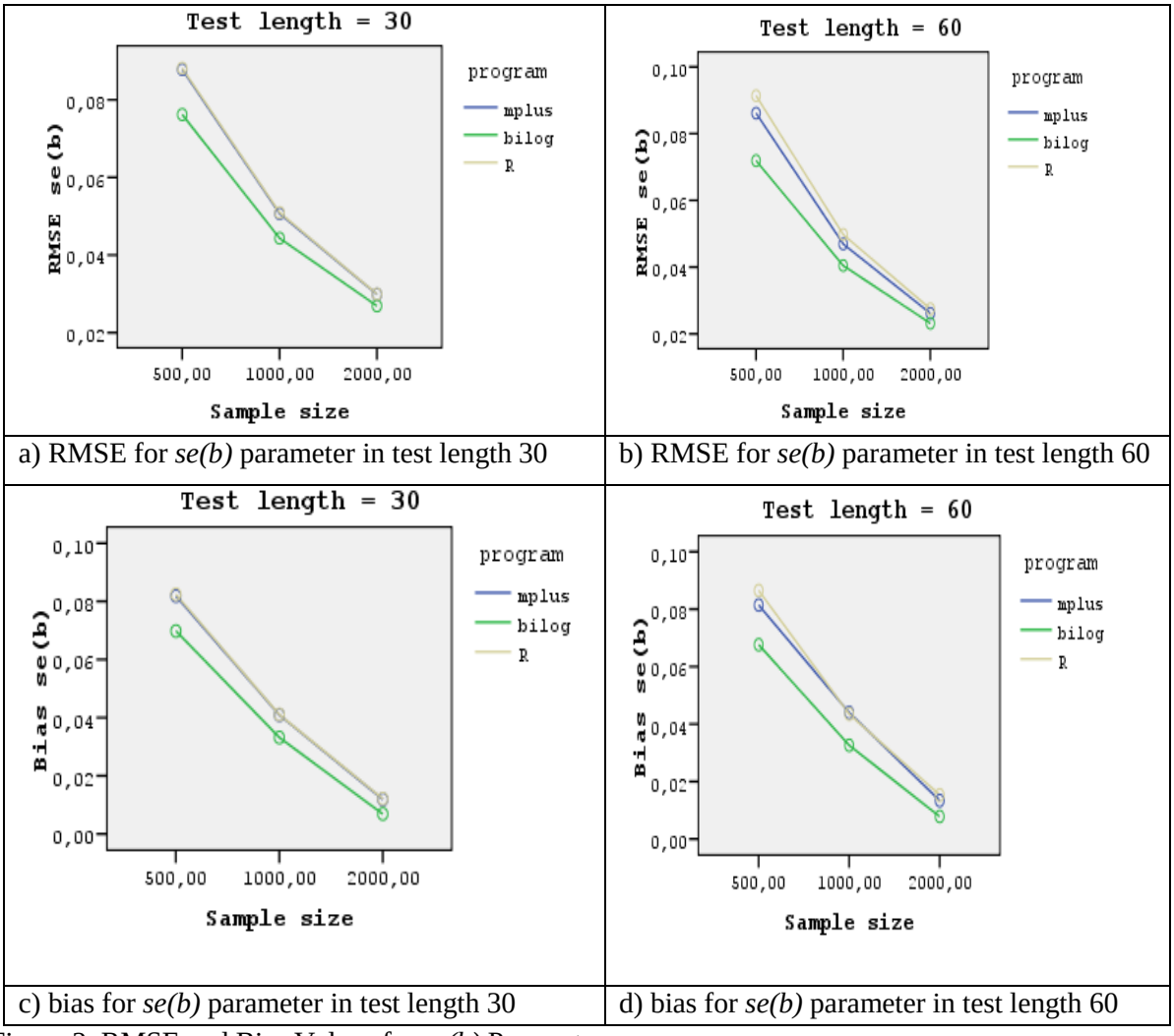
When we consider RMSE values for the *b* parameter at test length 60 in Figure 1b, we can say that BILOG-MG had the smallest and R (ltm) had the largest values. On the other hand, at the test length 1000, while Mplus had the smallest RMSE values, again R (ltm) had the largest values. When sample size was 2000, Mplus and BILOG-MG programs had similar and smaller RMSE values than R (ltm).

We can say that in all sample sizes at the test length 60, based on RMSE index, R (ltm) performed worse than other programs in terms of estimating *b* parameter.

The graphic in Figure 1c showed that at the test length 30, the smallest bias values for the *b* parameter were obtained by Mplus and the largest ones were obtained by BILOG-MG program. However, at the sample size 1000, R (ltm) had the smallest bias values and again BILOG-MG had the largest RMSE values. At the sample size 2000, while Mplus had the smallest bias values, again BILOG-MG had very close but larger bias than R (ltm). Also, when sample size increased from 500 to 1000, bias values of *b* parameter estimates from all programs increased but as sample size increased from 1000 to 2000, bias values decreased (see Figure 1c).

If we consider bias values for the *b* parameter at the test length of 60 and sample sizes of 500 and 1000, while the smallest bias values were obtained by Mplus, the largest ones got from R program. At the sample size of" 2000, bias values for b parameter estimates of R program were larger than other programs but BILOG-MG estimates had very close bias values to Mplus program (see Figure 1d).

For each of six conditions, the average of RMSE and bias values for the "se(*b*)" parameter over 50 replications were plotted in Figure 2.



a) RMSE for *se(b)* parameter in test length 30

b) RMSE for *se(b)* parameter in test length 60

c) bias for *se(b)* parameter in test length 30

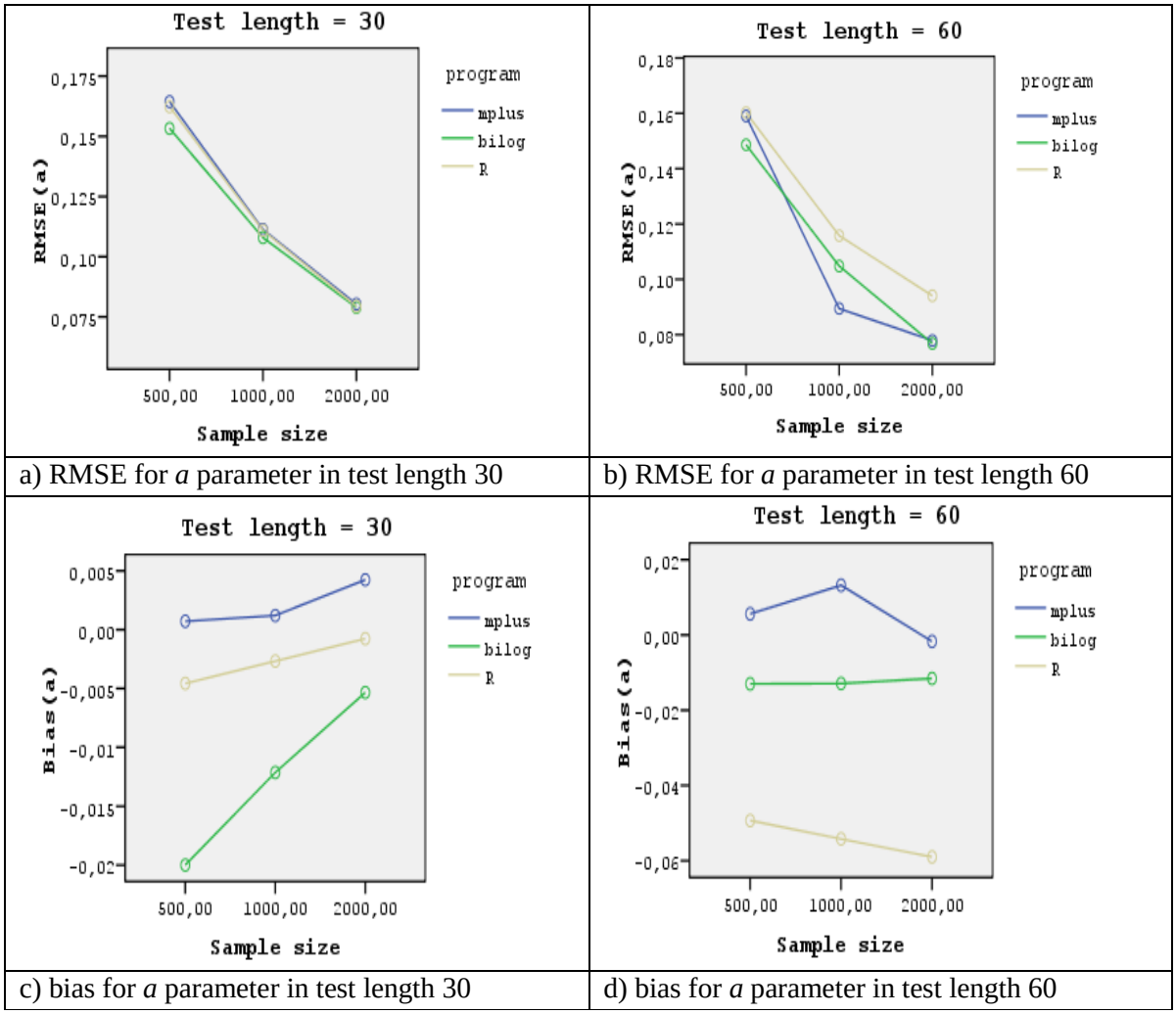
d) bias for *se(b)* parameter in test length 60

Figure 2. RMSE and Bias Values for *se(b)* Parameter

As seen in Figure 2, at the two test lengths as sample size increased, bias and RMSE values decreased for the *se(b)* estimates from all the programs. Considering the test length of 30 in Figure 2a and 2d, the

smallest RMSE and bias values for the *se(b)* parameter were obtained from BILOG-MG estimates at all the sample sizes. And Mplus and *R (ltm)* had similar but larger RMSE and bias values than BILOG-MG. According to results, we can say that at all sample sizes, BILOG-MG program performed best in estimating *se(b)* parameter. Similarly, at the test length of 60 and sample size of 500, again BILOG-MG had the smallest and R (ltm) had the largest RMSE and bias values for the *se(b)* parameter (see Figure 2b). At the sample size of 1000 and 2000, Mplus and R (ltm) had similar but larger RMSE and bias values than BILOG-MG program. However, at the sample size of 2000, the performance of three programs got very close to each other, BILOG-MG still estimated smaller RMSE and bias values for the *se(b)* parameter. In other words, we can say that BILOG-MG performed best in terms of estimating *se(b)* parameter at all the test lengths and sample sizes.

For each of six conditions, the average of RMSE and bias values for the "*a*" parameter over 50 replications are plotted in Figure 3.



a) RMSE for *a* parameter in test length 30

b) RMSE for *a* parameter in test length 60

c) bias for *a* parameter in test length 30

d) bias for *a* parameter in test length 60

Figure 3. RMSE and Bias Values for *a* Parameter

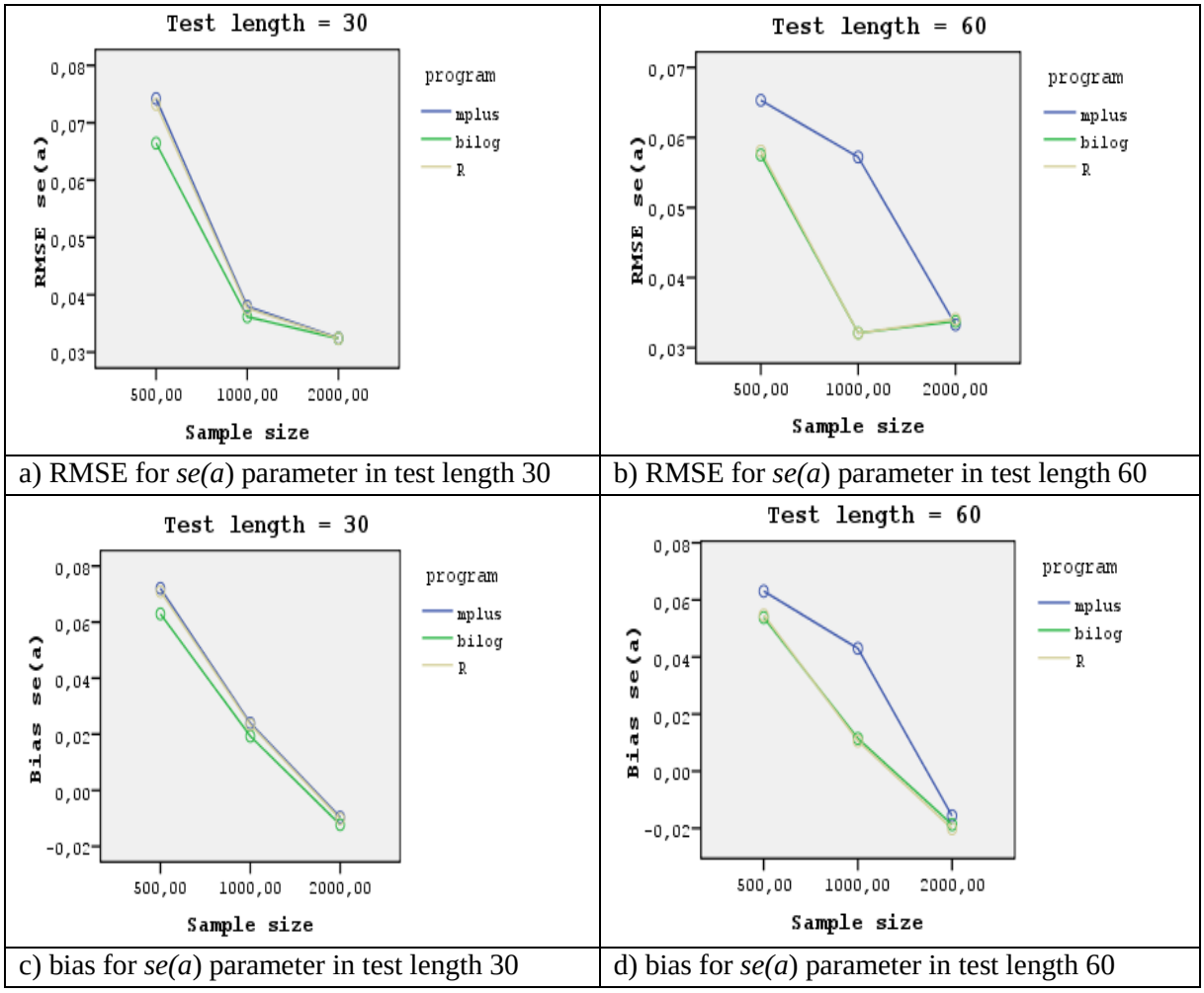
As shown in Figure 3a and 3b, when test lengths were 30 and 60, RMSE values of *a* parameter decreased as the sample size increased. This drop was sharper for Mplus and BILOG-MG programs when the number of item was 60. When test length was 30 and sample sizes were 500 and 1000, although BILOG-MG program had smaller RMSE values than other programs, at the test length of 2000, all of the three programs had similar RMSE values (see Figure 3a). We can say that while BILOG-MG had the best

performance at the sample size of 500 and 1000, at the sample size of 2000, all the programs performed similar in terms of estimating *a* parameter.

When test length increased to 60, programs performance changed due to sample size. For example, at the sample size of 500, Mplus and R (ltm) performed similar but they had larger RMSE values than BILOG-MG estimates. Under the condition where the sample size was 1000, the Mplus program had smallest and the R (ltm) had the largest RMSE values. At the sample size of 2000, while Mplus and BILOG-MG performed best, R (ltm) performed worst (see Figure 3b).

As shown in Figure 3c, for the test length 30, as sample sizes increased, bias values decreased in all programs except for Mplus. Also, Mplus had the smallest bias values and BILOG-MG was the largest bias values at all sample sizes. At the test length of 60, although BILOG-MG performed as well as Mplus program, generally Mplus had the smallest and R (ltm) had the largest bias values at all the sample sizes.

In Figure 4, the average of RMSE and bias values for the "*se(a)*" parameter over 50 replications are plotted.



a) RMSE for *se(a)* parameter in test length 30 | b) RMSE for *se(a)* parameter in test length 60

c) bias for *se(a)* parameter in test length 30 | d) bias for *se(a)* parameter in test length 60

Figure 4. RMSE and Bias Values for *se(a) P*arameter

As seen in Figure 4a, in all the programs, as sample size increased, RMSE values of *se(a)* parameter decreased in test length 30 conditions. At the sample size of 500, while BILOG-MG had the smallest RMSE values and it had the best performance, Mplus and R (ltm) had similar but larger RMSE values. When the sample size increased from 500 to 1000, RMSE values for *se(a)* were sharply decreased in all

the programs and although BILOG-MG estimates of *se(a)* had the smallest RMSE values, we can say that all of the three programs showed similar performance. And especially at the sample size of 2000, the performance of three programs is the same (see Figure 4a).

In conditions where test length was 60 and samples sizes were 500 and 1000, R (ltm) and BILOG-MG had smillar and smaller RMSE values than Mplus, but at the sample size of 2000, all the programs had similar RMSE values (see Figure 4b). Also we can say that as sample size increased from 500 to 1000, the RMSE values decreased in all programs. When sample size increased from1000 to 2000, RMSE values decreased for Mplus, but for BILOG-MG and R (ltm), RMSE values increased (see Figure 4b).

When we looked at the bias values in Figures 4c and 4d, we can see that at the test lengths of 30 and 60, as sample size increased, bias values for *se(a)* decreased in all the programs. At the test length of 30 and sample sizes of 500 and 1000, Mplus and R (ltm) programs had similar but larger bias values than BILOG-MG program but at the test length of 60 still Mplus had the largest bias values, BILOG-MG and R (ltm) had similar and smaller values than Mplus. On the other hand, at the sample size of 2000, for both of test lengths, we can say that all the programs had similar bias values for *se(a)* estimates.

According to Table 3 and Figure 4, when the number of items was 30, the RMSE values of *se (a)* decreased as the sample size increased in all the programs. When the sample size was 500, the smallest RMSE values were obtained by BILOG. All the programs showed similar performance when the sample size was 2000. When the number of item was 60, RMSE values of *se (a)* tended to decrease as the sample size increased. But when the sample size was 2000, the RMSE value of *se (a)* increased in BILOG and R (ltm) programs. The smallest RMSE values for *se (a)* were obtained in BILOG-MG and R (ltm). In all the three programs, while the number of items were 30 and 60, the bias values of *se (a)* decreased as the sample size increased. When test length was 30, the smallest bias values were obtained by BILOG-MG. When the number of items was 60, BILOG-MG and R (ltm) showed better and similar performance compared to Mplus.

DISCUSSION and CONCLUSION

The aim of this study was to investigate the effects of sample size and test length on parameter estimates and to compare the performance of Mplus, BILOG-MG and R (ltm) in terms of parameter estimation accuracy. The conclusions based on results can be listed as follows:

According overall results based on RMSE index, we can say that while Mplus was the best program in estimating *b* parameter, it was the worst program in estimating *se (a)* parameter. BILOG-MG was the best and R (ltm) was the less effective in estimating *se(b), a and se(a)* parameters. This result is consistent with the findings of Rahman and Chajewski (2014). The researchers compared the RMSE values for the parameter estimates obtained by BILOG, PARSCALE, IRTPRO, flexMIRT and ltm package in R software. They found that although the estimation results were within acceptable ranges, the R (ltm) showed the most erroneous estimation. With regard to bias index, Mplus was the best in estimating *b* and *a* parameters but it was the worst program in estimating *se(a)* parameter. On the other hand BILOG-MG was the best in estimating *se(a)* and *se(b)* parameters. Lastly, R (ltm) was the worst in estimating, *b, se(b)* and *a* parameters. Besides, Muthén (1999) noted that small differences between BILOG-MG and Mplus estimates can be ignored, because both programs use the ML estimation but BILOG uses the logit function (D=1.7) instead of the probit function.

In all test the lengths, as sample sizes increased, RMSE values decreased for all the parameter estimates. This finding supports the conclusion that the increasing sample size minimizes RMSE values for parameter estimation in the literature (Şahin & Anıl, 2017; Şahin & Colvin, 2015; Lord,1968; Ree & Jensen,1980). The consistency of the estimator increases as the sample size increases, and estimated parameters tend to approach to the true values (Thissen & Wainer, 1982). In addition, as the sample size increases, the standard errors of the sample decrease, therefore, RMSE values for parameter estimations can be reduced (Stone, 1992). As stated by Edelen and Reeve (2007), the standard errors of parameter estimations are also reduced as the sample size increases.

Based on RMSE index, at the test length of 30 and sample size of 500, BILOG-MG was the best performing program in estimating *b* parameter but as sample size increased to 1000 or to 2000, R (ltm) performed as well as BILOG-MG. According to Şahin & Colvin (2015), especially *b* parameters can be estimated most accurately by ltm for 1 PL, 2 PL and 3PL models. In our study, although the performance of Mplus was found to be closer to the other programs at sample size of 2000, generally it was the worst performing program in estimating *b* parameter. When test length increased to 60, at all of the sample sizes, R (ltm) was the less effective program in estimating *b* parameter and the performance of BILOG-MG and Mplus program was affected by the sample sizes. For example, while BILOG-MG performed better than Mplus at the sample size 500, Mplus performed better at sample size 1000 and both programs performed similar at the sample size of 2000.

In terms of bias index at the test length of 30, while Mplus was the best performing at sample sizes of 500 and 2000, R (ltm) was the best at sample size of 1000 and BILOG-MG was the low performing program in estimating *b* parameter. When test length was increased to 60, although the performance of BILOG-MG got very close to that of Mplus program at the sample size of 2000, Mplus was the best and R (ltm) was the worst performing program in estimating *b* parameter.

Another conclusion that can be drawn from this study according to RMSE and bias index for *se (b)* is that, BILOG-MG was the best performing program at all the test lengths and sample sizes. Although at the test length of 60, Mplus performed better than R (ltm) in some cases (i.e.at sample size 500), generally Mplus and R (ltm) showed similar performance. And another result is that as sample size increased, bias in estimating *se(b)* parameter decreased in all the programs. According to Toland (2008), the accuracy of the estimated *se(b)* in BILOG- MG is related to sample size for 2 PL model. He found that for sample size of 4000, consistent estimation of *se(b)* can be found throughout the range of difficulty parameters. But when sample size was 500, accuracy of *se(b)* decreased for larger *b* parameters in BILOG-MG. So he suggests that researchers can use BILOG-MG confidently for *se (b)* estimations in other applications with large sample sizes.

If we consider RMSE values for the *a* parameter, especially at the smallest sample sizes and for both test lengths, BILOG-MG was the best performing program. For the test length 30, at the sample sizes of 1000 and 2000, the performance of three the programs was very similar. At the test length of 60, although Mplus was the best performing program at sample size of 1000, BILOG-MG caught Mplus at sample size of 2000. Lastly, we can say that R (ltm) was the low performing program for test length 60.

In terms of bias values for *a* parameter, results showed that at the test length 30, Mplus was the best and BILOG-MG was the worst performed. At the test length 60, although BILOG-MG performed as well as Mplus program, generally Mplus performed best and R (ltm) performed the worst.

For *se(a)* parameter, based on RMSE index, at the test length 30, although R (ltm) and Mplus programs caught BILOG-MG's performance at sample sizes 1000 and 2000, generally BILOG-MG was the best. On the other hand, for the test length 60, although the three programs performed similar at the biggest sample size, BILOG-MG and R (ltm) performed similar and better than Mplus. According to Toland (2008), users of BILOG-MG can get reasonably accurate estimates of *se(a)* under the 2PL model for smaller values of *a* parameters (i.e., $a < 1.4$). These findings concur with the findings of the current study. This may be due to the fact that the true values of *a* parameter are less than 1.4 for only 4 items within 30 items and less than 1.4 for 13 items within 60 items.

In the previous studies, it is seen that RMSE values obtained for *a* parameter were between 0.11 and 0.15 and between 0.10 to 0.14 for *b* parameter. In this study, the RMSE values obtained from Mplus, BILOG-MG and R (ltm) were consistent with the previous studies, because they are in the same range as those obtained in previous studies (Gao & Chen, 2005; Kim, 2006; Yen, 1987). Therefore, it can be said that all the three programs can be used to estimate *a* and *b* parameters, because they predict *a* and *b* parameters close to their true values.

REFERENCES

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one, two and three parameter logistic models. *Applied Psychological Measurement, 11*(2), 111- 141.

Baker, F. B. (1990). Some observations on the metric of BILOG results. *Applied Psychological Measurement, 14*(2), 139–150. DOI: https://doi.org/10.1177/014662169001400203

Baker, F. B. (1998). An investigation of the item parameter recovery of a Gibbs sampling procedure. *Applied Psychological Measurement, 22*(2), 153–169. DOI: https://doi.org/10.1177/01466216980222005

Bulut, O. & Zopluoğlu, C. (2013). *Item parameter recovery of the graded response model using the R package ltm: A Monte Carlo simulation study*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. N.Y: CBS College Publishing Company.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*(1), 5–18. DOI: http://dx.doi.org/10.1007/s11136-007-9198-0

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-381. DOI: https://doi.org/10.1177/0013164498058003001

Foley, B. (2010). *Improving IRT parameter estimates with small sample sizes: Evaluating the efficacy of a new data augmentation technique.* Open Access Theses and Dissertations from the College of Education and Human Sciences. Paper 75

Gao, F. & Chen, L. (2005). Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three-parameter logistic model. *Applied Measurement in Education, 18*(4), 351-380.

Gübeş, N. Ö., Paek, I., & Cui, M. (2018). Örneklem büyüklüğünün ve test uzunluğunun MTK parametre kestirimine etkisi. *Pegem Atıf İndeksi*, 135-148.

Hambleton, R. K. (1989). *Principles and selected applications of item response theory*. In R. Linn (Ed.), Educational Measurment (3rd.ed., pp. 147-200). New York: Macmillan.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38–47. DOI: https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory principles and applications*. Boston: Kluwer-Nijhoff Publishing

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, Calif.: Sage Publications.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*(3), 249–260. http://dx.doi.org/10.1177/014662168200600301

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement, 43*(4), 355-381. DOI: https://doi.org/10.1111/j.1745-3984.2006.00021.x

Lim, R. G. & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item function. *Journal of Applied Psychology, 75*(2), 164–174.

Lord, F. M. (1968). An Analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*(2), 989-1020. DOI: https://doi.org/10.1002/j.2333-8504.1967.tb00987.x

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

Mislevy, R. J. & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*(1), 57-75.

Muthén, B. O. (1999). *IRT models in Mplus*. Retrieved from http://www.statmodel.com/discussion/messages/23/25.html

Muthén, L. K., & Muthén, B. O. (1998-2012). Mplus user's guide (8th ed.). Los Angeles, CA: Muthén & Muthén

Muthén, L. K., & Muthén, B. O. (2002). How To Use A Monte Carlo Study To Decide On Sample Size and Determine Power. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(4), 599-620.

Pan, T. (2012). *Comparison of four maximum likelihood methods in estimating the Rasch model*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, Canada.

Partchev, I. (2017). Package 'irtoys'. A collection of functions related to item response theory (IRT).

Patsula, L. N., & Gessaroli, M. E. (1995). *A comparison of item parameter estimates and ICCs produced with TESTGRAF and BILOG under different test lengths and sample sizes*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Proctor, T., Teo, K.-S., Hou & J., Hsieh (2005). *Comparison of Parameter Recovery in a 2 Parameter Logistic Item Response Model using MLE and Bayesian MCMC Methods.* Class project for 07P:148/22S:138 Bayesian Statistics,University of Iowa.

Rahman, N. & Chajewski, M. (2014). A Comparison and Validation of 2- and 3-PL IRT Calibrations in BILOG, PARSCALE, IRTPPRO, flexMIRT, and LTM (R). *National Council of Measurement in Education* at Philadephia.

Ree, M. J., & Jensen, H. E. (1980). Effects of sample size on linear equating of item characteristic curve parameters. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference*. Minneapolis, MN: University of Minnesota.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1–25.

Swaminathan, H. & Gifford, J. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss & R. D. Bock (Eds.), *New horizons in testing: latent trait test theory and computerized adaptive testing* (pp. 13–30). New York: Academic Press.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*(1), 1–16. DOI: http://dx.doi.org/10.1177/014662169201600101

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory and Practice*, 17(1), 321-335. DOI: 10.12738/estp.2017.1.0270

Şahin, F. & Colvin, K. (2015). *Evaluation of R package ltm with IRT dichotomous models*. NERA Conference Proceedings, 6.

Thissen, D. & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47*(4), 397–412. DOI: 10.1007/BF02293705

Toland, M. D. (2008). *Determining the accuracy of item parameter standart error of estimates in BILOG-MG3*. Doctoral dissertation. Available from ProQuest LLC (UMI Number 3317288)

Van der Linden, W. & Hambleton, R. K. (1997). *Handbook of modern item response theory*. Newyork: Springer-Verlag.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52*(2), 275–291. DOI: http://dx.doi.org/10.1007/BF02294241

Yen, W., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111-153). Westport, CT: Praeger Publishers.

Yoes, M. (1995). *An updated comparison of micro-computer based item parameter estimation procedures used with the 3-parameter IRT model*. Saint Paul, MN: Assessment Systems Corporation.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3: Item analysis and test scoring with binary logistic models*. Chicago, IL: Scientific Software.

İki Kategorili Puanlanan Maddelerde Madde Tepki Kuramına Dayalı Parametre Kestirimi: BILOG-MG, Mplus and R (ltm) Karşılaştırması

Giriş

Son yıllarda özellikle eğitim ve psikoloji alanlarında madde tepki kuramı (MTK) modellerinin kullanımı popülarite kazanmıştır (Foley, 2010). MTK'nın bireyin yeteneği ile maddeye verdiği yanıt arasındaki ilişkiyi modelleyebilme avantajı sunması klasik test kuramı (KTK) modellerine göre daha çok tercih edilmesini sağlamıştır (de Ayala, 2009; Hambleton, Swaminathan & Rogers, 1991; Yen & Fitzpatrick, 2006). KTK, bireyin testte verdiği doğru cevap sayısına odaklanmaktadır. Yani doğru cevap sayısı aynı olan iki birey sorunun zor ya da kolay olması dikkate alınmadan ölçülen özellik bakımından aynı puana sahip olmaktadır (Proctor, Teo, Hou & Hsieh, 2005). Oysa MTK, bireyin yeteneğine göre herhangi bir madde üzerinde göstereceği performansın olasılığı üzerine temellenmektedir ve madde parametrelerini gruptan bağımsız, yetenek parametrelerini ise maddeden bağımsız olarak kestirmektedir (Hambleton,

Swaminathan & Rogers, 1991). Bu nedenle MTK' ya dayalı madde ya da yetenek kestirimleri özellikle test geliştirme çalışmalarında adından sıklıkla söz ettirmektedir.

Test geliştirme çalışmalarında madde ve yetenek parametrelerini en doğru ve stabil şekilde kestirebilen modellerin ortaya konulması amaçlanmaktadır. Çünkü bireyin rapor edilen puanı, hakkında alınabilecek herhangi bir kararı etkileyebilmektedir. Bu nedenle araştırmacılar çeşitli koşullarda en doğru kestirim yapan modeli ortaya koymayı amaçlamaktadır (Rahman & Chajewski, 2014). Alan yazında MTK' ya dayalı test geliştirme çalışmalarında örneklem büyüklüğü ve test uzunluğunun parametre kestirimlerine olan etkisi sıklıkla araştırılan konu olarak ele alınmaktadır. MTK modelleri doğru parametre kestirimleri yapabilmek için büyük örneklemlere ihtiyaç duymaktadır (Hambleton, 1989; Hulin, Lissak & Drasgow, 1982). Her ne kadar minimum örneklem sayısı ve test uzunluğunun ne olması gerektiği konusunda kesin kurallar koyulamasa da (Foley, 2010) yapılan çalışmalar çeşitli koşullarda ulaşılması gereken örneklem sayısını ortaya koymaya yöneliktir (Lord, 1980; Patsula & Gessaroli, 1995; Yen, 1987; Yoes, 1995). Çalışmaların ortak noktası aslında örneklem sayısı ve test uzunluğunun özellikle karmaşık modellerde büyük olması gerektiği yönündedir.

Lord (1968) güçlük, ayırt edicilik ve şans parametrelerinin kestirildiği 3 parametreli lojistik modelde ayırt edicilik parametresini doğru kestirebilmek için en az 50 madde ve 1000 örneklem büyüklüğü gerektiğini belirtmiştir. Hulin ve diğerleri (1982) 200, 500, 1000 ve 2000 örneklem sayıları ile 15, 30 ve 60 sayıda maddeden oluşan test uzunluklarını dikkate alarak 2PL ve 3PL modele göre kestirimler yapmıştır. İki parametreli lojistik model için en az 500 örneklem ve 30 madde ayısına ihtiyaç duyulacağını belirtmiştir. Ayrıca 3PL model için örneklem sayısının 1000, madde sayısının ise 60 olmasını önermiştir. Ancak örneklem sayısı 2000, madde sayısı 30 olduğunda da çok benzer kestirim sonuçları elde etmiştir. Bu nedenle örneklem sayısının arttırılamadığı durumda madde sayısını artırmak bir yol olarak tercih edilebilmektedir.

Ancak, birçok test uygulamasında örneklem büyüklüğünü ya da test uzunluğunu arttırmak çok mümkün değildir. Bu nedenle çalışmalar artık örneklem büyüklüğü ya da test uzunluğuna göre en doğru modelin ve bilgisayar programının kullanımına yoğunlaşmaktadır. Baker (1987), parametre kestirimi ve kullanılan bilgisayar programının ayrılmaz bir bütün oluşturduğunu ve elde edilen madde parametre karakteristiklerinin programın altında yatan matematikten etkileneceğini belirtmiştir. Bu nedenle çeşitli zamanlarda teknolojinin sunduğu imkânlara bağlı olarak birçok bilgisayar programı kullanıma sunulmuştur. BILOG-MG (Zimowski, Muraki, Mislevy & Bock, 2003) iki kategorili maddelerde parametre kestirimi için yaygın bir şekilde kullanılan ve uzun geçmişe sahip olan programdır (Baker, 1990; Lim & Drasgow, 1990; Swaminathan & Gifford, 1983). Son zamanlarda MTK analizlerinin, açık kaynaklı program olan R programı (Rizopoulos, 2006, 2013) içerisindeki paketler (e.g. package ltm, irtoys) kullanılarak yürütüldüğüne rastlanmaktadır (Bulut & Zopluoğlu, 2013; Pan, 2012). R programı ücretsiz olduğu için yaygın şekilde kullanılmaktadır. Yine birçok analizi yapma imkânı sunan ve ücretli bir program olan Mplus (Muthén & Muthén, 1998-2012) son zamanlarda adından sıklıkla söz ettirmektedir ve örtük modelleri ortaya koymada tercih edilmektedir.

Bu bilgiler dikkate alındığında test uzunluğu ve örneklem büyüklüğüne ilişkin araştırmalara yer verilmesi gerektiği ve program türlerine göre elde edilen sonuçlarının karşılaştırılmasına ihtiyaç olduğu düşünülmektedir. Bu çalışma sözü geçen örneklem büyüklüğü ve test uzunluğu faktörlerinin MTK'nın 2PL modellerinde madde parametreleri ve madde kestirimlerine ait standart hata değerleri üzerine etkisinin araştırılması amacını taşımaktadır. Araştırmanın bir diğer amacı ise bu koşullar altında, alanyazında bu üçünün karşılaştırılmasına rastlanmadığı için, Mplus, BILOG ve R (ltm) programlarının parametre kestirimindeki performanslarını karşılaştırmaktır. Bu yönüyle ilgili araştırma MTK temel alınarak yapılan çalışmalarda yeterli örneklem büyüklüğünün ya da madde sayısının ne olması gerektiği konusundaki tartışmalara önemli katkıları olacağı düşünülmektedir. Öte yandan araştırmacılara eldeki verilere ya da kestirilecek parametrelere uygun olarak programlardan hangilerine ulaşmaları gerektiği konusunda fikir verebilecektir. Araştırma, parametrelere ilişkin standart hataları da karşılaştırmaya dâhil etmesi bakımından orijinallik özelliğini sağlamaktadır. Araştırmada simülasyon verileri kullanılmış ancak, veriler gerçek bir sınavdan kestirilen parametrelere uygun olarak üretilmiştir. Bu nedenle simülasyon sonuçları önceki çalışmalarla kıyaslanabilecek niteliktedir (Hulin ve diğerleri, 1982; Yen, 1987; Baker, 1998; Gao & Chen, 2005; Thissen & Wainer, 1982).

Tüm bunlar dikkate alındığında araştırmada ele alınan temel problem test uzunluğu ve örneklem büyüklüğü değiştiğinde parametre ve bunlara ait standart hata kestirimleri BILOG, Mplus ve R (ltm) programlarında nasıl değişmektedir? şeklinde belirlenmiştir.

Yöntem

Bu çalışmada kullanılan veriler R programında yetenek parametreleri aritmetik ortalaması 0, standart sapması 1 olan standart normal dağılım gösterecek şekilde üretilmiştir. TIMSS 2015 matematik uygulamasından hesaplanan madde parametreleri bu çalışmada verileri üretmek amacıyla kullanılmıştır.

Çalışmada örneklem büyüklüğü ve test uzunluğu simülasyon koşulları olarak ele alınmıştır. Örneklem büyüklüğü 500, 1000 ve 2000 test uzunluğu ise 30 ve 60 olacak şekilde 6 farklı koşul 50 tekrar yapılarak karşılaştırılmıştır. Bu çalışmada madde parametreleri 2 PL modele göre En Çok Olabilirlik Yöntemi (Maximum Likelihood Estimation-MLE) kestirim yöntemi kullanılarak elde edilmiştir. Veriler BILOG-MG, Mplus programlarında ve R programında irtoys paketinde ltm ile kestirilmiştir. Güçlük ve eğim (ayırt edicilik) parametreleri ve bunlara ait standart hataları (sh) karşılaştırmak amacıyla RMSE ve yanlılık indeksleri hesaplanmıştır.

Sonuç ve Tartışma

Bu araştırmanın amacı örneklem büyüklüğü ve test uzunluğunun parametre kestirimi üzerindeki etkisini incelemek ve Mplus, BILOG-MG ve R (ltm) programlarının parametre kestirimindeki performanslarını karşılaştırmaktır.

Araştırmadan elde edilen RMSE indeksleri dikkate alındığında Mplus programının *b* parametresini kestirmede en iyi, *sh (a)* parametresini kestirmede en düşük performansı sergilediği görülmüştür. BILOG-MG *sh(b), a* ve *sh(a)* parametrelerini en iyi kestiren program iken R (ltm) bu parametreleri kestirmede en düşük performansı sergilemiştir. Bu sonuç Rahman & Chajewski (2014)'ün bulgularıyla tutarlılık göstermektedir. Araştırmacılar BILOG, PARSCALE, IRTPRO, flexMIRT ve ltm (R) ile kestirdikleri parametrelere ilişkin RMSE değerlerini karşılaştırdıklarında kabul edilebilir derecede olsa da en hatalı kestirimin ltm programında olduğunu göstermişlerdir. Yanlılık indekslerine bakıldığında *b* ve *a* parametrelerini en yansız kestiren programın Mplus olduğu görülmüştür. Ancak bu program *sh (a*) parametresini en yanlı kestiren programdır. BILOG-MG programı *sh(a)* ve *sh(b)* parametresini en yansız kestiren program olmuştur. R (ltm) ise *b, sh(b)* ve *a* parametresini en yanlı kestiren programdır. Muthén'e (1999) göre, BILOG ve Mplus kestirimleri arasındaki küçük farklar göz ardı edilebilmektedir, çünkü her iki program da ML kestirim yöntemini, ancak BILOG programı probit fonksiyon yerine logit fonksiyonu (D=1.7) kullanmaktadır.

Araştırma bulguları tüm programlarda örneklem büyüklüğü arttıkça *a* ve *b* parametreleri ile bu parametrelerin standart hatalarına ilişkin kestirilen RMSE değerlerinin genel olarak düştüğünü göstermiştir. Bu bulgu alan yazında örneklem büyüklüğünün parametre kestirimine ilişkin RMSE değerlerini küçülttüğü sonucunu destekler niteliktedir (Şahin & Anıl, 2017; Şahin & Colvin, 2015; Lord,1968; Ree & Jensen,1980). Örneklem büyüklüğü arttıkça, kestiricinin tutarlılığı artmakta ve gerçek parametre değerine daha yakın kestirimler elde edilmektedir (Thissen & Wainer, 1982). Ayrıca, örneklem büyüklüğü arttıkça örneklem dağılımına ilişkin standart hatalar azalmakta dolayısıyla parametre kestirimlerine ilişkin RMSE değerleri azalmaktadır (Stone, 1992). Edelen & Reeve (2007)'nin de belirttiği gibi örneklem büyüklüğü arttıkça parametre kestirimlerine ait standart hatalar da küçülmektedir.

RMSE indekslerine göre test uzunluğu 30 ve örneklem büyüklüğü 500 olduğunda BILOG-MG programının *b* parametresini en iyi kestirdiği, ancak örneklem büyüklüğü 1000 ve 2000 olduğunda R (ltm) ile BILOG-MG'den daha iyi kestirimler elde edildiği görülmüştür. Şahin & Colvin (2015) de 1 PL, 2PL ve 3PL modellerde ltm paketinin *b* parametresini en doğru kestirdiğini belirtmiştir.

Bu çalışmada Mplus programının 2000 örneklem büyüklüğünde *b* parametresi için diğer programlara yakın kestirim sonuçları elde ettiği görülse de genel olarak *b* parametresini 30 madde sayısı ve 2000 örneklem büyüklüğünde en kötü kestirdiği sonucuna varılmıştır. Test uzunluğu 60 olduğunda tüm örneklem büyüklüklerinde R(ltm) *b* parametresini kestirmede en düşük performası sergilemiştir. BILOG-MG programı 500 örneklem büyüklüğünde Mplus'a göre *b* parametresini kestirmede daha iyi iken, 1000 örneklem büyüklüğünde Mplus programı BILOG-MG'ye göre daha iyidir. Örneklem büyüklüğü 2000 iken BILOG-MG ve Mplus benzer performans sergilemiştir.

Araştırmadan çıkan bir diğer sonuç *sh(b)* parametresini en iyi kestiren programın tüm örneklem büyüklüğü ve test uzunluklarında BILOG-MG olduğu yönündedir. Öte yandan örneklem büyüklüğü arttıkça *sh(b)* parametresine yönelik yanlılık indekslerinin tüm programlarda düştüğü görülmüştür. Toland (2008), *sh (b)* parametresinin BILOG-MG programında kestirim doğruluğunun 2 PL model için örneklem büyüklüğüne bağlı olduğunu belirtmiştir. Örneklem büyüklüğü 4000 olduğunda *sh(b)* için tutarlı sonuçlar elde ettiğini, ancak örneklem büyüklüğü 500 iken büyük *b* değerlerinde *sh(b)* parametresinin kestirim doğruluğunun azaldığını ifade etmiştir.

RMSE değerleri dikkate alınarak *a* parametresi incelendiğinde özellikle, küçük örneklemlerde 30 ve 60 madde sayısı koşullarında BILOG-MG programının en iyi performans sergilediği görülmüştür. Madde sayısı 30, örneklem büyüklükleri 1000 ve 2000 iken tüm programların perfromansı benzerdir. Madde sayısı 60 iken, örneklem büyüklüğü 1000 olduğunda Mplus en iyi kestirimi yaparken, 2000 örneklem büyüklüğünde BILOG-MG ve Mplus benzer performans göstermiştir. R (ltm) ise test uzunluğu 60 olduğunda en düşük performansı sergilemiştir.

a parametresi için yanlılık değerlerine bakıldığında test uzunluğu 30 olduğunda Mplus programının en iyi, BILOG-MG programının kestirim doğruluğunun en kötü olduğu görülmüştür. Ancak madde sayısı 60'a çıkarıldığında BILOG-MG, Mplus kadar iyi yansız kestirim yapabilmektedir. R (ltm) ise en yanlı kestirim sonuçlarına sahiptir. *sh (a)* parametresi için RMSE değerlerine bakıldığında test uzunluğu 30, örneklem büyüklükleri 1000 ve 2000 iken BILOG-MG en iyi performansı gösterirken, Mplus ve R(ltm)'nin performansları BILOG-MG'ye yakındır. Öte yandan test uzunluğu 60 ve örneklem sayısı büyük olduğunda BILOG-MG ve R (ltm) hem benzer hem de Mplus'tan daha doğru kestirim yapmaktadır. Toland (2008), BILOG-MG kullanıcılarının 2 PL modelde *a* parametresinin küçük değerleri ($a<1.4$) için *sh(a)*'nın kestirimine güvenebileceklerini belirtmiştir. Bu çalışmada elde ettiğimiz sonucun ilgili çalışma ile tutarlı olması, çalışmamızda *a* parametresinin gerçek değerlerinin genel olarak 30 madde içerisinde yalnızca 4 tanesinde ve 60 madde içerisinde 13 tanesinde 1.4 değerinden küçük olmasından kaynaklanıyor olabileceğini akla getirmektedir.

Daha önce yapılmış çalışmalarda (Gao & Chen, 2005; Kim, 2006; Yen, 1987), *a* parametresi için elde edilen RMSE değerlerinin 0.11 ile 0.15 arasında, *b* parametresi için 0.10 ile 0.14 arasında değiştiği belirtilmiştir. Bu çalışmada Mplus, BILOG-MG ve R (ltm) ile elde edilen RMSE değerleri yapılan çalışmalarla benzer aralıktadır. Dolayısıyla her üç programın da *a* ve *b* parametrelerini gerçek değere yakın kestirebilmesi nedeni ile kullanılabileceği önerilebilir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(1); 43-60

The Importance of Sample Weights and Plausible Values in Large-Scale Assessments

Serkan ARIKAN * Ferah ÖZER ** Vuslat ŞEKER *** Güneş ERTAŞ ****

Abstract
International large-scale assessments such as PISA (The Programme for International Student Assessment), PIAAC (The Programme for the International Assessment of Adult Competencies) and TIMSS (Trends in International Mathematics Science Study), play a key role in determining educational policies besides their primary objectives of measuring, evaluating and monitoring the educational process. Therefore, it is critical to analyze the data gathered from the large scale assessments using scientifically accurate statistical methods as the results have the potential to influence millions of stakeholders through major policy changes. Analysis of these data that consists of hundreds of different genuine variables requires expertise and using specific methods. This study illustrates issues to be considered while analyzing PISA, PIAAC and TIMSS data by presenting relevant syntax and exemplifying the possible incorrect results that might be encountered. In Turkey, there are very limited courses that focus on large scale data analysis. Workshops are also very limited to reach major groups. The aim of this study is to raise awareness related to sample weights and plausible values. Comparative findings of the study showed that without using sample weights and plausible values there is a high probability to get incorrect results. In this study, t-test and multiple regression analyses conducted by IDB Analyzer and multilevel regression analysis by Mplus were exemplified.

Keywords: Sample weights, plausible values, large scale assessment, IDB Analyzer, Mplus

INTRODUCTION

International large-scale assessments such as PISA (The Programme for International Student Assessment), PIAAC (The Programme for the International Assessment of Adult Competencies) and TIMSS (Trends in International Mathematics Science Study), play a key role in determining educational policies besides their primary objectives of measuring, evaluating and monitoring the educational process (Bialecki, Jakubowski, & Wisniewski, 2017; Figazzolo 2009; Novoa & Yariv-Mashal, 2003; Steiner-Khamsi & Waldow, 2018). In the early periods of these assessments, the developers highly emphasized that the aim of the assessment was mainly monitoring the process rather than cross-country comparisons (Landahl, 2018). Yet, in the following periods, cross-country comparisons raised the interest of both local and international media, which led the test results to be used as also for indicators of economic growth and rationales for policy reforms. Moreover, Addey, Sellar, Steiner-Khamsi, Lingard and Verger (2017) explained the reasons for participation of the countries to these tests as follows: to provide data-based information for policies, technical capacity-infrastructure building, to provide financial support and assistance, prominence in international relations, decision making in domestic politics, economic reasons, reforms to curriculum and teaching

* Asst. Prof, Boğaziçi University, Faculty of Education, İstanbul, serkan.arikan1@boun.edu.tr, ORCID ID: 0000-0001-9610-5496.
** Res. Assist., Boğaziçi University, İstanbul, Faculty of Education, ferah.ozer@boun.edu.tr, ORCID ID: 0000-0001-8621-3522.
*** Res. Assist., Boğaziçi University, İstanbul, Faculty of Education, vuslat.seker@boun.edu.tr, ORCID ID: 0000-0002-3279-5544.
**** Res. Assist., Boğaziçi University, İstanbul, Faculty of Education, gunes.ertas@boun.edu.tr, ORCID ID: 0000-0001-8785-7768.

To cite this article:
Arıkan, S., Özer, F., Şeker, V. & Ertaş, G. (2020). The Importance of Sample Weights and Plausible Values in Large-Scale Assessments. *Journal of Measurement and Evaluation in Education and Psychology, 11*(1), 43-60. doi: 10.21031/epod.602765

Received: 06.08.2019
Accepted: 14.02.2020

approaches. In addition to those, international organizations such as OECD (Organisation for Economic Co-operation and Development), UNESCO, World Bank utilize these assessment results to monitor educational policy reforms in countries and to determine for further investments/grants for developing countries (Addey & Sellar, 2018; Aydın, Selvitopu, & Kaya, 2018). In summary, to date, large-scale assessment data provide crucial information for the efficiency of countries' educational system elements and comparable data about the current student, teacher, and administrator profiles.

Regarding the main reason for the participation of the countries to large scale assessments (Adler, 2017), it is known that in recent years the data-driven results gathered from PISA, PIAAC, and TIMSS have been used for some major and minor educational policy reforms in different countries. In some cases, these major reforms include curricular changes, orientation and the integration of disadvantaged groups; whereas minor reforms include changes in textbooks, educational materials, integration of educational hardware-software and local school cultures. Specifically, it is known that France (Carvalho & Costa, 2015; Michel, 2017), Portugal (Carvalho & Costa, 2015), Poland (Bialecki et al., 2017), Hungary (Carvalho & Costa, 2015), Germany (Ertl, 2006), Sweden (Landahl, 2018), Israel (Pizmony-Levy, 2018) and Spain (Tiana Ferrer, 2017) utilized these source of data to legitimize recent radical policy reforms or curricular changes that were carried out by the different governmental institutions (Wiseman, 2013). Similarly, in Turkey major curricular reforms and changes on the national high-stakes exams have been made since the beginning of the 2000s. Especially in the curriculum changes of 2013 and 2018, the importance of providing learning environments and opportunities that promote higher cognitive skill development, such as analyzing, reasoning, and evaluating has been highly emphasized as an influence of PISA and TIMSS. In line with these policy changes, high-stake central exams were also affected by these major structural changes. For instance, High School Entrance Exam (LGS) has started to measure higher-order thinking skills along with subject matter knowledge (MEB, 2018). Indeed, the so-called national version of PISA administration, namely ABİDE, which aims to measure higher-order thinking skills such as critical thinking, problem-solving and interpretation could also be considered as one of the exemplary initiatives for recent reforms regarding PISA & TIMSS alignment.

Factors such as increased number of large scale assessments-related publications on local and international media (Martens & Niemann, 2010) and elicited media perception related to PISA (Michel, 2017) led the raised awareness on the public **(**Froese-Germain, 2010; Gür, Çelik & Özoğlu, 2012; Steiner-Khamsi & Waldow, 2018). In line with these factors, easy accessibility of the data, serving as a promising field to use the contemporary analysis methods, and providing opportunities for cross-cultural and cross-country comparisons also led the educators and researchers to study on this matter profoundly, which grounded for many national and international publications. In this vein, it is clear that data obtained from large scale assessments have a crucial mission to affect further educational policies. Considering crucial role and mission of large scale assessments, it is critical to analyze these data using accurate statistical methods. Analysis of these data that consists of hundreds of different genuine variables requires expertise. This study illustrates issues to be considered while analyzing PISA, PIAAC and TIMSS data by presenting relevant syntax and exemplifies the possible incorrect results that might be encountered when these issues are not taken into account. In this way, it is aimed to guide researchers studying large scale assessment data to use proper methods.

Large-Scale Tests

There are variety of large-scale assessments and the most widely used ones are PISA, PIAAC, and TIMSS. In the following sections, these assessments are briefly introduced.

Programme for international student assessment (PISA)

PISA is a program organized by the OECD in every three years since 2000 to measure 15-year-old students' performance on mathematics, science, and reading. PISA aims to reveal to what extent students have knowledge and skills needed for modern societies after they complete compulsory education (MEB, 2016a; OECD, 2018). There are three main subject areas in PISA: reading, mathematical literacy, and scientific literacy. PISA measures the degree to which students make use of their learning in these areas in different contexts. While PISA examined reading ability in more detail in 2000, 2009 and 2018, it focused on mathematics literacy in 2003 and 2012, and scientific literacy in 2006 and 2015. In addition, the program collects data from students, teachers, principals, and parents via questionnaires. In the latest PISA carried out in 2018, there were 76 member or nonmember countries. Turkey has been participating in PISA consistently since 2000.

Programme for the international assessment of adult competencies (PIAAC)

PIAAC aims to evaluate the key information processing skills needed for individuals aged 16-65 to participate in social life. The Survey of Adult Skills, as a product of the programme, aims to assess the adults' proficiency by focusing on three key information processing skills: literacy, numeracy, and problem-solving. It is assumed that adults who are proficient in those skills will be able to get benefit from the opportunities generated by technological and structural changes in modern societies (OECD, 2016). In addition to the survey of adult skills, PIAAC includes a comprehensive survey of participants' information related to socio-demographic characteristics. PIAAC was first implemented in 2011-2012 with the participation of 24 countries and on the second round in 2014-2015 with the participation of 9 more countries, the total number of participant countries had reached to 33. Turkey was among those 9 countries that participated on the second round of the study. According to the results of the report *Skills Matter: Further Results from the Survey of Adult Skills* published in 2016, Turkey was significantly below the OECD average (OECD, 2016; TEDMEM, 2016).

Trends in international mathematics science study (TIMSS)

TIMSS is an international study to evaluate the skills and knowledge gained in mathematics and science fields for the 4^{th} and 8^{th} grade students (MEB, 2016b; Mullis & Martin, 2017). TIMSS has been co-developed and administrated by Boston College and International Association for the Evaluation of Educational Achievement (IEA). Since its inauguration in 1995, the test was administrated in 1999, 2003, 2007, 2011, 2015 and 2019 consecutively every 4 years, with the increased number of participating countries in every year. Moreover, the expected number of countries for 2019 administration is likely to be 70 (Mullis & Martin, 2017). Turkey has been included in the TIMSS study in 1999, 2007, 2011, 2015 and 2019 (MEB, 2016b).

TIMSS generally focuses on curricular objective frameworks to evaluate the skills and knowledge gained in mathematics and science fields. Thus, TIMSS curriculum framework is basically three folded as follows: *intended curriculum* in national, social and educational contexts, *implemented curriculum* at home, school, teacher and classroom contexts; *attained curriculum* in student achievement and attitudes contexts. Within these contexts, the TIMSS evaluation framework basically consists of *subject matter dimension*, that focuses on the subject matter knowledge level and *cognitive dimension* that focuses on thinking processes. By providing detailed data among countries' mathematics and science curricula, TIMSS presents the opportunity to make cross-country comparisons as well as local comparisons (MEB, 2016b)

The Important Features of Large Scale Assessment Datasets

There are two important features of large scale assessment (LSA) datasets. The first one is the sample weights which are related to the sampling design of LSA's. The second one is the plausible values related to rotated test design used in the test administration (Rutkowski, Gonzalez, Joncas, & von Davier, 2010). The following section explains these concepts.

Sampling weights

Large scale assessments aim to choose the most representative sample generalizable to the population since it is not possible to use the entire population due to financial inadequacy and time limitations. The sample is useful the extent to which it estimates the characteristics of the population. The most common technique for clarifying the issue of differences between the distribution of characteristics in the sample and in the population is using sampling weights (Rust, 2013). In PISA 2015 technical report, the necessity of using sampling weights was highlighted as to ensure each student in the sample was represented with the correct number of students in the population (OECD, 2017). Sampling weights are used in studies that TÜİK (Turkey Statistics Institution) conducted at the national level and international large scale tests (PISA, TIMSS, & PIAAC, etc.).

In PISA and TIMSS, multistage sampling design is used for sample selection. The use of a multistage design has a significant impact on the precision of resulting estimates (Rust, 2013). In the first stage, schools are selected proportional to their size; and in the second stage classes and/or students are randomly selected from the selected school (LaRoche & Foy, 2016; OECD, 2017). The size of the school is determined by the number of students eligible to participate in the study. For instance, the number of students aged 15 in PISA and the number of students enrolled in 4^{th} or 8^{th} grade in TIMSS are considered to calculate the school size. In PIAAC, all non-institutionalized adults between the ages of 16 and 65 are considered.

Random sampling design is implemented in order to ensure that the sample selection is not biased and that each individual has an equal chance to be selected. Non-random sample designs may cause the bias, whether intentionally or unintentionally. In random sampling also, each individual's chance for selection may not always be equal in the population. In this case, sample weights are used to avoid the bias and to ensure the representativeness of all individuals in the population. A sample unit is determined according to the probability of selection of each individual in the sample. Sample weights are defined as the inverse of the probability of selection for the unit. In other words, if a group has a very low chance to be selected to the sample, the sample unit for the individual representing that group will be higher than the sample unit for the individual coming from the group having high chance to be selected (OECD, 2017). In the analysis, when the sample weights are taken into account for the mean scores of groups, the representation of the population is guaranteed and the estimations are precise. While analyzing the sample data, if the sample weights are used then the contribution of each student to statistical estimations will be proportional to the number of students represented in the population (Gonzales, 2012). Suppose that each individual has an equal chance to be selected among 300. Then, the probability of being selected among 30 individuals will be 1/10 and the weight of each individual will be 10. In this example, since the chance to be selected for each individual is equal, weights for each are also equal. The weights of 30 individuals add up to 300, the total number of individuals in the population. In this case, the weighted mean and the unweighted mean will be equal. For instance, suppose that a sample of 6 students is chosen from a population of 15 girls and 30 boys in a 45-student class. 3 boys and 3 girls are chosen for the sample. While boys are represented more than girls in the population, they are equally represented in the sample. The probability of selection of each 3 girls among 15 girls will be 3/15 = 0.2 and the probability of selection of 3 boys among 30 boys will be

3/30 = 0.1. According to this situation, the weight of each girl in the sample is 5 and the weight of each boy in the sample is 10. Let assume that girls took 8, 7, 7 points from the exam over 10 and boys took 5, 5, 4 points. In this case, while unweighted mean of the sample is [(8 + 7 + 7) + (5 + 5 + 4)]/6 = 6, the weighted mean of the sample which is [(8x5 + 7x5 + 7x5) + (5x10 + 5x10 + 4x10)]/45 = 5.56. Therefore, the weighted mean is 7 % lower than the unweighted mean. In the simplest way, as it is shown in the example, analysis without considering weights would mislead the estimations related to the population.

In multistage sample selection design, in an application that firstly schools are selected and then students are chosen from that school, school weight, within school weight and student weight are determined separately. For example, let the probability of selecting school *j* to be p_j and the probability of selecting students *i* at school *j* (under the condition of school *j* was selected) to be p_{ij}. Then the within school weight is $w_{ij}= 1/p_{ij}$ and the school weight is $w_j= 1/p_j$. In a population of 400 students from 10 different schools having 40 students, firstly 4 schools are randomly selected. Then, 10 students are chosen from each of those schools. The total number of students in the sample is 40. In this case, the probability of selection for each school (4 schools are selected from 10) is p_j = 4/10 = 0.4 and so the school weight is w_j= 2.5. The probability of selection for each student among 4 selected schools (10 students are chosen among 40 in each school) is p_{ij}= 10/40 = 0.25 and within school weight is w_{ij}= 4. Finally, in the case that firstly school is selected and the students are chosen within the school, the probability of selection for a student is $p*_{ij}= p_j$ x p_{ij}= 0.4 x 0.25 = 0.10 and the student weight is $w*_{ij}$= 10.

Since the data gathered from large scale assessments like PISA, PIAAC and TIMSS used multistage sampling, the methods and software that take into account sample weights must be used for all data analysis. The student weights in these data sets are W_FSTUWT (Final trimmed nonresponse adjusted student weight) in PISA, SPFWT0 (Final full sample weight) in PIAAC and TOTWGT (Total student weight) in TIMSS. In multilevel analysis, it is necessary to decompose these weights (Rutkowski et al., 2010). It is important to be aware that the results obtained without considering sample weights will be inaccurate (LaRoche & Foy, 2016; OECD, 2017; Rutkowski et al., 2010). Rutkowski et al. (2010) calculated that the mathematics mean score of Bulgaria as 463.63 when the sample weights were accurately used and 481.38 when sample weights were not used.

Plausible values

The large scale assessments like PISA and TIMSS aim to estimate the performance of population or subgroups in the populations instead of assessing the scores of individuals (Monseur & Adams, 2009; Von Davier, Gonzalez, & Mislevy, 2009). Calculating consistent and valid scores for individuals is not the purpose of large scale assessments. Therefore, the aim is to minimize the errors in population-level estimations (OECD, 2017). Furthermore, the rotated booklet design is used in order to minimize the test burden on individuals (Rutkowski et al., 2010). Students answer only certain parts of the test. However, as a group, student groups answer all of the questions. Therefore, student performance on large scale assessments is reported as plausible values (PVs).

Plausible value method accepts student ability as missing values (Rutkowski et al., 2010). The student ability distributions are estimated through Rubin's (1987) multiple imputation method. Within the distributions, random selections are made and these multiple assigned values are called plausible values (Rutkowski et al., 2010). Plausible values are precedent values for unobservable latent values

(Wu, 2005). Each student has an unobservable latent ability variable and multiple values are assigned to the variable (Laukaityte & Wiberg, 2017; Wu, 2005). OECD (2017) defines plausible values as randomly assigned numbers for individuals from the distribution of scores. The distribution is called marginal posterior distribution. Plausible values including random error variance components should not be considered as test scores, they should be used as defining population performance (OECD, 2017). In short, multiple values are assigned to each individual in order to minimize measurement error (Laukaityte & Wiberg, 2017). If the measurement error is small, multiple values assigned to an individual would be close to each other. On the contrary, if measurement error is large, multiple values assigned to an individual would be far from each other (Wu, 2005). Inferences from large scale assessments become more valid thanks to assigned plausible values and the results of the assessments contribute to the practice more productively (Laukaityte & Wiberg, 2017).

Five plausible values are used in many large scale assessment databases like PISA and TIMSS (OECD, 2017; Laukaityte & Wiberg, 2017). PISA started to report 10 plausible values since 2015. In PIAAC, 10 plausible values are reported. In the National Assessment of Educational Assessment (NAEP) database, 20 plausible values are used. The simulation studies conducted by Laukaityte and Wiberg (2017) showed that using multiple plausible values increases the accuracy of the estimation and decreases measurement error.

It is necessary to use methods and software that take into account plausible values in large scale assessments like PISA, PIAAC, and TIMSS. The researchers should be aware that the outcomes ignoring plausible values would be erroneous (LaRoche & Foy, 2015; OECD, 2017, Rutkowski et al., 2010).

Incorrect Data Analysis Approaches related to Large Scale Assessment Analysis

Rutkowski et al. (2010) listed two common incorrect data analysis approaches when LSA data is used. The first incorrect approach is to use only one of the plausible values. The second one is to take the average of all plausible values. For example, for TIMSS dataset, using only PV1 or averaging PV1 to PV5 are among these common incorrect data analysis approaches. Rutkowski et al. (2010) also added that taking the averages of plausible values creates more severe problems than taking only one plausible value. Therefore, they warned researchers not to use averages of plausible values. In the use of both incorrect approaches, standard errors will be estimated erroneously and p values will be affected. In addition to these aforementioned incorrect approaches, using plausible values as an indicator of a latent variable (such as math performance) in a structural equation model is another incorrect approach. In Turkey, there are studies that used correct approaches as well as incorrect approaches.

Present Study

The main purpose of this study is to raise awareness about LSA data analysis by explaining the structure and showing exemplary analysis. To fulfil this purpose 3 main research questions including group comparison with t-test, multiple linear regression, and multilevel regression were selected. The syntaxes of each analysis related to research questions were also provided in the appendices A-D. The research questions (RQs) of the study are as follows:

1) What are the effects of not taking into account the sample weights and plausible values in group comparison?
2) What are the effects of not taking into account the sample weights and plausible values in multiple regression?

3) Which procedures are used to take into account the sample weights and plausible values in multilevel regression?

To answer these research questions, the following sub-research questions were generated. For the RQ1, "Is there a statistically significant difference between mean TIMSS 2015 mathematics scores of boys and girls in Turkey?" and "Is there a statistically significant difference between mean PIAAC 2015 reading scores of adults who looked for a job last month and who did not look for a job last month in Turkey?"; for the RQ2, "Do disciplinary climate in science classes, epistemological beliefs, index of economic, social and cultural status, inquiry-based science teaching and learning practices, instrumental motivation, enjoyment of science, science self-efficacy, teacher-directed science instruction, teacher support in science classes predict PISA 2015 science performance of students in Turkey?"; for the RQ3 "Do student-level variables, parents make sure that time is allocated for the homework, parents check if the homework is completed, time spent on the homework; and teacher level variables, correcting assignments and giving feedback, letting students to correct their own homework, discussing homework in the classroom, monitoring completeness of the homework, using homework for grading predict TIMSS 2011 reasoning score of students in Turkey?" were used.

METHOD

Sample

In this study, PISA, PIAAC, and TIMSS datasets were used to introduce different LSA data. The sample used in the study is described in this section. In PISA 2015 dataset, there were 5895 students located in 187 schools from Turkey. The majority of students were 10th graders (MEB, 2016a). In PIAAC 2015 Turkey dataset, there were 5227 adults ranging from 16 to 65 years old (OECD, 2016). In TIMSS 2015 dataset, there were 6928 8th grade students located in 239 schools from Turkey (MEB, 2014). In TIMSS 2015 dataset, there were 6079 8th grade students located in 238 schools from Turkey (MEB, 2016b).

Instrument

PISA, PIAAC, and TIMSS have both tests to measure achievement or performance level and questionnaires to collect demographic and attitudinal data of participants. The first research question had two sub-research questions. The first sub-research question was related to the TIMSS 2015 dataset. Mathematics achievement in TIMSS was reported with 5 plausible values (BSMMAT01-BSMMAT05). Mathematics achievement was estimated using item response theory (IRT). The other variable of the research question, gender was taken from the questionnaire data (BSBG01). In the second sub-research question, PIAAC 2015 reading scores of the adults and whether they looked for a paid job was used as variables. Reading scores of adults were reported with 10 plausible values (PVLIT1- PVLIT10). The reading ability of the adults was estimated using IRT. The status of looking for paid job information (yes or no) was gathered from the adult questionnaire (C_Q02b).

In the second research question, the independent variables used in the model were disciplinary climate in science classes, epistemological beliefs, index of economic, social and cultural status, inquiry-based science teaching and learning practices, instrumental motivation, enjoyment of science, science self-efficacy, teacher-directed science instruction, teacher support in a science classes of PISA 2015 (DISCLISCI, EPIST, ESCS, IBTEACH, INSTSCIE, JOYSCIE, SCIEEFF, TDTEACH, TEACHSUP). These student-level independent variables are index scores of related questionnaire items. The science performance score was reported as 10 plausible values estimated by IRT (PV1SCIE-PV10SCIE).

In the last research question, TIMSS 2011 variables that were in the hierarchical structure, students nested in the classrooms, were used. Student level variables were parents make sure that time is allocated for the homework, parents check if the homework is completed, time spent on the homework (BSBG11C, BSBG11D, BSBM20B); and teacher level variables were correcting assignments and giving feedback, letting students to correct their own homework, discussing homework in the classroom, monitoring completeness of the homework and using homework for grading (BTBM25CA, BTBM25CB, BTBM25CC, BTBM25CD, BTBM25CE). The dependent variable, reasoning ability of the students, were estimated using IRT with 5 plausible values (BSMREA01-BSMREA05).

Data Analysis

In this section, how the analyses were performed and important concepts related to LSA data analysis were explained. The first research question was group comparison analysis. In both sub-research questions t-test was conducted as the grouping variables contained two categories. As explained in the introduction, LSA data analysis requires taking into account sample weights and plausible values. IEA's IDB Analyzer can conduct t-test by taking into account sample weights and plausible values (IEA, 2019). IDB Analyzer is an interphase program that can read SPSS files. In the first step, necessary variables including plausible values are selected. In the next step, the sample weight is selected. After these steps, IDB Analyzer produces an SPSS syntax and running the syntax produces the output. IDB Analyzer output does not give significance value (p-value), however, it reports t values. Using t value and the degrees of freedom, statistical significance can be decided. All of these values are reported in "*_sig.sav" output files. In the research question related to TIMSS, Total Student Weight (TOTWGT) was used. In the research question related to PIAAC, Final Full Sample Weight (SPFWT0) was used.

In the second research question, multiple linear regression was used as there were more than one independent variable to predict one dependent variable. IDB Analyzer also can conduct multiple regression by taking into account sample weights and plausible values. In PISA 2015, FINAL TRIMMED NONRESPONSE ADJUSTED STUDENT WEIGHT (W_FSTUWT) was used as a sample weight.

In the last research question, multilevel regression analysis was conducted as the research question contained student-level variables, as well as teacher-level variables. Mplus program was used as Mplus not only can take into account sample weights and plausible values but also multilevel structure of the variables (Muthen & Muthen, 2015). In order to take into account the sample weights, sample weights should be defined in the Mplus syntax. As Rutkowski et al. (2010) advised for multilevel analysis, sample weights were decomposed manually. For level 1 sample weights, the product of WGTADJ2*WGTFAC2*WGTADJ3*WGTFAC3 was used (CLASS WEIGHT ADJUSTMENT* CLASS WEIGHT FACTOR* STUDENT WEIGHT ADJUSTMENT* STUDENT WEIGHT FACTOR). For level 2 sample weights, the product of WGTADJ1* WGTFAC1 (SCHOOL WEIGHT ADJUSTMENT* SCHOOL WEIGHT FACTOR) was used. The product of level1 and level2 sample weights is equal to TOTAL STUDENT WEIGHT. Mplus requires creating separate text files that include one of the plausible values and the rest of the variables. For instance, if there are 5 plausible values, 5 text files that include one of the plausible values as one column and the rest of the variables in the other columns need to be created. Then the names of these text files are listed in a different text file which is the main input file and it is defined in MPLUS syntax (FILE = dataimputedlist.dat;). Also, the data structure should be stated in the syntax (TYPE = IMPUTATION;). Then, the relationships among variables should be defined.

RESULTS

This study aims to compare LSA data analysis with and without taking into account sample weights and plausible values. Also, it is aimed to guide researchers by showing LSA data analysis by providing syntaxes. The results of four main research questions were reported in the following sections comparatively.

Group Comparison Studies

In this section, two sub-research questions were analyzed. The first one is "Is there a statistically significant difference between mean TIMSS 2015 mathematics scores of boys and girls in Turkey?". t-test was conducted as the grouping variable, gender, contained two categories, boys and girls. With and without taking into account sample weights and plausible values were reported in Table 1.

When sample weights and plausible values were used, it was concluded that there was no statistically significant difference between mean TIMSS 2015 mathematics scores of boys and girls in Turkey (t=1.79, p>.05). This result is also the same as the TIMSS 2015 National Mathematics and Science Pre-Report (MEB, 2016b).

Table 1 also includes the results when each plausible value or the average of the plausible values were used. In all cases, there were statistically significant differences between mean TIMSS 2015 mathematics scores of boys and girls in Turkey. These findings totally contradict with the previous finding. Therefore, when sample weights and plausible values are not used, it is highly probable to obtain incorrect results.

Table 1. Comparison of Mathematics Scores of Girls and Boys

Method	Girls (SE)	Boys (SE)	Mean Difference (SE)	*t*
IDB Analyzer PV1-PV5	461.14 (4.80)	454.73 (5.31)	6.40 (3.57)	1.79
SPSS PV1	459.23 (1.90)	452.77 (1.86)	6.46 (2.66)	2.43*
SPSS PV2	460.50 (1.91)	452.87 (1.87)	7.63 (2.67)	2.85**
SPSS PV3	460.26 (1.91)	451.33 (1.91)	8.93 (2.70)	3.31**
SPSS PV4	458.04 (1.97)	449.01 (1.94)	9.03 (2.77)	3.26**
SPSS PV5	459.37 (1.94)	453.84 (1.90)	5.53 (2.72)	2.04*
SPSS PVmean	459.48 (1.87)	451.97 (1.83)	7.51 (2.62)	2.87**

*p < .05. **p < .01. ***p < .001. SE: Standard Error

In the second sub-research question, PIAAC dataset was used. The research question is "Is there a statistically significant difference between mean PIAAC 2015 reading scores of adults who looked for a job last month and who did not look for a job last month in Turkey?"

When sample weights and plausible values were used, it was concluded that there was no statistically significant difference between mean PIAAC 2015 reading scores of adults who looked for a job last month and who did not look for a job last month in Turkey (t=1.16, p>.05).

Table 2 also includes the results when each plausible value or the average of the plausible values was used. Among 11 cases, there were contradictory results. In 3 of these results, significant differences were found and in 8 of them, no difference was found. As similar to the first sub-research question, when sample weights and plausible values are not used, it is probable to obtain incorrect results.

In both sub-research questions, the difference in findings stems from standard errors. The standard errors were higher when sample weights and plausible values were taken into consideration than when they were not used. The change in the standard error directly affects the t value and the ultimate decision.

Tablo 2. Comparison of Reading Scores of Adults Who Looked For a Job and Not

Method	Looked for a job (SE)	Did not look for a job (SE)	Mean difference (SE)	*t*
IDB Analyzer PV1-PV10	226.11 (4.16)	221.05 (1.45)	5.06 (4.36)	1.16
SPSS PV1	229.06 (2.51)	223.90 (.83)	5.16 (2.73)	1.89
SPSS PV2	229.40 (2.59)	223.23 (.83)	6.17 (2.75)	2.25*
SPSS PV3	227.01 (2.57)	224.33 (.83)	2.67 (2.73)	.98
SPSS PV4	226.87 (2.45)	224.12 (.84)	2.76 (2.74)	1.01
SPSS PV5	226.52 (2.55)	222.94 (.83)	3.58 (2.71)	1.32
SPSS PV6	231.42 (2.58)	224.81 (.84)	6.61 (2.75)	2.40*
SPSS PV7	226.62 (2.55)	223.93 (.82)	2.70 (2.71)	1.00
SPSS PV8	226.73 (2.56)	223.70 (.83)	3.03 (2.72)	1.11
SPSS PV9	227.88 (2.51)	222.49 (.84)	5.39 (2.76)	1.95
SPSS PV10	231.14 (2.63)	222.82 (.84)	8.32 (2.75)	3.02**
SPSS PVmean	228.27 (2.34)	223.63 (.76)	4.64 (2.51)	1.85

*p < .05. **p < .01. ***p < .001. SE: Standard Error.

Single-Level Regression Study

In this section "Do disciplinary climate in science classes, epistemological beliefs, index of economic, social and cultural status, inquiry-based science teaching and learning practices, instrumental motivation, enjoyment of science, science self-efficacy, teacher-directed science instruction, teacher support in a science classes predict PISA 2015 science performance in Turkey?" sub-research question was investigated. The results were given in Table 3.

When sample weights and plausible values were taken into account instrumental motivation and teacher support in science classes could not predict the science performance of students. The disciplinary climate in science classes, epistemological beliefs, index of economic, social and cultural status, inquiry-based science teaching and learning practices, enjoyment of science, science self-efficacy, teacher-directed science instruction could predict science performance.

When sample weights and plausible values were not used, among 11 cases, 8 of them produced incorrect results. The main problem was that more variables were found to be significantly related to

the dependent variable which was also related to incorrect standard error estimation. Both using only PV1 or PVmean produced incorrect results. On general R square values were not changed dramatically however, R^2 of PVmean was higher. This example also illustrates that plausible values and sample weights should be used.

Tablo 3. Factors Predicting Science Performance

Method	discipline	beliefs	SES	Inquiry b. science	motivation	enjoy	Self-efficacy	Teacher-directed	support	R^2
IDB Analyzer PV1-PV10	.09***	.19***	.27***	-.19***	.03	.09***	.08***	.04*	.03	.20
SPSS PV1	.08***	.19***	.26***	-.18***	**.03***	.09***	.08***	.05***	**.03***	.19
SPSS PV2	.07***	.20***	.26***	-.18***	**.03***	.11***	.08***	.04**	.02	.20
SPSS PV3	.09***	.20***	.27***	-.19***	**.03***	.09***	.08***	.05***	.02	.20
SPSS PV4	.09***	.19***	.26***	-.19***	.02	.11***	.07***	.05***	.02	.20
SPSS PV5	.09***	.19***	.27***	-.18***	.03	.09***	.07***	.04**	.02	.20
SPSS PV6	.10***	.19***	.26***	-.18***	.03	.09***	.07***	.05***	.02	.19
SPSS PV7	.09***	.19***	.26***	-.20***	**.03***	.09***	.08***	.05***	.03	.20
SPSS PV8	.08***	.19***	.26***	-.18***	**.03***	.10***	.07***	.05***	.03	.19
SPSS PV9	.10***	.19***	.25***	-.20***	.02	.11***	.08***	.04**	**.03***	.20
SPSS PV10	.09***	.19***	.27***	-.19***	**.03***	.09***	.08***	.04**	.01	.20
SPSS PVort	.09***	.20***	.28***	-.20***	**.03***	.10***	.08***	.05***	.02	.22

*$p < .05$. **$p < .01$. ***$p < .001$.

Multilevel Prediction Study

The last sub-research question is "Do student-level variables, parents make sure that time is allocated for the homework, parents check if the homework is completed, time spent on the homework; and teacher level variables, correcting assignments and giving feedback, letting students correct their own homework, discussing homework in the classroom, monitoring completeness of the homework, using homework for grading predict TIMSS 2015 reasoning score in Turkey?". As both student level and teacher level variables were included in the model, multilevel regression was used. The results were given in Table 4.

The intraclass correlation was calculated as 0.32. This value represented that student scores were not independent and scores of the students in the same classrooms were related. Therefore, a multilevel regression analysis was necessary. Also, 32% of the total variance came from between classroom variance and 68% of the total variance came from within classroom variance. The variables of this research question could explain 4% of the variance in student level and 7% of the variance in teacher level. These explained variances were small which implied that the model was not a good one.

The results showed that among student-level variables, parents make sure that time is allocated for the homework and parents check if the homework is completed could predict reasoning scores of students. There was a positive relationship between parents make sure that time is allocated for the homework and reasoning scores. However, there was a negative relationship between parents check if the homework is completed and reasoning scores. Among teacher-level variables, there was a positive relationship between monitoring completeness of the homework and reasoning scores.

Table 4. Standard Coefficients of Multilevel Regression

Variables	Coefficient
Level-1	
time is allocated for the homework	.17***
parents check if the homework is completed	-.19***
time spent on the homework	-.03
Level-2	
correcting assignments	-.04
letting students correct homework	-.05
discussing homework	.10
monitoring completeness of the homework	.16*
grading	.08
Between-class explained variance	%7
Within-class explained variance	%4

*$p < .05$. **$p < .01$. ***$p < .001$.

DISCUSSION & CONCLUSION

It is known that large-scale assessment results are critical in determining educational policies, curriculum reforms and decision-making processes in the use of contemporary innovative practices in education (Hamilton, 2003). The large-scale assessment results also allow cross-country comparisons of various sizes and provide detailed information about the various elements included in the countries' own education system. As a result of its' crucial role in policymaking and the possible influence involving millions of stakeholders, it is required to analyze the data obtained from these tests properly. As it was seen in the cases of examples known as *PISA shock phenomenon* (Wiseman, 2013), misinterpretation of large-scale data sets through primitive and descriptive inferences led irrelevant and radical policy changes in some countries in the past. For instance, Germany's radical policy changes right after their inauguration of PISA 2000 results that were below the OECD average (Waldow, 2009) or Japan's sharp policy changes following the decreased performances in PISA 2000-2003 literacy and maths performance on PISA 2003-2006 could be examples for those misinterpretations (Wiseman, 2013). These instances support the argument that the analysis of the large-scale data sets requires the use of relevant techniques to be embraced (Wiseman, 2013).

As seen in the research questions, in the case of not using sample weights and plausible values appropriately may lead to incorrect results. For instance, as shown in research question 1, in the case of using proper methods of analysis with TIMSS 2015 data led no statistically significant differences between boys' and girls' math performance of Turkey sample. However, statistically significant difference between the groups could be found when the appropriate analysis was not conducted. Similarly, in the second research question, it was shown that multiple regression analysis results could be wrong in the case of not using sample weights and plausible values appropriately. Without taking into consideration of sample weights and plausible values led to 8 incorrect results out of 11 datasets. As Von Davier et al. (2009) and Rutkowski et al. (2010) emphasized within the context of Bulgaria's TIMSS 2007 performance instances, it is vital to use sample weights and plausible values to perform large-scale data set analysis.

Yet, it is seen from the relevant literature regarding the large scale assessment analysis, the awareness regarding embrace these accurate techniques is not as intended. Moreover, the undergraduate or graduate courses offered as well as workshops organized by either researchers or institutions that emphasize how to analyze these LSA data are rare in the national context. As a result of these, even

though there are some studies considering these features of LSA, there are also some studies that use inaccurately only one plausible value or the average of plausible values without using sample weights. In order to overcome these obstacles, this study exemplifies the importance of using sample weights and plausible values by providing the syntaxes. It is recommended for readers of large scale assessments to critically assess whether appropriate techniques are used or not before relying on the research findings. Also, researchers are required to carefully investigate the features of the software embraced in the analysis and to examine the technical reports in the literature for appropriate sample weight use as various sample weights are used in different data sets.

REFERENCES

Addey, C., & Sellar, S. (2018). Why do countries participate in PISA? Understanding the role of international large-scale assessments in global education policy. In A. Verger, M. Novelli & H. Kosar-Altınyeken (Eds.), *Global education policy and international development: New agendas, issues and policies*, (pp. 97-118). New York, NY: Bloomsbury Publishing.

Addey, C., Sellar, S., Steiner-Khamsi, G., Lingard, B., & Verger A. (2017). Forum discussion: The rise of international large-scale assessments and rationales for participation. *Compare, 47*(3), 434-452. doi:10.1080/03057925.2017.1301399

Aydın, A., Selvitopu, A., & Kaya, M. (2018). Eğitime yapılan yatırımlar ve PISA 2015 sonuçları karşılaştırmalı bir inceleme. *İlköğretim Online, 17*(3), 1283-1301.

Bialecki, I., Jakubowski, M., & Wiśniewski, J. (2017). Education policy in Poland: The impact of PISA (and other international studies). *European Journal of Education, 52*(2), 167-174.

Carvalho, L. M. & Estela Costa, E. (2015) Seeing education with one's own eyes and through PISA lenses: considerations of the reception of PISA in European countries, *Discourse: Studies in the Cultural Politics of Education, 36*(5), 638-646. doi:10.1080/01596306.2013.871449

Ertl, H. (2006). Educational standards and the changing discourse on education: The reception and consequences of the PISA study in Germany. *Oxford Review of Education*, *32*(5), 619-634.

Figazzolo, L. (2009). *Testing, ranking, reforming: Impact of PISA 2006 on the education policy debate*. Brussels: Education International.

Froese-Germain, B. (2010). The OECD, PISA and the impacts on educational policy. *Canadian Teachers' Federation (NJ1).* Retrieved from http://eric.ed.gov/?id=ED532562

Gonzalez, E. J. (2012). Rescaling sampling weights and selecting mini-samples from large-scale assessment databases. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 5*, 117–134.

Gür, B. S., Celik, Z., & Özoğlu, M. (2012). Policy options for Turkey: A critique of the interpretation and utilization of PISA results in Turkey. *Journal of Education Policy*, *27*(1), 1-21.

Hamilton, L. (2003). Assessment as a policy tool. *Review of research in education*, *27*(1), 25-68.

International Association for the Evaluation of Educational Achievement (IEA) (2019). IDB Analyzer (version 4.0). Hamburg, Germany: IEA Hamburg.

Landahl, J. (2018): De-scandalisation and international assessments: the reception of IEA surveys in Sweden during the 1970s. *Globalisation, Societies and Education, 16*(5), 566-576. doi:10.1080/14767724.2018.1531235

LaRoche, S., & Foy, P. (2016). Sample design in TIMSS Advanced 2015. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS Advanced 2015* (pp. 3.1–3.27). Erişim adresi http://timssandpirls.bc.edu/publications/timss/2015- a-methods/chapter-3.html

LaRoche, S., & Foy, P. (2016). Sample implementation in TIMSS 2015. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (pp. 5.1-5.175). Retrieved from http://timss.bc.edu/publications/timss/2015-methods/chapter-5.html

Laukaityte, I., & Wiberg, M. (2017). Using plausible values in secondary analysis in large-scale assessments. *Communications in Statistics-Theory and Methods*, *46*(22), 11341-11357.

Martens, K., & Niemann, D. (2010). Governance by comparison: How ratings & rankings impact national policy-making in education (No. 139). *TranState Working Paper*. Bremen: University of Bremen Collaborative Research Centre

Milli Eğitim Bakanlığı (MEB). (2014). *TIMSS 2011 ulusal matematik ve fen raporu 8. sınıflar*. Ankara: T.C. Milli Eğitim Bakanlığı Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü.

Milli Eğitim Bakanlığı (MEB). (2016a). *PISA 2015 ulusal raporu*. Ankara: T.C. Milli Eğitim Bakanlığı Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü.

Milli Eğitim Bakanlığı (MEB). (2016b). *TIMSS 2015 ulusal matematik ve fen bilimleri ön raporu 4. ve 8. sınıflar*. Ankara: MEB: Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü.

Milli Eğitim Bakanlığı (MEB). (2018). *2*018 Liselere Geçiş Sistemi (LGS): Merkezi sınavla yerleşen öğrencilerin performansı. *Eğitim Analiz ve Değerlendirme Raporları Serisi (No. 3)*. Ankara: T.C. Milli Eğitim Bakanlığı.

Michel, A. (2017). The contribution of PISA to the convergence of education policies in Europe. *European Journal of Education, 52*(2), 206-216.

Monseur, C., & Adams, R. (2009). Plausible values: How to deal with their limitations. *Journal of Applied Measurement, 10*(3), 1-15.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA). Retrieved from http://timssandpirls.bc.edu/timss2019/frameworks/

Muthén, L. K., & Muthén, B. O. (2015). *Mplus user's guide*. (7th ed.). Los Angeles, CA: Muthén and Muthén.

Novoa, A. & Yariv-Mashal, T. (2003). Comparative research in education: A mode of governance or a historical journey? *Comparative Education, 39*(4), 423–438.

The Organisation for Economic Co-operation and Development (OECD). (2016). *Skills matter: Further results from the survey of Adult Skills*. OECD Skills Studies. Paris: OECD Publishing. doi:10.1787/9789264258051-en.

The Organisation for Economic Co-operation and Development (OECD). (2017). *PISA 2015 Technical Report*. Paris: OECD Publishing. Retrieved from https://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf

The Organisation for Economic Co-operation and Development (OECD). (2019). *PISA 2018 Assessment and Analytical Framework*. PISA. Paris: OECD Publishing. doi:10.1787/b25efab8-en.

Pizmony-Levy, O. (2018). Compare globally, interpret locally: international assessments and news media in Israel. *Globalisation, Societies and Education, 16*(5), 577-595. doi:10.1080/14767724.20.18.1531236

Rubin, D. (1987). *Multiple imputation for nonresponse in sample surveys*. New York: John Wiley.

Rust, K. (2013). Sampling, weighting, and variance estimation in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (1st ed., pp. 117–154). New York, NY: Chapman and Hall/CRC Press.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher, 39*(2), 142-151.

Steiner-Khamsi, G. & Waldow, F. (2018). PISA for scandalisation, PISA for projection: the use of international large-scale assessments in education policy making – an introduction. *Globalisation, Societies and Education, 16*(5), 557-565. doi:10.1080/14767724.2018.1531234

Tiana Ferrer, A. (2017). PISA in Spain: Expectations, impact and debate. *European Journal of Education, 52*, 184-191.

TEDMEM. (2016). *OECD yetişkin becerileri araştırması: Türkiye ile ilgili sonuçlar*. Ankara: Türk Eğitim Derneği Yayınları.

Von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. *IERI Monograph Series, 2*(1), 9-36.

Waldow, F. (2009). What PISA did and did not do: Germany after the 'PISA-shock'. *European Educational Research Journal, 8*(3), 476-483.

Wiseman, A. (2013). Policy responses to PISA in comparative perspective. In H.D. Meye, & A. Benavot (Eds.) *PISA, power, and policy: The emergence of global educational governance*. (pp.303–322). Oxford: Symposium Books

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2-3), 114-128.

Appendix A. Syntax of The First Research Question-A

```
Include file =
"C:\Users\exper\AppData\Roaming\IEA\IDBAnalyzerV4\bin\Data\Templates\SPSS_Macros\JB_PV.i
easps".
JB_PV  infile="D:\idb\TIMSS_2015.sav"/
       cvar=IDCNTRY BSBG01 /
       almvars=/
       rootpv=BSMMAT0 /
       tailpv=/
       npv=5/
       wgt=TOTWGT/
       nrwgt=150 /
       rwgt=/
       jkz=JKZONE/
       jkr=JKREP/
       jk2type=FULL/
       nomiss=Y/
       method=JRR/
       kfac=0/
       shrtcut=N/
       viewcod=N/
       ndec=2/
       clean = Y/
       strctry = N/
       intavg = Y/
       graphs=Y/
       selcrit = /
       selvar = /
       outdir="D:\idb"/
       outfile="PVMath_gender".
```

Appendix B. Syntax of The First Research Question-B

```
Include file =
"C:\Users\exper\AppData\Roaming\IEA\IDBAnalyzerV4\bin\Data\Templates\SPSS_Macros\JB_PV.ieasps".
JB_PV   infile=" D:\idb\prgturp1.sav"/
        cvar=CNTRYID C_Q02A /
        almvars=/
        rootpv=PVLIT /
        tailpv=/
        npv=10/
        wgt=SPFWT0/
        nrwgt=80 /
        rwgt=SPFWT/
        jkz=/
        jkr=/
        jk2type=HALF/
        nomiss=Y/
        method=PIAAC/
        kfac=0/
        shrtcut=N/
        viewcod=N/
        ndec=2/
        clean = Y/
        strctry = N/
        intavg = Y/
        graphs=Y/
        selcrit = /
        selvar = /
        outdir=" D:\idb"/
        outfile="paidjoblook".
```

Appendix C. Syntax of The Second Research Question

```
include file =
"C:\Users\Toshibanb\AppData\Roaming\IEA\IDBAnalyzerV4\bin\Data\Templates\SPSS_Macros\JB_
RegGP.ieasps".
JB_RegGP infile="C:\idb\PISA_TUR2015.sav"/
        cvar=CNTRYID /
        convar=DISCLISCI EPIST ESCS IBTEACH INSTSCIE JOYSCIE SCIEEFF TDTEACH
TEACHSUP /
        catvar=/
        codings=/
        refcats=/
        ncats=/
        PVRoots=/
        PVTails=/
        dvar0=/
        rootpv=PV /
        tailpv=SCIE /
        npv=10/
        wgt=W_FSTUWT/
        nrwgt=80 /
        rwgt=W_FSTURWT/
        jkz=/
        jkr=/
        jk2type=/
        nomiss=Y/
        method=BRR/
        missing=listwise/
        kfac=0.5/
        shrtcut=N/
        viewcod=N/
        ndec=2/
        clean = Y/
        strctry = N/
        viewprgs=Y/
        viewlbl=Y/
        qcstats=Y/
        newout=Y/
        intavg = Y/
        selcrit = /
        selvar = /
        outdir="C:\idb"/
        outfile="regression".
```

Appendix D. Syntax of The Third Research Question

```
TITLE: this is an example of a two-level
        regression analysis
```

```
DATA: FILE = dataimputedlist.dat;
        !Create a file list;

      TYPE = IMPUTATION;
      !Define that your data has multiple imputation;

 VARIABLE:

  NAMES = IDSCHOOL IDSTUD BSBG11C BSBG11D BSBM20B
  BTBM25CA BTBM25CB BTBM25CC BTBM25CD BTBM25CE
  REAPV WGTADJ1WGTFAC1 WGTADJ2WGTFAC2WGTADJ3WGTFAC3;

  USEVARIABLES ARE IDSCHOOL BSBG11C BSBG11D BSBM20B
  BTBM25CA BTBM25CB BTBM25CC BTBM25CD BTBM25CE
  REAPV WGTADJ1WGTFAC1 WGTADJ2WGTFAC2WGTADJ3WGTFAC3;

  CLUSTER = IDSCHOOL;
  !Define Cluster Variable here;

  MISSING = ALL (9999);

  WEIGHT = WGTADJ1WGTFAC1;
  BWEIGHT = WGTADJ2WGTFAC2WGTADJ3WGTFAC3;
  !Define Sample Weights Here;

  WITHIN = BSBG11C BSBG11D BSBM20B;
   !Define Level1 variables here;

  BETWEEN = BTBM25CA BTBM25CB BTBM25CC BTBM25CD BTBM25CE;
  !Define Level2 variables here;

ANALYSIS:     TYPE = TWOLEVEL;
!Define number of level here;

MODEL:
        %WITHIN%
        REAPV on BSBG11C BSBG11D BSBM20B;
!Define Level1 relationships here;

        %BETWEEN%
        REAPV on BTBM25CA BTBM25CB BTBM25CC BTBM25CD BTBM25CE;
!Define Level2 relationships here;

OUTPUT: STANDARDIZED;
!For standardized coefficients;
```

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(1); 61-75

Simultaneous Estimation of Overall Score and Subscores Using MIRT, HO-IRT and Bi-factor Model on TIMSS Data

Ayşenur ERDEMİR * Hakan Yavuz ATAR **

Abstract
In educational testing, there is an increasing interest in the simultaneous estimation of the overall scores and subscores. This study aims to compare the reliability and precision of the simultaneous estimation of overall scores and sub-scores using MIRT, HO-IRT and Bi-factor models. TIMSS 2015 mathematics scores have been used as a data set in this study. The TIMSS 2015 mathematics test consists of 35 items, four of which are polytomously scored (0-1-2), and the rest of the items are dichotomously scored (0-1). The four content domains include number (14 items), algebra (9 items), geometry (6 items), and data and change (6 items). Ability parameters were estimated using the BMIRT software. The results showed that the MIRT and HO-IRT methods performed similarly in terms of precision and reliability for subscore estimates. The MIRT maximum information method had the smallest standard error of measurement for the overall score estimates. All three methods performed similarly in terms of the overall score reliability. The findings suggest that among the three methods compared, HO-IRT appears to be a better choice in the simultaneous estimation of the overall score and subscores for the data from TIMSS 2015. Recommendations for the testing practices and future research are provided.

Key Words: TIMSS, subscores, multidimensional item response theory, higher-order item response theory, bi-factor model.

INTRODUCTION

Many tests in educational and psychological testing generally measure more than one ability, which makes them multidimensional inherently (Reckase, 1985; 1997). Tests may be inherently multidimensional due to the intended content or construct structure of the tests (Ackerman, Gierl, & Walker, 2003). Tests consisting of different content domains often measure a primary ability and additional abilities; thus, each item measures the primary ability and one additional secondary ability. Content categories can be considered as the source of secondary abilities. That is, while the primary ability is the estimated overall score, subscores for content categories are considered secondary abilities (DeMars, 2005). Subscores estimated from secondary abilities have been of substantial importance recently (DeMars, 2005; Reckase & Xu, 2015; Sinharay, Haberman, & Wainer, 2011; Wedman & Lyren, 2015). It is because of the potential diagnostic value of the subscores in future remedial work in which students have a chance to know their weaknesses and strengths in different content domains that the test measures (Haberman & Sinharay, 2010). Haberman (2008) and Sinharay (2010) focused on the added value of subscores over the total score by using Classical Test Theory methods. Brennan (2012) suggested the utility index similar to Haberman's method. Besides, the subscore augmentation method developed by Wainer, Sheehan, and Wang (2000) is used to examine whether getting information from other portions of the test (augmented subscore) estimates the subscore more accurately.

* Res Assist., Gazi University, Gazi Faculty of Education, Ankara-Turkey, erdemiraysenur@gmail.com, ORCID ID: 0000-0001-9656-0878
** Prof. PhD., Gazi University, Gazi Faculty of Education, Ankara-Turkey, hakanatar@gazi.edu.tr, ORCID ID: 0000-0001-5372-1926

To cite this article:
Erdemir, A.. & Atar, H. Y. (2020). Simultaneous estimation of overall score and subscores using MIRT, HO-IRT, and Bi-factor Model on TIMSS data. *Journal of Measurement and Evaluation in Education and Psychology, 11*(1), 61-75. doi: 10.21031/epod.645478

Received: 11.11.2019
Accepted: 25.02.2020

The psychometric quality of subscores is also of importance when they are utilized by policymakers, test takers, and educators for the purpose of diagnosis and admission (Haberman, 2008; Monaghan, 2006). According to the Standard 1.14 of the Standards of Educational and Psychological Testing (2014, p.27), "When a test provides more than one score, the distinctiveness and reliability of the separate scores should be demonstrated." Over the years, researchers have examined the methods arguing the psychometric quality of subscores (de la Torre & Patz, 2005; DeMars, 2005; Fan, 2016; Haberman, 2008; Haberman & Sinharay, 2010; Longabach, 2015; Md Desa, 2012; Shin, 2007; Sinharay, 2010; Stone, Ye, Zhu & Lane, 2010; Wang, Chen, & Cheng, 2004; Yao, 2014; Yao & Boughton, 2007).

In multidimensional tests, when the overall score is reported, it shows the test-takers' achievement levels concerning the overall construct of the test subject. Subscores, on the other hand, give additional information about the strengths and weaknesses of test-takers in the domain abilities while the overall score presents a general profile of the test-takers. For example, the TOEFL test, which is the English-language test, has four content domains (reading, listening, speaking, and writing). For this test, test-takers receive four subscores related to each skill and a total score as a representative of general English-language ability. Since many tests have a multidimensional structure, the interest in estimating and reporting overall scores and subscores simultaneously has increased (Liu & Liu, 2017). Simultaneous estimation of those scores provides test takers and educators with more detailed information about the primary and secondary ability levels of students (Yao, 2010). More clearly, as opposed to the separate estimation of the primary and secondary abilities, simultaneous estimation means one can have the information on those abilities with one single analysis.

There are studies discussing the methods estimating the overall score and subscores simultaneously (de la Torre & Song, 2009; de la Torre & Song, 2010; Liu, Li, & Liu, 2018; Soysal & Kelecioğlu, 2018; Yao, 2010). In all these studies, it is emphasized that the reliability of scores is very important when the overall scores and subscores need to be reported. Yao (2010) states that the simple averaging method is the most commonly used method to obtain the overall score by averaging the domain scores. She also indicates that simply averaging the domain scores ignores (a) different maximum raw score points of different domains, (b) correlation between the domain abilities, and (c) the possibility of having a different relationship between overall scores and domain scores at different score points. In order to overcome these problems, Yao (2010) proposed using the Multidimensional Item Response Theory (MIRT) maximum information method for the overall score instead of the simple averaging method. The proposed method does not assume any linear relationship between the overall score and domain scores. In the study, subscores were estimated by using MIRT, and the overall scores were estimated by using the MIRT maximum information method. Estimated overall and subscores were compared to those obtained from the Higher-Order Item Response Theory (HO-IRT), Bi-factor, and unidimensional IRT methods. It is found that the MIRT method provides reliable subscores similar to the HO-IRT method and also reliable overall score. The MIRT maximum information method produced overall scores with the smallest standard error of measurement (Yao, 2010).

de la Torre and Song (2009) also proposed using Higher-order Item Response Theory approach for simultaneous estimation of overall and domain abilities. The HO-IRT method assumes a linear relationship between the overall score and the domain score, unlike the MIRT method. In the study, the HO-IRT method was compared with the unidimensional IRT (UIRT) in which the overall ability is estimated using all items ignoring the multidimensional structure of the data, and the domain abilities are estimated using corresponding subsets of items, separately. The findings of the study show that the overall and domain abilities can be estimated more efficiently by using the HO-IRT method. Additionally, in the HO-IRT framework, it is possible to obtain efficient overall and domain ability estimates with small sample sizes and small number of items (de la Torre & Song, 2010).

To estimate the overall score and domain scores based on the bi-factor model, Liu et al. (2018) introduced six methods in the framework of the bi-factor model and compared them with the MIRT method. The weights of the general and domain factors were calculated in different ways in those six bi-factor methods. It is found that the most accurate and reliable overall and domain scores in most conditions were obtained using Bi-factor-M4 and Bi-factor-M6 methods, weights of which were computed using discrimination parameters for a specific domain. In the bi-factor methods, the domain-

specific factors are orthogonal to the general factor and each other, unlike the MIRT and HO-IRT methods.

Related research regarding simultaneous estimation of the overall and subscores seems to be few in number (de la Torre & Song, 2010; Liu et al., 2018; Soysal & Kelecioğlu, 2018; Yao, 2010). The present study aims to contribute to the related research. The purpose of the study is to investigate by using which method simultaneous estimation of the overall score and subscores yields more accurate and reliable ability estimates. For this purpose, MIRT, HO-IRT, and bi-factor general model, the most suggested methods in literature, were used in the study. This study also differs from earlier research in that it runs the analysis on mixed-format data, including both dichotomously and polytomously scored items, whereas all other studies used data consisting only dichotomously or polytomously scored items. At this point, using mixed-format data is thought to be important since tests containing a mixture of multiple-choice and constructed-response items are used in many testing situations (Lane, 2005; Yao & Schwarz, 2006).

Ability Estimation with Multiple Dimensions

Multidimensional Item Response Theory

Multidimensional Item Response Theory is a method that provides "a reasonably accurate representation of the relationship between persons' locations in a multidimensional space and the probabilities of their responses to a test item" (Reckase, 2009, p. 53) with a particular mathematical expression. An essential distinction between MIRT models related to the structure of the data is whether the probability of responses to any test item is influenced by one latent dimension or not. If this is the case, the structure of the data is defined as between-item dimensionality (simple-structure). If responses to one item are affected by more than one ability, then, it is denoted as within-item dimensionality (complex structure; Adams, Wilson, & Wang, 1997). In this study, the data were assumed to follow a simple structure because each item was modeled as depending on one specific ability dimension.

Additionally, there are several models within MIRT varying basically in terms of the number of possible score points for the items: MIRT models for dichotomously scored items and MIRT models for polytomously scored items. All of the MIRT models can be considered as generalizations of unidimensional IRT models (Reckase, 1997). However, many tests contain both dichotomously and polytomously scored items on the same test form, which creates a need to use different item response models together (Yao & Schwarz, 2006). TIMSS mathematics achievement test also contains mixed item types. Therefore, in the present study, the TIMSS data were examined using the multidimensional three-parameter logistic (M-3PL) model for dichotomously scored items and the multidimensional two-parameter partial credit model (M-2PPC) applied to polytomously scored items as suggested in the study of Yao & Schwarz (2006). For a dichotomous item *j*, the probability of a correct response to item *j* for an examinee with ability $\vec{\boldsymbol{\theta}}_{\mathbf{i}} = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iD})$ for the M-3PL model (Reckase, 1997) is

$$P_{ij1} = P(x_{ij} = 1 \mid \vec{\boldsymbol{\theta}}_{\mathbf{i}}, \vec{\boldsymbol{\beta}}_{\mathbf{j}}) = \beta_{3j} + \frac{1 - \beta_{3j}}{1 + e^{(-\vec{\boldsymbol{\beta}}_{\mathbf{2j}} \odot \vec{\boldsymbol{\theta}}_{\mathbf{i}}^{\mathbf{T}} + \beta_{1j})}}, \quad (1)$$

where

x_{ij} = the response of examinee *i* to item *j*

$\vec{\boldsymbol{\beta}}_{\mathbf{j}}$ = the parameters for the j^{th} item $(\vec{\beta}_{2j}, \beta_{1j}, \beta_{3j})$

$\vec{\boldsymbol{\beta}}_{\mathbf{2j}}$ = a vector of dimension D of item discrimination parameters $(\beta_{2j1}, \ldots, \beta_{2jD})$

β_{1j} = the scale difficulty parameter

β_{3j} = the scale guessing parameter

$\vec{\boldsymbol{\beta}}_{\mathbf{2j}} \odot \vec{\theta}_i^T$ = a dot product of two vectors.

For a polytomous item *j*, the probability of a response *k*−1 to item *j* for an examinee with ability $\vec{\boldsymbol{\theta}}_{\boldsymbol{i}}$ for the M-2PPC model (Yao & Schwarz, 2006) is

$$P_{ijk} = P(x_{ij} = k-1 \mid \vec{\boldsymbol{\theta}}_{\mathbf{i}}, \vec{\boldsymbol{\beta}}_{\mathbf{j}}) = \frac{e^{(k-1)\vec{\boldsymbol{\beta}}_{\mathbf{2j}} \odot \vec{\boldsymbol{\theta}}_{\mathbf{i}} - \sum_{t=1}^{k} \beta_{\delta_t j}}}{\sum_{m=1}^{K_j} e^{\left((m-1)\vec{\boldsymbol{\beta}}_{\mathbf{2j}} \odot \vec{\boldsymbol{\theta}}_{\mathbf{i}}^{\mathbf{T}} - \sum_{t=1}^{m} \beta_{\delta_t j}\right)}}, \quad (2)$$

where

x_{ij} = the response of examinee *i* to item *j* (0, …, $K_j - 1$)

$\vec{\boldsymbol{\beta}}_{\mathbf{j}}$ = the parameters for the *j*th item ($\vec{\boldsymbol{\beta}}_{\mathbf{2j}}$, $\beta_{\delta_2 j}$, …, $\beta_{\delta_{K_j} j}$)

$\vec{\boldsymbol{\beta}}_{\mathbf{2j}}$ = a vector of dimension D of item discrimination parameters (β_{2j1}, …, β_{2jD})

$\beta_{\delta_k j}$ = the threshold parameters for $k = 1, 2, \ldots, K_j$; $\beta_{1j} = 0$ and K_j = the number of response categories for the *j*th item.

Higher-Order Item Response Theory

de la Torre and Song (2009) proposed a higher-order multidimensional IRT approach in which overall and domain abilities can be specified simultaneously. In this model, the first order describes domain-specific abilities, while the second-order can be viewed as the overall ability. It is considered that each domain is unidimensional; the second-order ability contains all the domain abilities, so the overall ability is also viewed as unidimensional. de la Torre and Hong (2010) stated that a test is deemed multi-unidimensional in the HO-IRT framework.

The HO-IRT method uses a hierarchical Bayesian framework (de la Torre et al., 2011), and the domain abilities are considered as linear functions of the overall ability, expressed as

$$\theta_i^{(d)} = \lambda^{(d)}\theta_i + \varepsilon_{id}, \quad (3)$$

where

θ_i = the overall ability,

$\theta_i^{(d)}$ = the domain-specific abilities, $d = 1, 2, \ldots, D$,

$\lambda^{(d)}$ = the latent coefficient in regressing the ability *d* on the overall ability,

ε_{id} = the error term following a normal distribution with a mean of zero and variance of $1 - \lambda^{(d)2}$, and $|\lambda^{(d)}| \leq 1$.

The latent regression coefficient, $\lambda^{(d)}$, also means the correlation between the overall and domain abilities. Mathematically, $\lambda^{(d)}$ can have negative values, but it is generally expected to be positive since domain abilities are typically related to the overall ability.

Focusing on estimating abilities of test-takers (Equation 3), the model parameters that need to be estimated are the overall ability, domain abilities, and the latent regression parameters $\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(D)}$. With a hierarchical Bayesian framework, the model formulation is expressed as follows (de la Torre & Song, 2009):

$$\theta_i \sim N(0,1), \quad (4)$$

$$\lambda^{(d)} \sim U(-1.0, 1.0), \quad (5)$$

and

$$\theta_i^{(d)} \mid \theta_i, \lambda^{(d)} \sim N\left(\lambda^{(d)}\theta_i, 1 - \lambda^{(d)2}\right). \quad (6)$$

The model parameters are estimated by using MCMC sampling procedure. First, the overall ability θ_i is sampled from a normal distribution (Equation 4), and the regression coefficient is sampled from a uniform distribution (Equation 5). Then, based on the estimated overall ability and the regression

coefficients, the MCMC procedure samples the domain abilities with the sixth equation (de la Torre & Hong, 2010; de la Torre & Song, 2009).

Bi-factor General Model

The bi-factor model (Gibbons & Hedeker, 1992) defines a general factor on which all the items load and domain-specific factors on which the items related to that dimension load. The domain-specific factors are orthogonal to the general factor. The method provides estimates of the overall ability and domain abilities at the same time. It is considered that the domain factors are nuisance traits within the Bi-factor framework, which yields a more meaningful overall ability (DeMars, 2013; Yao, 2010).

Cai, Yang, and Hansen (2011) demonstrated the factor pattern of the standard item bi-factor measurement structure as

$$\begin{pmatrix} a_{10} & a_{11} & 0 \\ a_{20} & a_{21} & 0 \\ a_{30} & a_{31} & 0 \\ a_{40} & 0 & a_{42} \\ a_{50} & 0 & a_{52} \\ a_{60} & 0 & a_{62} \end{pmatrix}.$$

As seen in the pattern, there are six items, one general and two domain-specific factors. The *a*s are the indicators of item discrimination parameters, which are similar to the factor loadings. The first factor is the general factor, and the last two columns refer to the domain factors (Cai et al., 2011).

As defined in Liu et al.'s (2018) study, in the vector of item discrimination parameters, only the one for the general factor (β_{aj}) and one discrimination parameter of s^{th} subscale (β_{sj}) have values other than zero. The ability vector of each examinee includes one overall ability for the general factor (θ_{ia}) and domain-specific abilities for S specific factors ($\theta_{i1}, \dots, \theta_{is}, \dots, \theta_{iS}$).

Based on the Bi-factor model, estimation of the overall score and domain scores can be expressed as follows:

$$\theta_{overall} = w_{1a}\theta_{ia} + \sum_{s=1}^{S} w_{1s}\theta_{is} \quad (7)$$

and

$$\theta_{domain\_s} = w_{2a}\theta_{ia} + w_{2s}\theta_{is}, \quad (8)$$

where

w_{1a} = weight of the general factor for the overall score

w_{1s} = weight of the domain factors for the overall score

w_{2a} = weight of the general factor for the domain scores

w_{2s} = weight of the domain factors for the domain scores.

Thus, the overall score (Equation 7)) is a weighted composite of the general factor (θ_{ia}) and all domain factors (($\theta_{i1}, \dots, \theta_{is}, \dots, \theta_{iS}$), while the domain score (Equation 8) for the s^{th} factor is a weighted composite of the general factor (θ_{ia}) and the relevant domain-specific factor (θ_{is}). In the current study, the Bi-factor general model was employed by using 1 and 0 as the weights, as in the study of Yao (2010): $w_{1a} = 1, w_{1s} = 0$ and $w_{2a} = 0, w_{2s} = 1$. In this method, the general factor represents the overall score, while the domain-specific factors represent subscores.

METHOD

Data Description

Eighth graders' responses to the mathematics test in Trends in International Mathematics and Science Study (TIMSS) 2015 were used in this study. Each country's data from the 1st booklet of mathematics achievement test were merged into a whole data set. The reason behind choosing 1st booklet is that it is the booklet that has the largest number of polytomously-scores items (four items). For handling missing data, the listwise deletion method was utilized because the researchers aimed to analyze the data consisting of the subjects who answered all of the items The final version of the data consists of 5732 students from all the countries who were administered the 1st assessment booklet in TIMSS 2015. Table 1 shows the distribution of scoring types and contents for the chosen test form for the current study.

Table 1. Scoring Types and Content Distribution for The Data

Content domain	Scoring types	Number of items
Number	Dichotomously-scored	11
	Polytomously-scored	3
Algebra	Dichotomously-scored	9
Geometry	Dichotomously-scored	5
	Polytomously-scored	1
Data and Chance	Dichotomously-scored	6

As shown in Table 1, the test has four content domains, which are number (14 items), algebra (9 items), geometry (6 items), and data and change (6 items). The total number of items is 35, four of which are polytomously scored (0-1-2), and the rest of the items are dichotomously scored (0-1).

Data Analysis

Dimensionality analysis

In order to improve interpretations and uses of scores, the dimensional structure of the data is essential to get evidence of validity (Reckase & Xu, 2015). Dimensionality shows the relationship between a test and response patterns, which gives clues about the latent structure measured by the test. Wainer and Thissen (1996) mention the fixed and random forms of dimensionality. While random dimensionality is a concept explaining the possibility of encountering some "unexpected" dimensions, fixed dimensionality is a somewhat "expected" situation. In particular, it is usual to see multidimensionality in scores when the test has multiple content domains. It can be assumed that the data have a multidimensional structure when the test has content domains. Under this circumstance, it is said that it might be more reasonable and effective to use confirmatory dimensionality assessment (Zhang, 2016). Therefore, confirmatory methods were used to assess the dimensionality structure of the data in this study. Confirmatory Factor Analysis (CFA) and content-based confirmatory mode of Poly-DETECT (Zhang & Stout, 1999a, 1999b; Zhang, 2007) were the methods utilized as dimensionality analysis in the current study.

The poly-DETECT analysis was done through the *sirt* package (Robitzsch, 2018). The result of the analysis gives the indices DETECT, ASSI and RATIO. The information about the evaluation of these indices is presented in Table 2 (Jang & Roussos, 2007; Zhang, 2007):

Table 2. Dimensionality Indices of the Poly-DETECT Analysis and Their Evaluation

Index	Critical Values	Explanation
DETECT	DETECT > 1.00	Strong multidimensionality
	.40 < DETECT < 1.00	Moderate multidimensionality
	.20 < DETECT < .40	Weak multidimensionality
	DETECT < .20	Essential unidimensionality
ASSI	ASSI=1	Maximum value under simple structure
	ASSI > .25	Essential deviation from unidimensionality
	ASSI < .25	Essential unidimensionality
RATIO	RATIO=1	Maximum value under simple structure
	RATIO > .36	Essential deviation from unidimensionality
	RATIO < .36	Essential unidimensionality

The DETECT index shows the amount of multidimensionality on a test. The DETECT value of 1 or more indicates strong multidimensionality; values of 0.4 to 1 indicate moderate to large multidimensionality; values below 0.4 indicate moderate to weak multidimensionality, and values below 0.2 indicate unidimensionality. For ASSI and RATIO indices, the critical values are 0.25 and 0.36, respectively. ASSI and RATIO values smaller than those critical values indicate that the data is essentially unidimensional. On the other hand, the data that has the ASSI and RATIO values higher than the critical values are considered to be multidimensional.

MPlus software program was used to conduct the Confirmatory Factor Analysis. Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and RMSEA (Root Mean Square Error of Approximation) are the fit indices used to test model fit. It is reported that the model fits quite well with the data when CFI and TLI have values more than 0.95, and RMSEA has a value lower than 0.05 (Hu & Bentler, 1999; Tabachnick & Fidell, 2013, p.720-723).

Estimating overall score and subscores

Three estimation methods (MIRT, HO-IRT, and Bi-factor) were used to obtain the overall score (mathematics achievement) and subscores (number, algebra, geometry, and data and chance) for 5732 test takers who were administered the first booklet of TIMSS 2015. Ability parameters for the methods were estimated using the BMIRT software (Yao, 2003; Yao, 2013; Yao, Lewis, & Zhang, 2008). In the present study, the data were analyzed using the M-3PL model for dichotomously-scored items, and the M-2PPC applied to polytomously-scored items for all of the estimation methods. The following are brief explanations of the estimation methods and what they estimate in the context of the current data:

- MIRT: the simple structure MIRT analysis was used to estimate abilities based on four content domains. It gives four thetas (θ), each of which represents single subscore. The overall score was obtained by domain scores using maximum information method as in Yao (2010).
- HO-IRT: It is assumed that there is a linear relationship between the overall score and subscores, so the parameters for the overall ability and domain abilities were estimated simultaneously.
- Bi-factor: The Bi-factor general model estimated five abilities. The first one was the general dimension, and the other four abilities were content-specific dimensions, respectively. In the bi-factor model, content-specific dimensions are orthogonal to each other and the general dimension, and there is no correlation between dimensions.

The default priors of BMIRT software were used for the analyses in this study. The mean and variance of the ability prior distribution were 0.0 and 1.0, respectively. The priors were taken to be lognormal for the discrimination parameters with a mean of 1.5 and variance of 1.5. For the difficulty or threshold parameters, a standard normal distribution with a mean of 0.0 and variance of 1.5 was used. Guessing parameter c had prior beta (α, β) distribution, in which $\alpha = 100$ and $\beta = 400$.

Evaluation criteria

The conditional standard error of measurement (cSEM) was used to evaluate the accuracy of overall scores and subscores. The BMIRT program calculated the cSEM values for each student's ability parameters under studied methods estimating the overall and domain scores simultaneously. Then, the analysis of variance (ANOVA) on repeated-measures data for the cSEM was conducted to examine whether there is a significant difference among the mean errors calculated by estimation methods.

The other criterion for the evaluation of methods is reliability. A method proposed by de la Torre & Patz (2005) called Bayesian marginal ability or empirical reliability (Brown & Croudace, 2015) was applied for this study. The reliability of test *d* can be obtained from

$$\rho_d = \frac{var\,(\hat{\theta}_d)}{var\,(\hat{\theta}_d)+\overline{Pvar\,(\hat{\theta}_d)}}. \quad (9)$$

The observed (Equation 10) and marginal posterior (Equation 11) variance of the overall or domain ability estimates are computed from the estimated ability scores $\hat{\theta}$ and their standard errors (SE) in a sample of N test takers:

$$var\left(\hat{\theta}_d\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i - \overline{\hat{\theta}}\right)^2 \quad (10)$$

$$\overline{Pvar\,(\hat{\theta}_d)} = \frac{1}{N}\sum_{i=1}^{N} SE^2\left(\hat{\theta}_i\right). \quad (11)$$

For this study, reliability measures for one overall score and four subscores were obtained from the equations above for each studied methods. Higher marginal reliability indicates higher reliability of scores from the methods tested (Md Desa, 2012).

RESULTS

Dimensionality Analysis

Poly-DETECT (confirmatory mode) and Confirmatory Factor Analysis were conducted in order to examine the multidimensionality due to the content domains for mixed-format TIMSS data used in this study. Table 3 shows the results of the content-based Poly-DETECT analysis.

Table 3. The Results of the Poly-DETECT Analysis

Index	Value	Corresponding Classification	
DETECT	0.406	Moderate multidimensionality	.40 < DETECT < 1.00
ASSI	0.459	Essential deviation from unidimensionality	ASSI > .25 RATIO > .36
RATIO	0.522		

As seen in Table 3, the results yielded an essential deviation from unidimensionality in which ASSI = .459 and RATIO = 0.522. DETECT index, which is .406, means moderate multidimensionality. The values of indices obtained from the Poly-DETECT analysis provide evidence of multidimensionality for the current data.

A four-factor model was tested through CFA. The content domains with related items were taken as factors, and the model fit was evaluated. Fit indices for the data and the associated criteria are presented in Table 4.

Table 4. CFA Model Fit Indices and Associated Criteria

Index	Value	Good Fit
TLI	0.974	TLI ≥ 0.95
CFI	0.975	CFI ≥ 0.95
RMSEA	0.037	RMSEA ≤ 0.05

CFI and TLI indicated that the model fits the data well (≥ 0.95). Likewise, the RMSEA value (≤ 0.05) showed a good fit (Table 4). According to the results of CFA, the four-factor model had a good fit with the present data, which supported content-based multidimensionality. After providing evidence of the content-based multidimensionality of the data, the overall and domain abilities were obtained with the aforementioned methods.

Precision of Estimates

The selected three methods (MIRT, HO-IRT, and Bi-factor) for the current study were used through running the BMIRT program to estimate the overall and subscores simultaneously. BMIRT also provided standard errors for the estimated scores. The means for standard errors for the overall and domain ability estimates under each estimation method are summarized in Table 5.

Table 5. The Means and Standard Deviations for the Standard Errors for the Overall and Domain Abilities

	Domain				
Method	Number (14 items)	Algebra (9 items)	Geometry (6 items)	Data and Chance (6items)	Overall
MIRT	0.376 (0.125)	0.511 (0.130)	0.545 (0.142)	0.586 (0.149)	0.295 (0.124)
HO-IRT	0.332 (0.103)	0.410 (0.120)	0.422 (0.133)	0.443 (0.140)	0.474 (0.050)
Bi-factor	0.670 (0.164)	0.820 (0.163)	0.849 (0.168)	0.898 (0.178)	0.322 (0.135)

Table 5 shows the means and standard deviations for the standard errors for each ability. Generally, MIRT and HO-IRT yielded similar results, but the HO-IRT estimation method performed slightly better than MIRT for domain abilities. The Bi-factor model gave the worst standard errors for the domain abilities among all the methods and similar to the MIRT for the overall ability. The repeated-measures ANOVA results whether the difference between standard errors are statistically significant are presented in Table 6.

Table 6. The Repeated-measures ANOVA results for the Standard Errors

Ability	Source	Sum of Squares	*df*	Mean Square	*F*	Partial η^2	Pairwise comparison
Number	Methods	386.536	1.726	223.918	15465.323*	.730	All pairwise
	Error	143.239	9893.087	.014			HOIRT<MIRT<BF
Algebra	Methods	521.582	1.885	276.701	15288.071*	.727	All pairwise
	Error	195.524	10802.949	.018			HOIRT<MIRT<BF
Geometry	Methods	552.440	1.909	289.387	14196.309*	.712	All pairwise
	Error	223.018	10940.494	.020			HOIRT<MIRT<BF
Data and chance	Methods	621.124	1.925	322.731	13418.317*	.701	All pairwise
	Error	265.284	11029.804	.024			HOIRT<MIRT<BF
Overall	Methods	105.937	1.692	62.613	8162.767*	.588	All pairwise
	Error	74.377	9696.490	.008			MIRT<BF<HOIRT

*$p < .001$

The repeated-measures ANOVA with a Greenhouse-Geisser correction determined that mean standard errors differed statistically significantly when the estimation method was changed for the domain ability estimates ($F_{(1.726, 9893.087)\ number} = 15465.323$, $p < .05$, *partial* $\eta^2 = .73$; $F_{(1.885, 10802.949)\ algebra} = 15288.071$, $p < .05$, *partial* $\eta^2 = .727$; $F_{(1.909, 10940.494)\ geometry} = 14196.309$, $p < .05$, *partial* $\eta^2 = .712$; $F_{(1.925, 11029.804)\ data\ and\ chance} = 13418.317$, $p < .05$, *partial* $\eta^2 = .701$). Post hoc tests using the Bonferroni correction revealed that all pairwise comparisons were statistically significantly different from each other. According to the

results in Table 4, the HO-IRT method had the lowest standard errors for all domain abilities, and MIRT had the second-lowest standard errors. Domain abilities from the Bi-factor model were not as accurate as the other two methods.

Therefore, it can be concluded that HO-IRT elicited a statistically significant reduction in standard errors of domain ability estimates. Likewise, the overall ability results showed that the standard errors were significantly affected by the type of estimation method ($F_{(1.692,\ 9696.490)\ overall} = 8162.767$, $p < .05$, *partial* $\eta^2 = .588$). Post hoc tests using the Bonferroni correction revealed that all pairwise comparisons were significantly different from each other. The HO-IRT had the highest mean for standard errors. The MIRT and Bi-factor model had low and similar standard errors for the overall ability. In general, the three estimation methods were significantly different for all the abilities, including the overall and domain abilities.

Reliability of Scores

The overall and four domain ability estimates from the studied methods were compared in terms of marginal reliability. Estimated reliability coefficients are presented in Table 7.

Table 7. Marginal Reliability Coefficients

	Domain				
Method	Number (14 items)	Algebra (9 items)	Geometry (6 items)	Data and Chance (6items)	Overall
MIRT	0.847	0.722	0.682	0.635	0.816
HO-IRT	0.894	0.838	0.824	0.809	0.815
Bi-factor	0.539	0.301	0.253	0.161	0.876

Table 7 presents the Bayesian marginal reliability of the overall score and subscores based on four content domains. In general, MIRT and HO-IRT had substantially higher reliability across all content domains compared to the reliability of the Bi-factor model. The reliability of the Bi-factor model was extremely low for the domain scores, especially for geometry (i.e., 0.253) and data and chance (i.e. 0.161). In addition, the reliability of domains varied slightly between domains for MIRT and HO-IRT. The reliability coefficient of HO-IRT subscores was for number, 0.894; for algebra, 0.838; for geometry, 0.824, and for data and chance, .809. It can be concluded that HO-IRT was the most reliable method of estimating subscores, followed by MIRT, for all content domains for the data used in the current study. Furthermore, the reliabilities of all methods decreased as the number of items in the domains decreased. The reliability of the overall score was for MIRT, 0.816; for HO-IRT, 0.815, and for Bi-factor, 0.876. Unlike the subscores, the Bi-factor model was the most reliable method for the overall score estimation. The other two methods (MIRT and HO-IRT) also estimated the overall score with high reliability.

DISCUSSION and CONCLUSION

When the overall and domain abilities are reported to the test takers and used by the authorities, it is important to obtain accurate and reliable estimates of the overall score and subscores. The overall scores are useful in reporting the test-takers' general achievement and taking important decisions such as rank-ordering the test takers. On the other hand, the subscores provide test takers, teachers, or policymakers with more diagnostic information such as strengths and weaknesses in each domain. The simultaneous estimation of those scores can be another solution to both of the needs.

This study examined three methods of estimating the overall score and subscores simultaneously in the same model, including MIRT, HO-IRT, and Bi-factor, and compared the reliability and precision of these methods across the overall and domain ability estimates. For this purpose, the real data of mixed item types from TIMSS 2015 were used. The results of Poly-DETECT and CFA provided evidence for the content-based multidimensional structure of the data.

The study showed that the MIRT and HO-IRT methods performed similarly in terms of precision and reliability for subscore estimates. However, HO-IRT had slightly lower standard errors and higher reliability than MIRT. Likewise, de la Torre and Song (2009) stated that domain ability estimates can be more efficient by using the HO-IRT model. In addition, Yao (2010) found that MIRT and HO-IRT were quite similar in terms of estimating subscores. The precise ability estimation and reliable scores by using HO-IRT also supported the use of subscores for reporting for the current data. The Bi-factor general model had the highest standard errors and lowest reliability estimates for the domain scores. Liu et al. (2018) also did not recommend the Bi-factor, the original factor method, for reporting scores. They proposed six other methods of reporting overall and subscores as weighted composite scores of the overall and domain-specific factors in a bi-factor model.

For the overall ability estimation, the MIRT maximum information method and Bi-factor model outperformed the HO-IRT method with regard to standard errors. The MIRT maximum information method had the smallest standard error of measurement for the overall score estimates, as in the study of Yao (2010). While all three methods performed similarly and relatively good in terms of the overall score reliability, the reliability of Bi-factor model was a bit higher than the other two methods.

The analyses of the current study suggested that overall, HO-IRT seems the best solution for the simultaneous estimation of the overall and subscores for the data from TIMSS 2015. Soysal and Kelecioğlu (2018) also recommended the use of HO-IRT in estimation of overall and subscores in their study.

In the present study, only real data were used to examine the relative performance of the three methods, since the true model for the data was not known. Therefore, it is quite possible to get different results for other samples. It is suggested that future research can be done by using other real data. It is also advisable that when the simultaneous estimation of the overall and domain abilities must be done in testing practices, the relative performance of the estimation methods should be checked before reporting the scores to test takers.

REFERENCES

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22*(3), 37-51. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x

Adams, R., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23. https://doi.org/10.1177/0146621697211001

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

Brennan, R. L. (2012). *Utility indexes for decisions about subscores* (CASMA Research Report No. 33). Iowa City, IA: University of Iowa, Center for Advanced Studies in Measurement and Assessment. http://www.education.uiowa.edu/docs/default-source/casma---research/33utility-revised.pdf?sfvrsn=2

Brown, A. & Croudace, T. (2015). Scoring and estimating score precision using multidimensional IRT. In Reise, S. P. & Revicki, D. A. (Eds.). *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment (a volume in the Multivariate Applications Series).* New York: Routledge/Taylor & Francis Group.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*(3), 221–248. http://doi.org/10.1037/a0023350

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have : A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, *30*(3), 295–311. https://doi.org/10.3102/10769986030003295

de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size: A higher-order IRT model approach. *Applied Psychological Measurement, 34*, 267-285. https://doi.org/10.1177/0146621608329501

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order ırt model approach. *Applied Psychological Measurement*, *33*(8), 620–639. http://doi.org/10.1177/0146621608326423

de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscoring. *Applied Psychological Measurement, 35*(4), 296–316. http://doi.org/10.1177/0146621610378653

DeMars, C.E. (2005, August). *Scoring subscales using multidimensional item response theory models.* Poster presented at the annual meeting of the American Psychology Association. https://eric.ed.gov/?id=ED496242

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing, 13*(4), 354-378. https://doi.org/10.1080/15305058.2013.799067

Fan, F. (2016). *Subscore Reliability and Classification Consistency : A Comparison of Five Methods* (Doctoral dissertation, University of Massachusetts Amherst). https://scholarworks.umass.edu/dissertations_2/857/

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*(3), 423–436. http://doi.org/10.1007/BF02295430

Haberman, S. J. (2008). When can subscale scores have value? *Journal of Educational and Behavioral Statistics, 33*(2), 204–229. https://doi.org/10.3102/1076998607302636

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75*, 209– 227. https://doi.org/10.1007/s11336-010-9158-4

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structural analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1- 55. https://doi.org/10.1080/10705519909540118

Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement, 44*(1), 1-21. https://doi.org/10.1111/j.1745-3984.2007.00024.x

Lane, S. (2005, April). *Status and future directions for performance assessments in education.* Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Liu, Y., Li, Z., & Liu, H. (2018). Reporting Valid and Reliable Overall Scores and Domain Scores Using Bi-Factor Model. *Applied Psychological Measurement, 43*(7), 562–576. https://doi.org/10.1177/0146621618813093

Liu, Y., & Liu, H. (2017). Reporting overall scores and domain scores of bi-factor models. *Acta Psychologica Sinica, 49*(9), 1234. http://doi.org/10.3724/SP.J.1041.2017.01234

Longabach, T. (2015). *A comparison of subscore reporting methods for a state assessment of English language proficiency* (Doctoral dissertation, University of Kansas). https://kuscholarworks.ku.edu/handle/1808/19517

Md Desa, Z. N. D. (2012). *Bi-factor multidimensional item response theory modeling for subscore estimation, reliability, and classification* (Doctoral dissertation, University of Kansas). http://kuscholarworks.ku.edu/dspace/handle/1808/10126

Monaghan, W. (2006). *The facts about subscale scores* (ETS R&D Connections No. 4). Princeton, NJ: Educational Testing Service. https://www.ets.org/Media/Research/pdf/RD_Connections4.pdf

Reckase, M. D. (1985). The difficulty of items that measure more than one ability. *Applied Psychological Measurement, 9*(4), 401-412. https://doi.org/10.1177/014662168500900409

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden &R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). NY: Springer.

Reckase, M. D. (2009). *Multidimensional item response theory (statistics for social and behavioral sciences).* New York: Springer.

Reckase, M. D., & Xu, J.-R. (2015). The Evidence for a Subscore Structure in a Test of English Language Competency for English Language Learners. *Educational and Psychological Measurement, 75*(5), 805–825. https://doi.org/10.1177/0013164414554416

Robitzsch, A. (2019). Supplementary Item Response Theory Models Version. R-project, Package 'sirt' manual. https://cran.r-project.org/web/packages/sirt/sirt.pdf

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47*(2), 150–174. https://doi.org/10.1111/j.1745-3984.2010.00106.x

Sinharay, S., Haberman, S. J., & Wainer, H. (2011). Do Adjusted Subscores Lack Validity? Don't Blame the Messenger. *Educational and Psychological Measurement, 71*(5), 789–797. https://doi.org/10.1177/0013164410391782

Soysal, S., & Kelecioğlu, H. (2018). Toplam Test ve Alt Test Puanlarının Kestiriminin Hiyerarşik Madde Tepki Kuramı Modelleri ile Karşılaştırılması. *Journal of Measurement and Evaluation in Education and Psychology , 9*(2), 178-201. https://doi.org/10.21031/epod.404089

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education, 23*(1), 63–86. https://doi.org/10.1080/08957340903423651

Tabachnick B. G. & Fidel, L. S. (2013). *Using multivariate statistics (4th ed.).* MA: Allyn & Bacon, Inc.

Wainer, H., Sheehan, K. M., & Wang, X. (2000). Some paths toward making Praxis scores more useful. *Journal of Educational Measurement, 37*(2), 113–140. https://doi.org/10.1111/j.1745-3984.2000.tb01079.x

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practices, 15*(1), 22–29. https://doi.org/10.1111/j.1745-3992.1996.tb00803.x

Wang, W.-C., Chen, P.-H., & Cheng, Y.-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9*(1), 116–136. http://doi.org/10.1037/1082-989X.9.1.116

Wedman, J., & Lyren, P.- E. (2015). Methods for examining the psychometric quality of subscores: A review and application. *Practical Assessment, Research and Evaluation, 20*(1). https://scholarworks.umass.edu/pare/vol20/iss1/21/

Yao, L. (2003). BMIRT: Bayesian multivariate item response theory [computer software]. Monterey, CA: DefenseManpower Data Center. Downloaded from https://www.bmirt.com/6271.html

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement, 47*(3), 339–360. http://doi.org/10.1111/j.1745-3984.2010.00117

Yao, L. (2013). The BMIRT toolkit. Monterey. https://www.bmirt.com/8220.html

Yao, L. (2014). Multidimensional item response theory for score reporting. In Y. Cheng, & H.- H. Chang (Eds.) *Advances in modern international testing: Transition from summative to formative assessment.* Charlotte, NC: Information Age.

Yao, L., & Boughton, K. A. (2007). A Multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*(2), 83–105. http://doi.org/10.1177/0146621606291559

Yao, L., Lewis, D., & Zhang, L. (2008, April). *An introduction to the application of BMIRT: Bayesian multivariate item response theory software.* Training session presented at the meeting of the National Council on Measurement in Education, New York, NY.

Yao, L., & Schwarz R. (2006). Amultidimensional partial credit model with associated item and test statistics: An application to mixed format tests. *Applied Psychological Measurement. 30*(6), 469—492. https://doi.org/10.1177/0146621605284537

Zhang, J. (2007). Conditional covariance theory and DETECT for polytomous items. *Psychometrika, 72*(1), 69-91. https://doi.org/10.1007/s11336-004-1257-7

Zhang, J., & Stout, W. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64*(2), 129-152. https://doi.org/10.1007/BF02294532

Zhang, J., & Stout, W. (1999b). The theoretical DETECT index of dimensionality and its applicationto approximate simple structure. *Psychometrika, 64*(2), 213-249. https://doi.org/10.1007/BF02294536

Zhang, M. (2016). *Exploring dimensionality of scores for mixed-format tests* (Doctoral Dissertation, University of Iowa). https://ir.uiowa.edu/etd/2171/

Çok Boyutlu MTK, İkinci-düzey MTK ve Bifaktör Modelleri ile TIMSS Verisi için Toplam ve Alt Puanların Birlikte Kestirilmesi

Giriş

Eğitimde ölçme işlemi gerçekleştirilirken bir testin farklı yetenekleri ölçmesi yaygın bir durumdur. Bir testin alt testlerden oluştuğu durumlarda hâlihazırda birçok boyutluluk söz konusudur (Ackerman, Gierl, & Walker, 2003). Bu durumlarda test hem genel yeteneği hem de alt alanlar ile ilgili yetenekleri ölçer. Toplam puana ek olarak alt puanların da raporlanmasına ilişkin artan bir ilgi vardır. Toplam puan genele ilişkin bilgi verirken alt puanlar yanıtlayıcılara güçlü ve zayıf yönlerini detaylı bir şekilde verebilmesi açısından tanılayıcı bir değere sahiptir (Haberman & Sinharay, 2010).

Testlerin çoğunun çok boyutlu bir yapıya sahip olması ve alt alanlardan oluşması, yanıtlayıcılara ve eğitimcilere daha doğru bilgi sağlayan toplam puan ve alt puanların birlikte kestirimine olan ilgiyi arttırmıştır (Liu & Liu, 2017). Az sayıda çalışma toplam puan ve alt puanların birlikte kestirildiği yöntemleri ele almıştır (de la Torre & Song, 2009; Liu, Li, & Liu, 2018; Soysal & Kelecioğlu, 2018; Yao, 2010). De la Torre ve Song (2009) bu puanların birlikte kestiriminin sağlandığı ikinci-düzey madde tepki kuramı (MTK) yöntemini önermişlerdir. Yao (2010) çalışmasında toplam puan ve alt puanların

birlikte raporlanabildiği dört yöntemi (tek boyutlu MTK, çok boyutlu MTK, ikinci-düzey MTK ve Bifaktör model) karşılaştırmıştır. Liu ve diğerleri (2018) 6 yeni bifaktör model önermiş ve bunları çok boyutlu MTK yöntemi ile karşılaştırmıştır.

Bu çalışmanın amacı, daha doğru ve güvenilir kestirimler elde etmek amacıyla toplam puan ve alt puanların birlikte kestirildiği yöntemlerin incelenmesidir. Bu kapsamda ele alınan yöntemler, çok boyutlu MTK, ikinci-düzey MTK ve Bifaktör modeldir. Bu çalışmanın az sayıda çalışma bulunan Alana katkı sağlayacağı düşünülmektedir. Ayrıca yapılan çalışmalardan farklı olarak ikili ve çoklu puanlanan maddelerin bir arada kullanıldığı karma-format bir test üzerinden analizlerin gerçekleştirilmiş olması önemli görülmektedir.

Yöntem

Sekizinci sınıflara uygulanan TIMSS 2015 matematik başarı testi birinci kitapçığında yer alan 35 maddeye verilen yanıtlar çalışma verisi olarak kullanılmıştır. Kayıp veri ile baş etme yöntemi olarak liste bazında silme kullanılmış ve kalan 5732 öğrenci verisi analize alınmıştır. TIMSS matematik başarı testi konu temelli dört alt alandan oluşmaktadır: sayılar (14 madde), cebir (9 madde), geometri (6 madde) ve veri ve olasılık (6 madde). Testi oluşturan 35 maddeden dördü çoklu puanlanırken geri kalan 31 madde ikili puanlanmaktadır.

Veri analizi için öncelikle boyutluluk analizi yapılmıştır. Bu amaçla Poly-DETECT ve doğrulayıcı faktör analizleri gerçekleştirilmiştir. İlgili veri için toplam puan ve alt puan kestirimleri ve bunlara ilişkin hatalar, BMIRT programı kullanılarak elde edilmiştir. Yöntemlerin değerlendirilmesi için kriter olarak ele alınan indeksler yetenek kestirimlerine ilişkin standart hatalar ve güvenirlik değerleridir. Standart hata ortalamaları arasındaki fark tekrarlı ölçümler için ANOVA ile değerlendirilirken toplam puan ve alt puanlar için güvenirlik kestirimi marjinal güvenirlik indeksi ile hesaplanmış ve yorumlanmıştır.

Sonuç ve Tartışma

Çalışma verisinin boyut yapısının incelenmesi amacıyla yapılan Poly-DETECT analizi sonuçları tek boyutluluktan sapma olduğunu göstermektedir (DETECT>.40; ASSI>.25; RATIO>.36). Dört alt testin her birinin bir faktör olarak ele alındığı modelin test edildiği doğrulayıcı faktör analizi sonuçları modelin veri ile uyumlu olduğunu göstermektedir (CFI>.95; TLI>.95; RMSEA<.05). Bu bulgular alt alan bazında çok boyutluluğun olduğunu kanıtlamaktadır.

Alt puan bazında yetenek parametrelerine ilişkin hataların ortalamasına bakıldığında çok boyutlu MTK yöntemi ile elde edilen yeteneklerin en düşük hata ile kestirildiği, en yüksek hata ortalamalarının Bifaktör model altında elde edildiği görülmektedir. Toplam puan için ise çok boyutlu MTK ve Bifaktör yöntemlerinin birbirine yakın ve düşük hata ortalamasına sahip olduğu ve ikinci-düzey MTK yönteminin diğer iki kestirim yönteminden az miktarda daha fazla hata ortalaması değerine sahip olduğu sonucuna ulaşılmıştır. Tekrarlı ölçümler için ANOVA sonuçları alt puanlar için elde edilen hata ortalamalarının kestirim yöntemine göre birbirinden anlamlı olarak farklılaştığını göstermektedir estimates ($F_{(1.726,\ 9893.087)\ \text{sayılar}} = 15465.323$, $p < .05$, *kısmi* $\eta 2 = .73$; $F_{(1.885,\ 10802.949)\ \text{cebir}} = 15288.071$, $p < .05$, *kısmi* $\eta 2 = .727$; $F_{(1.909,\ 10940.494)\ \text{geometri}} = 14196.309$, $p < .05$, *kısmi* $\eta 2 = .712$; $F_{(1.925,\ 11029.804)\ \text{very ve olasılık}} = 13418.317$, $p < .05$, *kısmi* $\eta 2 = .701$). Daha sonra yapılan ikili karşılaştırmalar, bütün ikili karşılaştırmalar istatistiksel olarak anlamlı olduğu bulunmuştur. Bu bulgu, alt puanlar için hata ortalamaları dikkate alındığında, ikinci-düzey MTK yönteminin anlamlı olarak diğer yöntemlerden daha az hata ile yetenek kestirimi yaptığını göstermektedir. Çalışma verisi için Bifaktör model ile kestirilen alt puanlar ise diğer iki yöntem kadar doğru değildir. Benzer şekilde, toplam puan bazında ise yetenek parametrelerine ilişkin hataların ortalamaları yöntemlere göre birbirinden anlamlı olarak farklılaşmaktadır ($F_{(1.692,\ 9696.490)\ \text{toplam}} = 8162.767$, $p < .05$, *kısmi* $\eta^2 = .588$). Analiz sonrasında yapılan ikili karşılaştırmalar bütün çiftlerin birbirinden anlamlı olarak farklılaştığını göstermektedir. Çalışma verisi için standart hata ortalaması en yüksek olan yöntem ikinci-düzey MTK'dir. Çok boyutlu MTK ve Bifaktör modele ilişkin standart hata ortalamaları birbirine yakın ve görece düşüktür.

Bir diğer değerlendirme kriteri olan güvenirlik için çalışmada ele alınan bütün yöntemlere göre elde edilen toplam puan ve alt puanlar için marjinal güvenirlik katsayısı hesaplanmıştır. Genel olarak bakıldığında, bütün alt alanlar için çok boyutlu MTK ve ikinci-düzey MTK yöntemleri ile elde edilen puanlara ilişkin güvenirlik değerleri, Bifaktör model ile elde edilen puanlara ilişkin güvenirlik değerlerinden yüksektir. İkinci-düzey MTK ile kestirilen alt puanlara ilişkin güvenirlik kestirimleri diğerlerinden daha yüksek ve hepsi 0,80'den yüksektir. Toplam puanlar için güvenirlik kestirimleri ise çok boyutlu MTK için 0,816, ikinci-düzey MTK için 0.815 ve Bifaktör model için 0.876 olup her üçü için de görece yüksek ve birbirine yakındır. Bifaktör model ile kestirilen güvenirlik ise diğerlerinden biraz daha yüksektir.

Sonuçlar genel olarak ele alındığında, çok boyutlu MTK ve ikinci-düzey MTK yöntemleri, alt puanların kestirim doğruluğu ve güvenirlik açısından benzer özellikler göstermektedir. Fakat ikinci-düzey MTK yöntemi, çok boyutlu MTK yönteminden nispeten daha düşük standart hata ortalamalarına ve daha yüksek güvenirlik kestirimlerine sahiptir. Benzer şekilde, de la Torre ve Song (2009) da çalışmalarında, ikinci-düzey MTK kullanıldığında alt puan kestirimlerinin daha etkili olduğunu belirtmişlerdir. Yao (2010) da çok boyutlu MTK ve ikinci-düzey MTK yöntemlerinin birbirine benzer sonuçlar ürettiğini bulmuştur. Bu çalışma kapsamında Bifaktör genel model, alt puan kestirimleri için en yüksek hataya ve en düşük güvenirliğe sahiptir. Liu ve diğerleri (2018) de elde ettiği sonuçlar ile puanların raporlanmasında orijinal faktör yöntemi olan Bifaktör modelin kullanılmasını tavsiye etmediğini belirtmektedir. Toplam puan kestirimi için ise çalışmada ele alınan üç yöntemin de birbirine yakın değerler vermesine rağmen en düşük hata ile yapılan kestirimin çok boyutlu MTK'ye ait olduğu görülmektedir. Güvenirlik değerleri incelendiğinde ise ilgili üç yöntemin de yüksek güvenirliğe sahip olmakla birlikte en yüksek güvenirlik kestiriminin Bifaktör model ile elde edildiği bulunmuştur.

Özetle, bu çalışma kapsamında gerçekleştirilen analizler, TIMSS 2015 verisi için toplam puan ve alt puanların birlikte kestirildiği yöntemlerden ikinci-düzey MTK yönteminin kullanılmasını önermektedir. Soysal ve Kelecioğlu (2018) da çalışmalarının bulguları doğrultusunda geniş ölçekli testlerde toplam puan ve alt puanların birlikte kestirilmesi için ikinci-düzey MTK'nin kullanılabileceğini önermektedir.

Bu çalışmada, verilere ilişkin gerçek model bilinmediğinden, üç yöntemin göreceli performansını incelemek için yalnızca gerçek veriler kullanılmıştır. Bu nedenle, diğer örneklemler için farklı sonuçlar elde edilmesi olası görünmektedir. Başka gerçek veriler kullanılarak araştırmanın tekrarlanabileceği önerilmektedir. Ayrıca, test uygulamalarında toplam ve alt puanların eşzamanlı olarak kestirilmesi gerektiğinde, puanları yanıtlayıcılara bildirmeden önce ilgili yöntemlerin göreceli performanslarının kontrol edilmesi önerilmektedir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(1); 76-97

Changes in Literacy of Students in Turkey by Years and School Types: Performance of Students in PISA Applications

H. Eren SUNA * Hande TANBERKAN ** Mahmut ÖZER ***

Abstract
The assessment of students' academic achievement via international monitoring studies provides important insights to participating countries. Besides the cognitive performance of students, educational equity is one of the emphasized topics within the scope of Programme for International Student Assessment (PISA) study. Results regarding educational equity are quite important in Turkey because academic achievement differences among school types are relatively high in Turkey. Although a wide range of studies is conducted to examine the performance differences between school types in Turkey, it is observed that most studies focus on mean scores of school types. The aim of this study is to examine the change in student ratios at a basic- and advanced level of proficiency by school types in PISA applications between 2003 and 2018. Results show that approximately all students in science high schools and social sciences high schools have basic proficiency in all literacy fields and throughout PISA 2003 and PISA 2018. The ratio of students with basic proficiency in Anatolian high schools and Anatolian imam hatip high schools tends to be increased. However, the ratio of students with advance proficiency seems to be low in all school types in Turkey except science high schools. Steps to decrease the achievement differences between school types in Turkey within the scope of findings are suggested.

Keywords: Academic achievement, educational equity, PISA proficiency, school types, school tracking

INTRODUCTION

The assessment of students' academic achievement and literacy levels through international monitoring studies provides important feedback to the participating countries about their educational processes. These monitoring studies allow participating countries to assess the status of their students in cognitive and affective areas within the framework of international criteria. Today, the Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS) and International Computer and Information Literacy Study (ICILS) focusing on students' academic skills and Study on Social and Emotional Skills, and International Civic and Citizenship Education Study (ICCS) focusing on their cognitive skills are examples of these monitoring efforts (Australian Council for Educational Research-ACER, 2014; Hopfenbeck et al., 2018; International Association for the Evaluation of Educational Achievement-IEA, 2010; Rutkowski, Rutkowski and von Davier, 2014; Thomson, 2019).

Today, one of the most important goals of education is to provide students with the ability to use the knowledge and skills they have acquired at school in their daily lives and apply them in the situations they are unfamiliar with (Malik, 2018). In this way, the knowledge and skills acquired by the students are transferred from the theoretical context to real life, and it makes it easier for students to internalize these skills (Organization for Economic Cooperation and Development-OECD, 2019a). These skills, which are defined as literacy, include students going beyond theoretical knowledge, making decisions, and solving problems in various situations (Darling-Hammond, 2014; Hopfenbeck et al., 2018). Literacy

* PhD., Ministry of National Education, Ankara-Türkiye, herensuna@gmail.com, ORCID ID:0000-0002-6874-7472
** PhD., Ministry of National Education, Ankara-Türkiye, handetanberkan@gmail.com, ORCID ID: 0000-0001-7142-5397
*** Prof. PhD., Ministry of National Education, Ankara-Türkiye, mahmutozer2002@yahoo.com, ORCID ID: 0000-0001-8722-8670

To cite this article:
Suna, H. E., Tanberkan, H. & Ozer, M. (2020). Changes in literacy of students in Turkey by years and school types: Performance of students in PISA applications. *Journal of Measurement and Evaluation in Education and Psychology, 11*(1), 76-97. doi: 10.21031/epod.702191

Received: 11.03.2020
Accepted: 20.03.2020

is also considered important for students to be successful in business life in the long term and to participate actively in lifelong learning processes (OECD, 2019a; OECD, 2019b; Ozer, 2019b).

PISA, which has been implemented by OECD since 2000, international monitoring study with the highest participation in which students' literacy is assessed in mathematics, science, and reading (OECD, 2019a). PISA is implemented in three-year periods, and in each PISA application, one of the reading, mathematics, and science literacy is considered as the primary area. In addition to cognitive tests, student, teacher, and school-level surveys are conducted, and detailed information about the education systems of the participating countries is obtained. In this way, PISA provides essential findings of the literacy performance of students as well as the relationship between many educational variables, such as school characteristics, family, and student characteristics, with student performance (National Economic and Social Council-NESC, 2012). In the selected major area, detailed analyses are carried out in terms of student performance and various educational and economic indicators.

One of the main topics focused on PISA study is equality in education. In this context, the relationship between various socioeconomic and demographic information obtained through questionnaires and literacy performance of students is examined (OECD, 2019a; OECD, 2019b). Equality in education is evaluated academically under two main titles: access to education and quality of education (Ferreira, Gignoux and Aran, 2010; Önder and Güçlü, 2014). Equality in access to education is generally analysed with basic statistics in the field of education such as schooling rates, attendance and dropout rates, distribution of student and school types. Academic achievement studies conducted at the national and international scale provide important findings to measure the impact of school-level characteristics (Hanushek and Wößmann, 2007; Scheerens, 1992).

Achievement differences within- and between schools and the performance of students in different gender groups and socioeconomic levels presented in PISA results are reported in detail (OECD, 2016; OECD, 2019). Therefore, PISA results provide valuable feedback to the participating countries about the educational equality of opportunity as well as the literacy of the students.

The differences arising from school-related factors in terms of literacy evaluated within the scope of PISA are the indicators taken into consideration in terms of equality in education (Eğitimde Reform Girişimi-ERG, 2009; Levin, 2003). Acquiring basic literacy to students regardless of the type of school has vital importance in ensuring educational equality. The fact that school characteristics have a stronger effect on students' academic outcomes than many variables (Greenwald, Hedges and Laine, 1996; Wang, Haertel and Walberg, 1993) requires determining the level of explanation of student performances within- and between-school differences, and detailed studies in which these results are interpreted (OECD, 2007). Results of within- and between-school differences are evaluated in the context of educational equality of opportunities (Inter-American Development Bank, 2012; OECD, 2014). Countries both conduct detailed studies on differences between proficiency levels and focus on the reflections of these differences to school types in literacy areas.

Turkey have participated in PISA regularly since 2003. The fact that the academic achievement differences existed for a long time at the levels of both secondary school and high school is a common finding of national and international studies. Studies which focus on PISA results of Turkey is mostly dependent on mean scores of school types (Albayrak, 2009; Ataş and Karadağ, 2017; Berberoğlu and Kalender, 2005; Çiftçi, 2006; Erdoğan, 2018). However, this is the first comparative study which focuses on the distribution of students to proficiency levels by school types in Turkey. Accordingly, the variation between the student distributions to proficiency levels in PISA applications by school types is examined in this study. Besides mean scores, interpretation of the student distribution to proficiency levels becomes important due to the fact that students at both ends of Turkey's performance scale are high. Therefore, this study is critical because it focuses on literacy performance changes of Turkish students in PISA applications and examines this change on the distribution of students' level of proficiency. The study findings will provide detailed feedback on the change of student ratios with basic and advanced qualifications by years and school types. Findings of the study provide detailed insights

about the variation of student ratios at basic- and advanced level of proficiency by school types and years.

The Achievement Differences between School Types in Turkey

Academic studies have been performed for a long time to identify school-related factors which affect students' academic skills. It has been empirically demonstrated that various factors and family characteristics of schools have had a significant impact on student achievement since the 1960s. In the Coleman report (1966), which is the first example frequently emphasized in this regard, school characteristics were shown to be related to student achievement. Although the advanced statistical and methodological methods commonly used today are not used, the results obtained in the Coleman report have also been confirmed in the studies performed after (Coleman, Hoffer and Kilgore, 1982; Coleman and Hoffer, 1987; Mortimore et al., 1988; Rosenholtz, 1985; Scheerens and Creemers, 1989).

The main reason for simultaneous examining the effects of school and family characteristics on student achievement is that these variables are related. According to Bourdieu (1986), factors such as the condition of the family in the social structure, the resources it has, and the educational level of the family members determine the academic achievement of the students to a considerable extent. The fact that students from more rooted, wealthier and more educated families are also more successful academically, is explained by the concept of social reproduction (Bourdieu, 1986; Bourdieu and Passeron, 2010; Ozer and Perc, 2020). The characteristics of the families can also be effective in the selection of schools where students will continue their education. Therefore, if there are significant differences in academic achievement among these school types, it is possible that the distribution of students to school types is related to family characteristics.

The fact that there are considerable differences between the academic skills of students in different types of schools is shown by academic studies in Turkey for a long time. The results of PISA 2003, which Turkey participated in PISA for the first time, showed that Turkey is the country where the between-school differences explain the student performance ratio at maximum level (OECD, 2007). Çiftçi (2006) showed in PISA 2003 that one of the factors that have a significant effect on Turkish students' mathematical literacy performance is school type. It has been found that students in science high schools, Anatolian high schools and private high schools perform significantly better in mathematics compared to other students. Berberoğlu and Kalender (2005) aimed to determine the academic achievement differences between school types by using the Student Selection Exam (ÖSS) results and PISA results. The findings of the study showed that there were significant and considerable achievement differences between school types in both the ÖSS and PISA context. Alacacı and Erbaş (2010) aimed to determine the effects of school-related and student characteristics on student performance by controlling the family characteristics and demographic characteristics of Turkish students in PISA 2006. The results showed that even when the family and demographic characteristics are controlled, 55% of the variance in student performance is explained by school characteristics. Yalçın and Tavşancıl (2014) analysed the data in three PISA applications between 2003 and 2009 by data envelopment method and examined the school effect on student achievement. In the study, it was determined that the significant performance differences between the school types continued at a similar level in all three applications, the most effective school type among the secondary education institutions was science high schools and the lowest effective school type was the vocational high schools. Albayrak (2009) aimed to determine the variables that affect the science performance of Turkish students in PISA 2006. Findings of the study showed that one of the effective factors on students' literacy performance is the type of school. The science literacy scores of students in science high schools and Anatolian high schools, which accept students with high placement scores, were found to be significantly higher than students in other schools. Özdemir (2016) examined the effect of socioeconomic variables on students' mathematics literacy scores in order to examine the status of the Turkish education system on equality. With PISA 2012 Turkey sample data, results show that type of school is the factor that leads to biggest difference on student performance in mathematics. Erdoğan (2018) and Ataş and Karadağ (2017) analysed PISA 2015

data for Turkey with hierarchical linear modelling and showed that school type has a significant effect on the reading literacy of the students.

The findings on academic achievement differences between school types in Turkey is not limited to international monitoring studies. It is also possible to observe considerable achievement differences between the school types in the monitoring studies performed to assess the academic performances of students and the results of the stage-transition examinations. In High School Entrance Examination (LGS), it is found that the performance of students differentiated significantly by secondary school types and high school types they are placed (Ministry of National Education-Milli Eğitim Bakanlığı-MEB, 2018a). One of the obvious examples of the difference between the academic performances of students in different high schools can be seen in the results of the 2018 University Entrance Examination (Ölçme, Seçme ve Yerleştirme Merkezi-ÖSYM, 2018). It can be seen in the results of earlier versions of University Entrance Examination (ÖSS and ÖYS), which were conducted in 1995 that academic achievement differences have remained in existence for a long time between high school types (Köse, 1999). In the 8th grade application carried out in 2016 within the scope of the Monitoring and Evaluation of Academic Skills (ABİDE) project, it was emphasized that there are significant and considerable differences in all areas between the performances of students in different secondary school types (MEB, 2016). Literacy differences between school types between schools are examined via proficiency distributions of students in PISA rather than mean scores in contrast to other studies.

Proficiency Levels in PISA Studies

In PISA, mean scores, rankings, status according to the OECD average and distribution of students at proficiency levels are used to assess the status of the participating countries in terms of literacy. All of these statistics provide information from different perspectives in terms of students' literacy. However, the distribution of students in their level of proficiency provides more detailed information about the current status of students in terms of literacy compared to other statistics (OECD, 2019a). In countries where there is no significant difference between their mean scores, the distribution of students by their level of proficiency and their mean scores by socioeconomic levels can differ significantly. This situation creates the possibility of ignoring detailed educational indicators only if the focus is on ranking or mean score of countries (Gür, Çelik and Özoğlu, 2012; Ozer, 2020; Woessman, 2016).

Proficiency levels provide a concrete relationship between the scores of students in each literacy field and their cognitive skills in this field (OECD, 2017; OECD, 2019a). According to the scores of students in mathematics, science, and reading, it is determined which level of proficiency they are and what cognitive skills they have in these fields (OECD, 2017). Establishing proficiency levels is an important step in PISA test development processes. Student performances in literacy are assessed on a continuous scale in the fields of mathematics, science, and reading. In addition, creating cut-off points and proficiency levels to define student skills provides concrete feedback to participating countries. Each proficiency level defines the capabilities and skills that students can do in the relevant literacy field. As the proficiency levels are defined to cover a certain score range, it is natural to expect a partial difference between the skills of the students at the lower limit and the upper limit of this range. Despite this, the proficiency levels allow valid predictions about the capabilities and skills of all students at that level (OECD, 2017). As of PISA 2009, six proficiency levels are used in the fields of mathematics, science, and reading literacy (NESC, 2012; OECD, 2019a).

In PISA applications, the second proficiency level is considered to be the minimum level expected to be achieved in order to demonstrate basic skills in the related field (OECD, 2019a). OECD defines the second level of proficiency as "the level that students should reach in order to solve practical problems and use their capacities" (OECD, 2019a, p.89). The second level of qualification is also considered to be the minimum qualification level that every student should achieve in the United Nations Sustainable Development Goals at the secondary education level (OECD, 2019a). It provides important feedback to

the participating countries in terms of the level of students who have a basic level of cognitive skills in mathematics, science, and reading literacy. In fact, the OECD lists the participating countries in PISA reports in addition to their mean scores in terms of student ratio of having basic literacy. The fifth and sixth proficiency levels within the framework of PISA represent the highest level of performance. In this context, the ratio of students at the level of five and sixth proficiency provides vital feedback in terms of the ratio of students at advanced proficiency levels (top performer) within the total. Participating countries are also ranked according to the ratio of students at advanced proficiency levels (OECD, 2016; OECD, 2019a).

Proficiency levels are determined in PISA applications which it is the major field (mathematics, science, and reading) (OECD, 2017). Therefore, proficiency levels in the field of reading were determined in 2000, when the first PISA application was conducted, proficiency levels in mathematics in 2003, and proficiency levels in science in 2006. After defining proficiency levels, they do not remain constant and can be updated throughout PISA applications. For example, in PISA 2003 and PISA 2006 applications, five proficiency levels have been defined in the field of reading literacy. PISA 2009 is the first application in which six proficiency levels are defined in all fields. In PISA 2018, all updates and comparability analyses related to proficiency levels were carried out, and how to make proficiency level comparisons in the most appropriate way was determined again. In line with the results, comparisons were made in the PISA 2018 report between 2003-2018 in the field of mathematics, between 2006-2018 in the field of science, and between 2009-2018 in the field of reading (OECD, 2019a).

Purpose of the Study

The aim of this study is to examine the change in student ratios at basic- and advanced level of proficiency by school types in PISA applications between 2003 and 2018. For this purpose, answers to the following questions were sought:

1. Is there any significant difference between students with basic proficiency ratios by type of school in Turkey in PISA applications between 2003 and 2018?
   1.a. Is there any significant difference between the student ratios at second and higher proficiency levels by school types in PISA applications between 2003 and 2018 in mathematics literacy?
   1.b. Is there any significant difference between the student ratios at second and higher proficiency levels by school types in PISA applications between 2006 and 2018 in science literacy?
   1.c. Is there any significant difference between the student ratios at second and higher proficiency levels by school types in PISA applications between 2009 and 2018 in reading literacy?
2. Is there any significant difference between students with advanced proficiency ratios by type of school in Turkey in PISA applications between 2003 and 2018?
   2.a. Is there any significant difference between the student ratios at the fifth and sixth proficiency levels by school types in PISA applications between 2003 and 2018 in mathematics literacy?
   2.b. Is there any significant difference between the student ratios at the fifth and sixth proficiency levels by school types in PISA applications between 2006 and 2018 in science literacy?
   2.c. Is there any significant difference between the student ratios at the fifth and sixth proficiency levels by school types in PISA applications between 2009 and 2018 in reading literacy?

METHOD

Research Design

This study in which the change of students' distribution at PISA proficiency levels in PISA studies between 2003-2018 by the school types has been performed in the correlational design. In the research, the current situation is examined without any intervention, and this situation reveals the descriptive structure of the study (Karasar, 2005). Comparisons between school types and years lead to the correlational aspect of the study.

Population and Sample

The research population is constituted by students who are 15 and continuing formal education in the years 2003, 2006, 2009, 2012, 2015, and 2018 in Turkey. In PISA applications, students are selected by stratified sampling. Participating countries and economies are expected to identify labels that best represent 15-year-old students (OECD, 2017). The international research centre determines the schools to be applied through random sampling among the schools in the relevant levels. Following the determination of the relevant schools, students studying in these schools are also selected randomly. Schools located in different types of schools in 12 regions covered by Turkey Statistical Region Units Classification (Turkey-İBBS 1) created by socioeconomic level similarity in Turkey are included in the sampling process.

The data of all students in Turkey sample of PISA practices between 2003 and 2018 were used in the research. The number of students participating in the PISA survey between 2003 and 2018 ranged from 4.855 to 6.890 in Turkey. The distribution of students by school type in Turkey sample of PISA applications between 2003 and 2018 is shown in Table 1.

Table 1. Distribution of Students by School Type in PISA Turkey Sample between 2003 and 2008.

School Type	*PISA 2003*		*PISA 2006*		*PISA 2009*		*PISA 2012*		*PISA 2015*		*PISA 2018*	
	f	*%*	*f*	*%*	*f*	*%*	*f*	*%*	*f*	*%*	*f*	*%*
Anatolian High School	3238	66.69	2824	57.14	2659	53.22	2719	56.08	2155	36.56	3013	43.73
Anatolian Fine Arts H. School	-	-	-	-	32	0.64	-	-	40	0.68	42	0.61
Anatolian İmam Hatip High School	-	-	-	-	-	-	-	-	906	15.37	943	13.69
Multi Program Anatolian H. School	-	-	278	5.63	268	5.36	178	3.67	285	4.83	273	3.96
Science High School	63	1.30	35	0.71	100	2.00	35	0.72	40	0.68	291	4.22
Vocational and Technical Anatolian High School	1435	29.56	1689	34.18	1800	36.03	1693	34.92	2268	38.47	2143	31.10
Secondary School	119	2.45	116	2.35	137	2.74	120	2.48	121	2.05	22	0.32
Police College	-	-	-	-	-	-	68	1.40	-	-	-	-
Social Sciences High School	-	-	-	-	-	-	35	0.72	80	1.36	163	2.37
Total	4855	100	4942	100	4996	100	4848	100	5895	100	6890	100

As seen in Table 1, between 2003 and 2018, the change in 15-year-old student population in Turkey has led to changes in the distribution of students within the sample by the school types. Similar to the student population, there were important changes in school types during this period. Despite these changes, in order to make comparisons between school types, existing school types in 2009 and before have been converted to current school types within the scope of the research, as shown in Table 2. The similarity between the old school types and the current school types is taken into consideration in this transformation.

Table 2. Current School Types and Old School Types Before PISA 2015

Old School Type	*Current School Type*
Anatolian Teacher High School	Anatolian High School
General High School	Anatolian High School
Foreign Language Weighted High School	Anatolian High School
Anatolian Vocational High School	Vocational and Technical Anatolian High School
Anatolian Technical High School	Vocational and Technical Anatolian High School
Vocational High School	Vocational and Technical Anatolian High School
Technical High School	Vocational and Technical Anatolian High School

Data Collection Instruments

In the research, reading, mathematics, and science tests applied within the scope of the PISA 2003, PISA 2006, PISA 2009, PISA 2012, PISA 2015, and PISA 2018 research were used. The tests used in the PISA research consist of open-ended, short-answer, and multiple-choice items. Each subtest contains items developed for different proficiency levels. As an indicator of the students' performance in the tests, plausible values are calculated for each student (OECD, 2017). Until the PISA 2015 application, while calculating five possible values from each of the fields of mathematics, science, and reading, the possible values calculated on and after PISA 2015 were increased to ten. As the Turkey samples participating in PISA between 2003 and 2008 were taken into consideration in the study, the first plausible value (1st plausible value), which is calculated as common to all applications, was taken into account.

Data Analysis

In this study, firstly, the proficiency levels of Turkish students in six PISA applications were determined by considering the first plausible values in each field. Then, the ratio of students who have basic proficiency in each PISA application is calculated by adding the student ratios at second and higher proficiency levels. A similar practice was used in the calculation of the students at the advanced proficiency levels by summing the student ratios at the fifth and sixth proficiency levels in each PISA application.

In successive PISA applications, the ratio of changes in the proficiency level distributions was - examined with the z test method for independent sample ratios. The z test is a statistic that is also used in cases where the sample sizes are not equal, and the significance of the difference between the ratios calculated in independent samples is tested (Schumacker, 2015). The aim of the study is to compare the type of school at the secondary level; thus the students at secondary school level in Turkey sample were excluded from the study. Since Anatolian fine arts high school is included in the sample in PISA 2009 and not included in PISA 2012, the changes in this school type were examined only between PISA 2015 and PISA 2018.

RESULTS

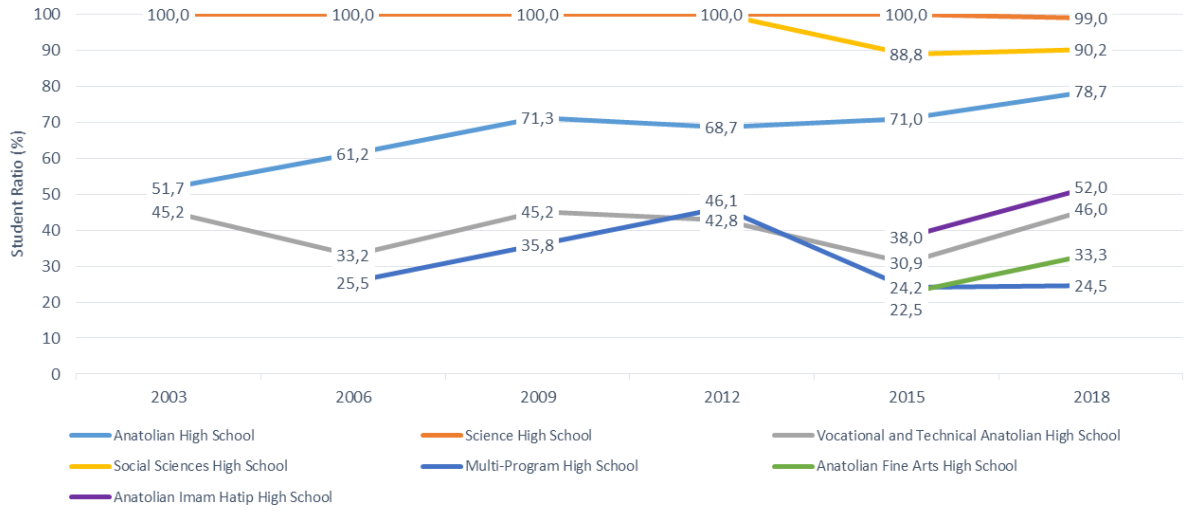
Findings of The First Research Question

Is there a significant difference between students with basic proficiency ratios by type of school in Turkey in PISA applications between 2003 and 2018?

First, findings regarding the sub-question of mathematics literacy, are presented below.

In Graph 1, the distribution of students with basic math proficiency by years and school types in Turkey between 2003 and 2018 PISA applications is given. Table 3 shows the results of the z test regarding the significance of the differences between the ratios given in Graph1.

Graph 1. Distribution of Turkish Students with Basic Mathematics Proficiency in PISA Applications by Years and School Types



As seen in Graph 1, that ratio of students having basic proficiency in mathematical literacy in Turkey shows significant differences from one PISA application to another. School types are categorized into four groups as those who tend to increase according to the performance of the students over the years, those who have a tendency to decrease, those who remain at a similar level and those who show multiple changes.

It is seen that the students whose performance has increased over the years in terms of mathematics literacy performance take education in Anatolian high schools and Anatolian imam hatip high schools. The ratio of students with basic mathematical literacy showed an overall increasing trend in Anatolian high schools between 2003 and 2018, and the ratio, which was calculated as 51.7% in 2003, reached 78.7% in 2018. Similarly, the ratio of students with basic mathematical literacy among the students studying in Anatolian imam hatip high schools increased from 38% in 2015 to 52% in 2018. While the ratio of students with basic mathematics literacy among the students studying in Anatolian fine arts high schools was 22.5% in 2015, this ratio increased to 33.3% in 2018; however, it is found that the increase was not significant.

It was determined that the mathematical literacy performances of the students in social sciences high schools decreased significantly over the years. In PISA 2012 application, despite the fact that all students performed on and above the basic proficiency level in mathematics literacy, the ratio of students with this proficiency in PISA 2015 was 88.8% and in PISA 2018, it was 90.2%.

PISA mathematics literacy performances of students in vocational and technical Anatolian high school and multi-program Anatolian high schools have reached the level in 2003 with significant increases and decreases over the years. The ratio of vocational and technical Anatolian high school students with basic mathematics literacy dropped to 30.9% between 2009 and 2015, then increased again in 2018 and reached 46%. The ratio of multi-program Anatolian high school students with basic mathematical literacy increased significantly between 2006 and 2012, but decreased significantly in 2015. In PISA 2018, it was determined that 24.5% of students studying in multi-program Anatolian high school have basic mathematics literacy and this ratio is very close to the level of 2006.

Science high schools are the only type of school whose performance does not change significantly between PISA 2003 and PISA 2018 applications. The ratio of students with basic mathematical literacy in science high schools varies between 99% and 100%.

Table 3. z-Test Results Regarding the Ratio of Students with Basic Mathematics Literacy in PISA Applications by School Types*

School Type	*2006-2003*	*2009-2006*	*2012-2009*	*2015-2012*	*2018-2015*
Anatolian High School	7.400*	7.904*	-2.081*	1.706	6.311*
Anatolian İmam Hatip High School	-	-	-	-	6.045*
Anatolian Fine Arts High School	-	-	-	-	1.092
Multi Program Anatolian High School	-	2.607*	2.164*	-4.880*	0.091
Science High School	x	x	x	x	-0.645
Vocational and Technical Anatolian High School	6.903*	7.258*	-1.430	-7.721*	10.346*
Social Sciences High School	-	-	-	-2.067*	0.035

**p<0.05*
-: School type not represented in PISA sample
x: Significance test is not performed since there is no ratio change between years.

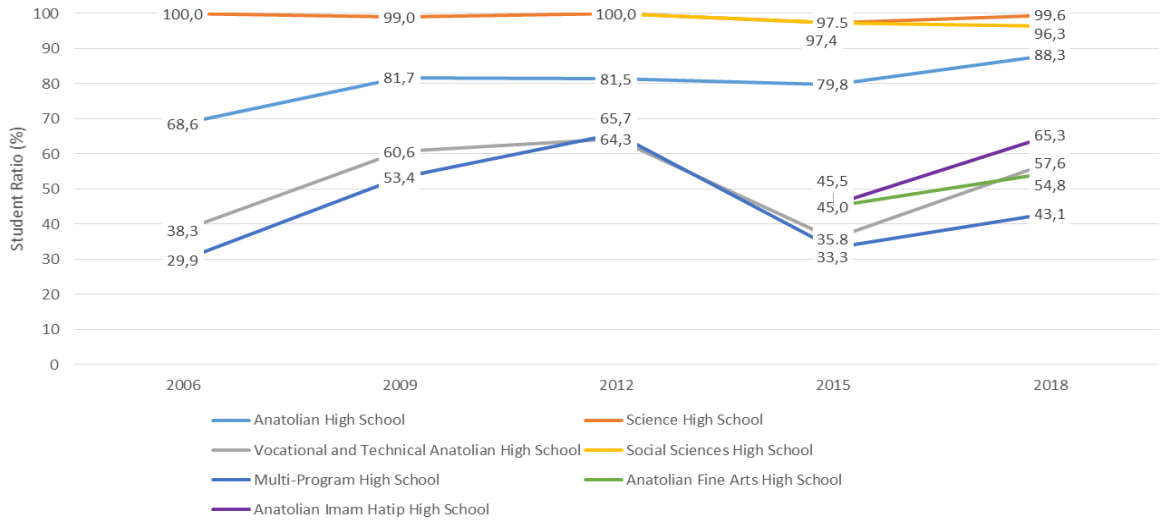
Secondly, findings regarding the sub-question of ***science literacy*** are presented below.

In Graph 2, the distribution of students with basic science proficiency by years and school types in Turkey between 2006 and 2018 PISA applications is given. Table 4 shows the z-test results regarding the significance of the difference between the ratios given in Graph 2.

As seen in Graph 2, the ratios of students having basic science literacy by school types show significant differences from one PISA application to another. The school type with the highest ratio of students with basic science literacy in all five applications between 2006 and 2018 is science high school. Multi-program Anatolian high school is the type of school with the lowest ratio of students reaching basic science literacy in all applications except 2012.

The ratio of students with basic science literacy in Anatolian high schools and Anatolian imam hatip high schools tends to increase. While the ratio of students with basic science literacy in Anatolian high schools in 2006 was 68.6%, this ratio reached 88.3% in 2018. Similarly, the ratio of students with basic science literacy among Anatolian imam hatip high school students was calculated as 45.5% in 2015 and 65.3% in 2018. The ratio of students with basic proficiency in Anatolian fine arts high school increased from 45% to 54.8% in 2018, but it was determined that this increase was not significant.

Graph 2. Distribution of Turkish Students with Basic Science Proficiency in PISA Applications by Years and School Types



The ratio of students with basic science literacy among students studying in science high schools and social sciences high schools does not differ significantly between PISA applications. The ratio of students with basic science literacy in PISA practices between 2006 and 2018 ranged from 97.5% to 100% in science high schools and 96.3% to 100% in social sciences high schools. In other words, almost all students studying in science high schools and social sciences high schools between 2006 and 2018 have basic science literacy.

The ratios of students in vocational and technical Anatolian high schools and multi-program Anatolian high schools having basic science literacy varied in PISA applications between 2006 and 2018. The ratio of students with basic science literacy among the students studying in vocational and technical Anatolian high schools was calculated as 38.3% in 2006, increasing and decreasing over the years, reaching 57.6% in 2018. In multi-program Anatolian high schools, the ratio of students with basic science literacy was calculated as 29.9% in 2006 and reached 43.1% in 2018 after changes in different directions.

Table 4. z-Test Results Regarding the Ratio of Students with Basic Science Literacy in PISA Applications by School Types*

School Type	*2009-2006*	*2012-2009*	*2015-2012*	*2018-2015*
Anatolian High School	10.738*	-0.217	-2.917*	9.381*
Anatolian İmam Hatip High School	-	-	-	8.073*
Anatolian Fine Arts High School	-	-	-	0.884
Multi Program Anatolian High School	5.574*	2.595*	-6.807*	2.491*
Science High School	-0.594	0.594	-0.942	1.650
Vocational and Technical Anatolian High School	13.365*	2.026*	-17.293*	15.063*
Social Sciences High School	-	-	0.944	-0.485

**p<0.05*

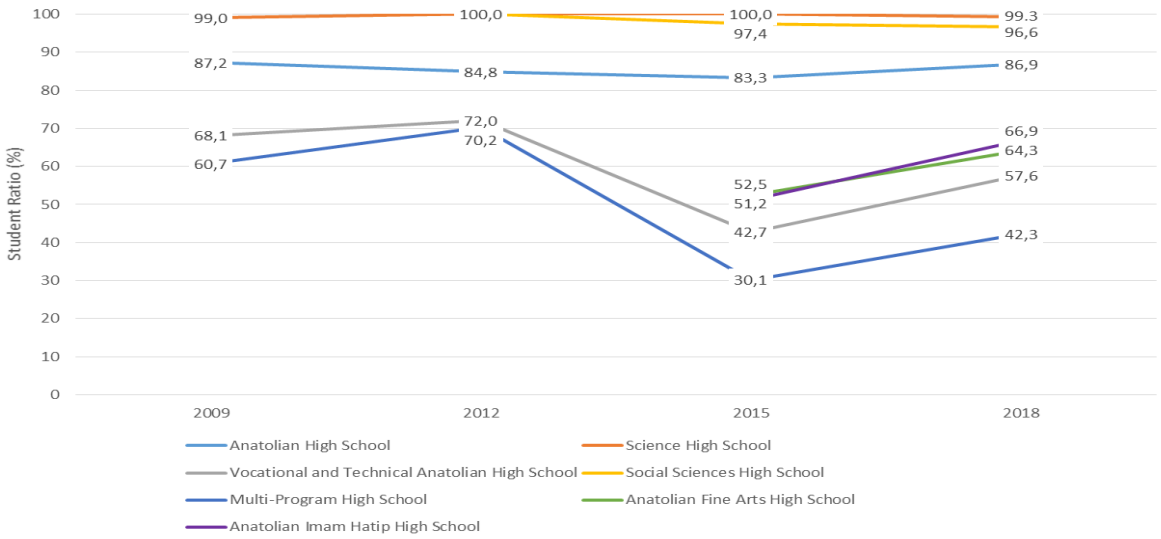
-: The school type was not represented in the PISA sample.

Lastly, findings related to the sub-question of reading literacy are presented below. The distribution of the students having basic reading proficiency in Turkey based on years and school types between 2009 and 2018 is given in Graph 3. Table 5 shows the results of the z test regarding the significance of the difference between the ratios given in Graph 3.

As can be seen in Chart 3, the ratio of students having basic reading literacy by school types shows significant differences from one PISA application to another. It is the school type science high school with the highest ratio of students with basic science literacy in all four PISA applications between 2009 and 2018. Multi-program Anatolian high school is the type of school with the lowest ratio of students reaching basic science literacy level in all applications.

The ratio of students studying at the Anatolian imam hatip high schools tends to increase over the years. The ratio of students with basic reading literacy in this school type was calculated as 51.2% in 2015 and 66.9% in 2018. While 52.5% of Anatolian fine arts high school students had basic literacy in 2015, this ratio reached 64.3% in 2018; however, it is found that this increase was not significant.

Graph 3. The Distribution of Turkish Students with Basic Reading Literacy in PISA Applications by Years and School Types



The ratio of students studying in science and social sciences high schools having basic reading literacy between 2003 and 2018 varies between 96.6% and 100%. In other words, almost all students in science high schools between 2003 and 2018 and social science high schools between 2012 and 2018 have basic reading literacy.

The ratio of having basic reading literacy among the students in vocational and technical Anatolian high schools and multi-program Anatolian high schools has been increasing and decreasing over the years. In PISA 2009, the ratio of vocational and technical Anatolian high school students with basic reading literacy has been calculated as 68.1%, this ratio has decreased to 42.7% in 2015 and reached 57.6% in 2018. While the ratio of students with basic reading literacy among multi-program Anatolian high school students was 60.7% in 2009, this ratio was calculated as 42.3% in 2018. The ratios of multi-program Anatolian high school students with basic reading literacy in this time interval varied considerably, between 30.1% and 70.2%.

Unlike other fields, the ratio of having basic reading literacy among Anatolian high school students did not increase significantly and remained close to 87.2% calculated in PISA 2009.

Table 5. z-Test Results Regarding the Ratio of Students with Basic Reading Literacy in PISA Applications by School Types

School Type	*2012-2009*	*2015-2012*	*2018-2015*
Anatolian High School	-2.255*	-2.332*	3.585*
Anatolian İmam Hatip High School	-	-	5.779*
Anatolian Fine Arts High School	-	-	1.083
Multi Program Anatolian High School	2.111*	-8.264*	2.236*
Science High School	0.594	x	0.492
Vocational and Technical Anatolian High School	2.332*	-18.119*	9.878*
Social Sciences High School	-	-0.944	-0.485

**p<0.05.*

-: School type not represented in PISA sample

x: Significance test is not performed since there is no ratio change between years.

Findings of The Second Research Question

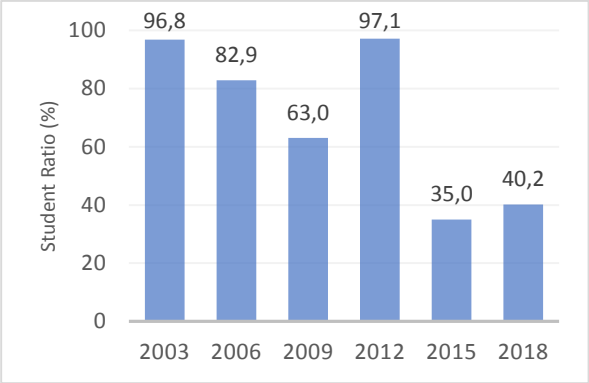
Is there any significant difference between students with advanced proficiency ratios by type of school in Turkey in PISA applications between 2003 and 2018?

Firstly, findings related to sub-question of mathematics literacy are presented below.

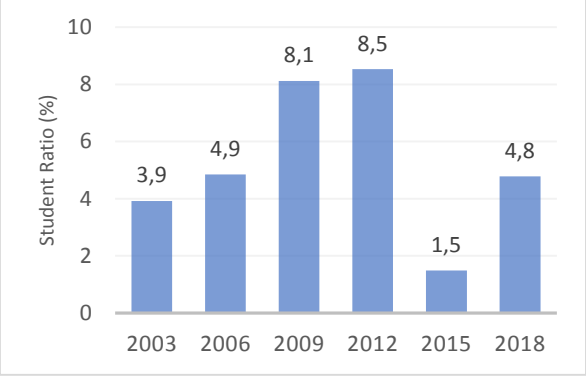
In Graph 4, the distribution of students with advanced maths proficiency by years and school types in Turkey sample between 2003 and 2018 PISA applications is given. Table 6 shows the z-test results regarding the significance of the difference between the ratios given in Graph 4.

Graph 4. The Distribution of Turkish Students with Advanced Mathematical literacy in PISA Applications by Years and School Types

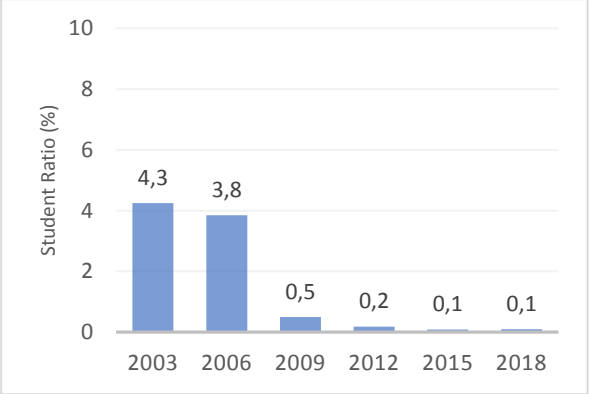
a. Science High Schools



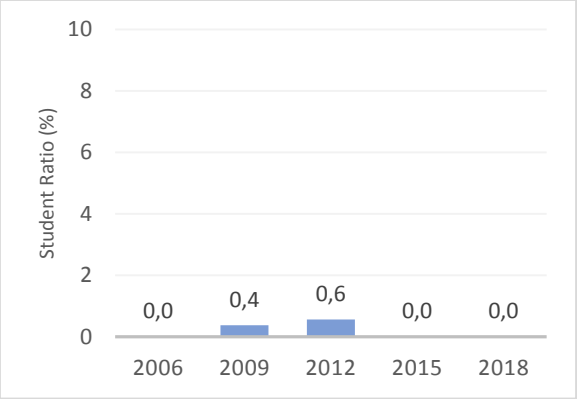
b. Anatolian High Schools



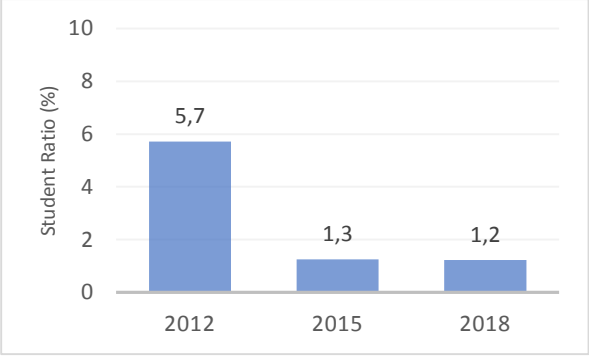
c. Vocational and Technical Anatolian High Schools



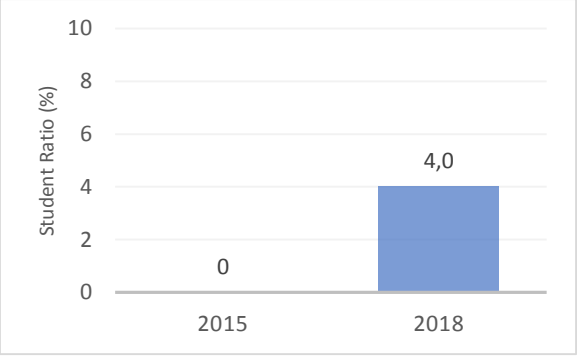
d. Multiple Programs High Schools



e. Social Sciences High Schools



f. Anatolian İmam Hatip High Schools



As seen in Graph 4, there are significant differences between school types in terms of student ratios with advanced mathematical literacy. In addition, it has been determined that school types have significant time-dependent changes in terms of student ratios with advanced mathematical literacy.

Anatolian fine arts high school and multi-program Anatolian high schools constitute the types of schools in which the ratio of students with advanced mathematics literacy is below 1% in all PISA applications. In PISA 2015 and 2018, the proficiency levels of the students in Anatolian fine arts high schools in mathematics literacy range from the sixth level to the fourth level. As a result, it was determined that students in Anatolian fine arts high schools could not reach advanced mathematics literacy proficiency levels. It was determined that 0.4% of multi-program Anatolian high school students in PISA 2009 had advanced mathematics literacy in PISA 2012 and 0.6% in PISA 2012. In PISA 2006, PISA 2015 and PISA 2018, it is seen that students in this high school type do not reach advanced mathematics literacy levels.

According to Graph 4, the ratio of students with advanced mathematics literacy in vocational and technical Anatolian high schools tends to decrease over time. In vocational and technical Anatolian high schools, the relevant ratio was calculated as 4.3% in 2003, and this ratio decreased to 0.1% in 2015 and 2018 applications. While the ratio of students with advanced literacy in mathematics literacy was 5.7% in social sciences high schools in 2012, this ratio was calculated as 1.2% in 2015 and 2018, but it is seen through Table 6 that this decrease is not significant.

The ratio of students with advanced mathematics literacy in Anatolian and science high schools varied between 2003 and 2018. While the ratio of having advanced mathematics literacy among Anatolian high school students in PISA 2013 was 3.9%, this ratio increased up to 8.5% in PISA 2012. The ratio of having advanced mathematics literacy among Anatolian high school students decreased sharply to 1.5% in PISA 2015 and reached 4.8% in PISA 2018 with a significant increase. The ratio of advanced mathematics literacy among students studying in science high schools varies greatly between 35% and 97.1% in different PISA applications. The change is particularly noticeable in PISA applications between 2012 and 2018. While 97.1% of science high school students had advanced mathematics literacy in PISA 2012, this ratio decreased to 35% in 2015 and reached 40.2% in 2018.

In the Anatolian imam hatip high schools, which were included in the sample as a school type for the first time in PISA 2015, students could not reach advanced mathematics literacy levels. However, the ratio of having advanced mathematics literacy among Anatolian imam hatip high school students reached 2.3%, with a significant increase in PISA 2018.

Table 6. z-Test Results Regarding the Ratio of Students with Advanced Mathematics Literacy in PISA Applications by School Types

School Type	*2006-2003*	*2009-2006*	*2012-2009*	*2015-2012*	*2018-2015*
Anatolian High School	1.768	4.934*	0.543	-10.796*	6.438*
Anatolian İmam Hatip High School	-	-	-	-	6.105*
Anatolian Fine Arts High School	-	-	-	-	x
Multi Program Anatolian High School	-	1.019	0.292	-1.267	x
Science High School	-2.420*	-2.170*	3.866*	-5.594*	0.631
Vocational and Technical Anatolian High School	-0.57	-6.860*	-1.63	-0.781	0.057
Social Sciences High School	-	-	-	-1.382	0.015

**p<0.05*

-: School type not represented in PISA sample

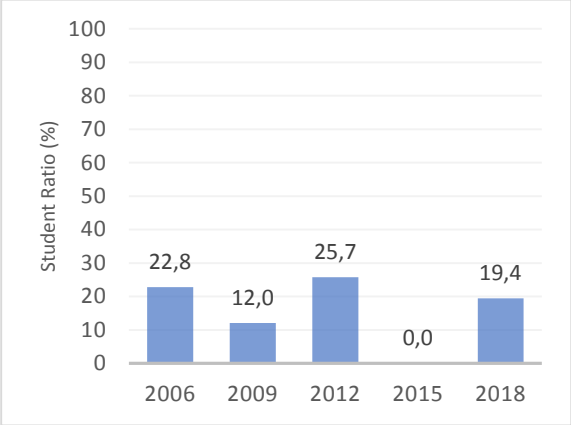
x: Significance test is not performed since there is no ratio change between years.

Secondly, findings related to the sub-question of ***science literacy*** are presented below.

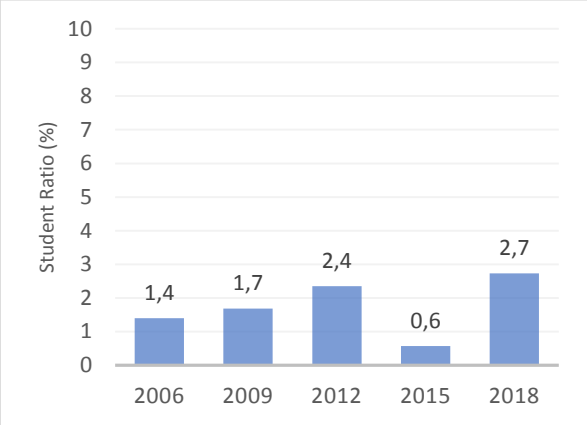
In Graph 5, the distribution of students with advanced science proficiency by years and school types in Turkey between 2006 and 2018 PISA applications is given. Table 7 shows the z-test results of the significance of the difference between the ratios given in Graph 5.

Graph 5. The Distribution of Turkish Students with Advanced Science Literacy in PISA Applications by Years and School Types

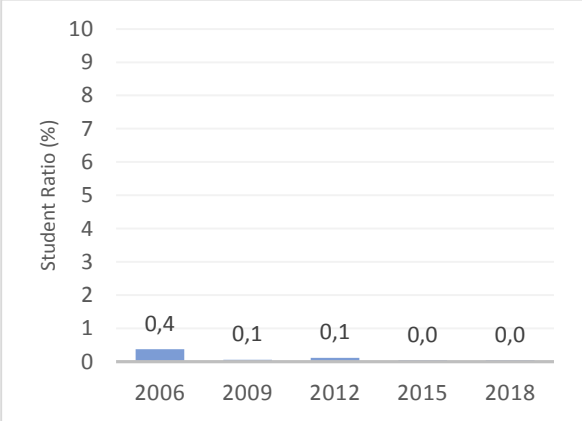
a. Science High Schools



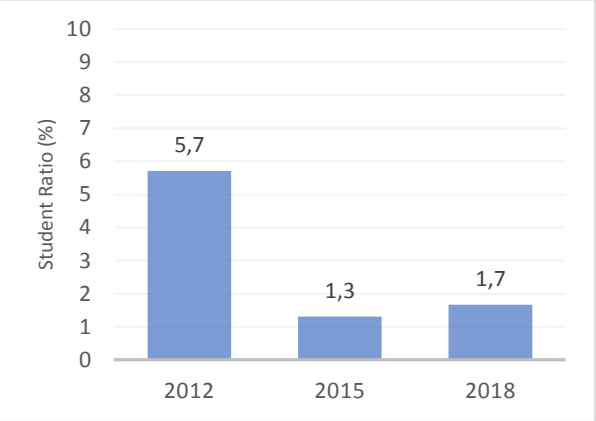
b. Anatolian High Schools



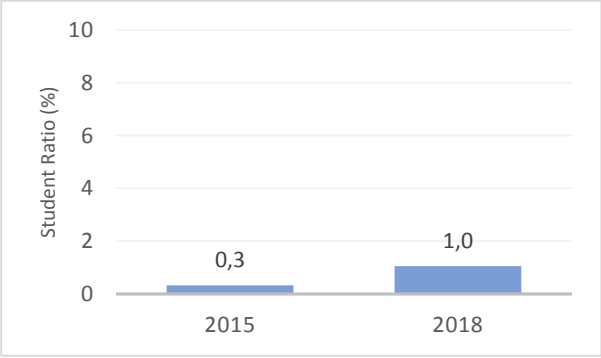
c. Vocational and Technical Anatolian High Schools



d. Social Sciences High Schools



e. Anatolian Imam Hatip High Schools



As can be seen in Graph 5, there are significant differences between the types of schools in terms of the ratio of students with advanced science literacy. It is determined that the ratio of students with advanced science literacy over the years within the school types changed significantly.

Students in multi-program Anatolian high schools and Anatolian fine arts high schools could not reach advanced science proficiency levels in PISA applications between 2006 and 2018. The ratio of students with advanced science proficiency among vocational and technical Anatolian high school students varies between 0.1% and 0.4% in 2006 and 2012 applications. It was determined that the ratio of students with advanced science proficiency among the students in Anatolian imam hatip high schools was 0.3% in 2015 and 1.0% in 2018.

It was determined that the ratio of social science high school students having advanced science proficiency tends to decrease, but the decrease in Graph 5 is not significant.

Anatolian high schools and science high schools are the types of schools where the ratio of students with advanced science literacy differs significantly in different directions. The ratio of students with advanced science literacy among Anatolian high school students varied between 0.6% and 2.7% in 2006 and 2018. Significant changes have also been observed in science high schools in terms of the ratio of students with advanced science literacy. In 2015, science high school students could not reach advanced proficiency in the field of science, and in 2018, 19.4% of the students reached their advanced proficiency levels.

Table 7. z-Test Results Regarding the Ratio of Students with Advanced Science Literacy in PISA Applications by School Types

School Type	*2009-2006*	*2012-2009*	*2015-2012*	*2018-2015*
Anatolian High School	1.388	2.714*	-5.743*	5.290*
Anatolian İmam Hatip High School	-	-	-	3.508*
Anatolian Fine Arts High School	-	-	-	x
Multi Program Anatolian High School	x	x	x	x
Science High School	-1.221	1.587	-3.419*	2.912*
Vocational and Technical Anatolian High School	-3.173*	1.037	-1.268	0.040
Social Sciences High School	-	-	-1.382	0.034

**p<0.05*

-: School type not represented in PISA sample

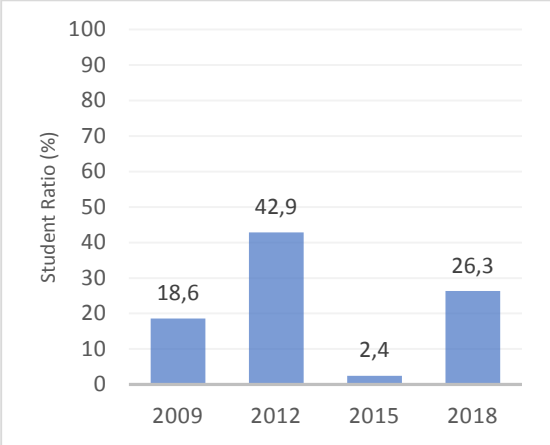
x: Significance test is not performed since there is no ratio change between years.

Lastly, findings related to the sub-question of reading literacy are presented below.

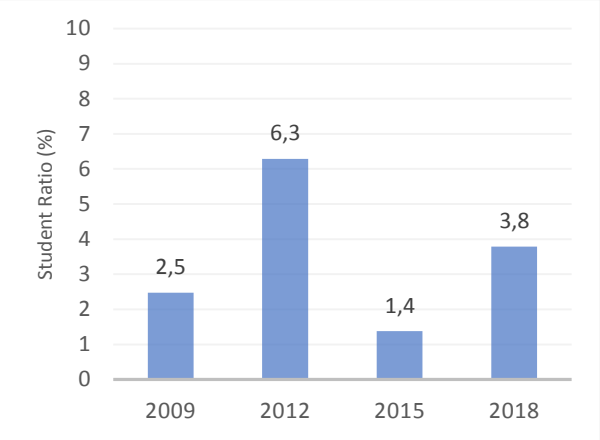
In Graph 6, the distribution of students with advanced reading literacy by years and school types in Turkey between 2009 and 2018 PISA applications is given. Table 8 shows the z-test results regarding the significance of the difference between the ratios given in Graph 6.

Graph 6. The Distribution of Turkish Students with Advanced Reading Literacy in PISA Applications by Years and School Types

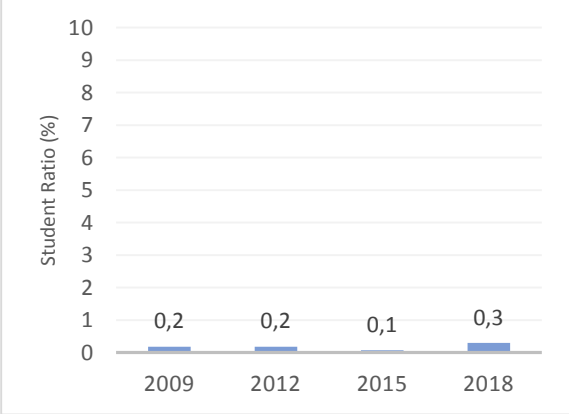
a. Science High Schools



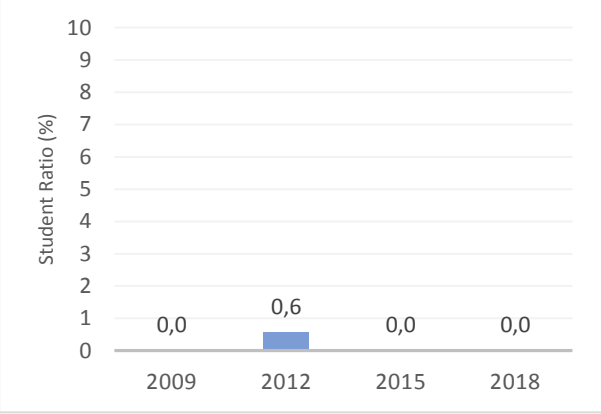
b. Anatolian High Schools



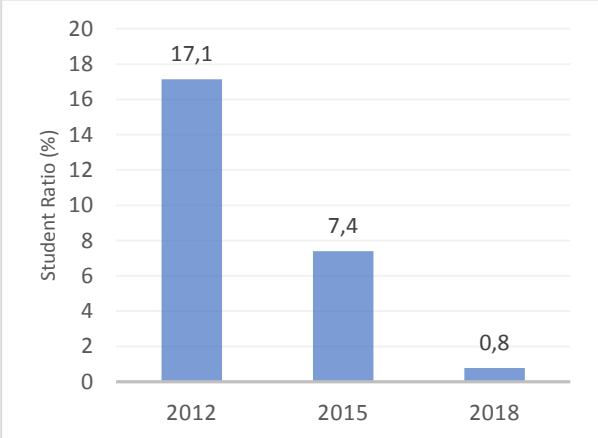
c. Vocational and Technical Anatolian High Schools



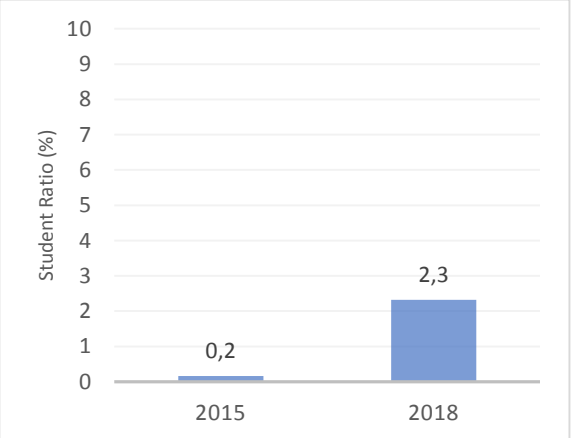
d. Multi-program High Schools



e. Social Sciences High Schools



f. Anatolian Imam Hatip High Schools



According to Graph 6, student ratios of advanced proficiency in PISA reading literacy change significantly between PISA applications by school types. Similarly, there are significant changes of ratios within school types between the PISA applications. School types are categorized into four groups as those who do not show any significant difference from one application to another in terms of the ratio of students with advanced proficiency in reading literacy, those with an increasing trend, those with a decreasing trend and those with multiple changes.

As can be seen in Graph 6, students in Anatolian fine arts high schools could not reach advanced proficiency in reading literacy between 2009 and 2018. In multi-program Anatolian high schools, only 0.6% of students have advanced reading literacy in PISA 2012.

The ratio of students with advanced proficiency in reading literacy among the students in the Anatolian imam hatip high school was calculated as 0.2% in PISA 2015, this ratio increased significantly and reached 2.3% in PISA 2018. In the vocational and technical Anatolian high schools, the ratio of students with advanced reading literacy changed between 0.1% and 0.3% in four PISA applications, and the increase in PISA 2018 was found to be significant.

There was a significant decrease in the ratio of students with advanced reading literacy in the social sciences high schools between PISA 2012 and PISA 2018. The ratio of students having advanced proficiency in reading has decreased from 17.1% to 7.4% in PISA 2015, and from 7.4% to 0.8% in PISA 2018.

Anatolian high schools and science high schools are the types of schools in which there are two-way changes between PISA applications in terms of student ratios with advanced reading literacy. The ratio of those who have advanced reading literacy among Anatolian high school students varies between 1.4% and 6.3%. The ratio of science high school students with advanced reading literacy ranged from 2.4% to 42.9%.

Table 8. z-Test Results Regarding the Ratio of Students with Advanced Reading Literacy in PISA Applications by School Types

School Type	*2012-2009*	*2015-2012*	*2018-2015*
Anatolian High School	6.720*	-8.948*	5.249*
Anatolian İmam Hatip High School	-	-	5.863*
Anatolian Fine Arts High School	-	-	x
Multi Program Anatolian High School	1.228	-1.267	x
Science High School	2.799*	-4.256*	3.013*
Vocational and Technical Anatolian High School	0.075	-1.305	2.204*
Social Sciences High School	-	-1.556	-3.016*

**p<0.05*

-: School type not represented in PISA sample

x: Significance test is not performed since there is no ratio change between applications.

DISCUSSION and CONCLUSION

Turkey participates in the PISA studies regularly since the year of 2003. It is emphasized in both national and international reports that the performance of Turkey is on an increasing trend between PISA 2003 and PISA 2012 (MEB, 2010; MEB, 2013; MEB, 2019a, OECD, 2019a). However, the performance of Turkey decreased dramatically in PISA 2015 in all literacy fields. It is reasonable to infer that possible reasons for this decrease are low-representatives of PISA 2015 sample in terms of school type distribution which can be seen in Table 1, and computer-based application of PISA in Turkey for the first time in PISA 2015. On the other side, Turkey is one of the three country which increases its performance significantly in all literacy fields. Also, the mean scores of Turkey reached their maximum

levels in science and mathematics since PISA 2003. It is emphasized by OECD that the increasing trend of performance of Turkey continues in PISA 2018, and the decrease in PISA 2015 is considered as an "anomaly" (OECD, 2019a). Therefore, Turkey continues to improve literacy performance in PISA despite the growing population of 15-years-olds (OECD, 2019a).

Between-school and within-school academic achievement differences are important elements evaluated in the framework of equal opportunities in education. Regardless of the type of school in which they are, providing the students with the necessary opportunities to gain the expected cognitive skills is an important step taken to ensure equal opportunities in education systems (Önder and Güçlü, 2014; Turan, Açıkalın and Şişman, 2007). Huge achievement differences among schools lead to decrease in homogeneity within schools, and thus, low-performing students cannot have academic support which they need (Lavy, Paserman and Schlosser, 2011; Mendolia, Paloyo and Walker, 2018). So it is the ideal that there are no huge differences between schools and students with diverse academic performance levels take education within schools together. Educating the students with heterogenic academic performance levels within schools also increases the contribution of peer-education to academic achievement (Brunello, 2004; Hanushek and Woessmann, 2006; Ozer and Perc, 2020). In this case, students can choose the type of school they will continue their education in line with their interests and abilities rather than a career path or employment opportunities. Also, in this case, the pressure of the examinations and methods used in determining the schools in which students will continue their education have a low level on education systems. In the 2023 Education Vision announced in 2018, the Ministry of National Education has determined to reduce the differences in success among schools as one of the main goals (MEB, 2018b).

Differences in academic achievement between-schools and within-schools has long been a controversial issue in Turkey (Alacacı and Erbaş, 2010; Albayrak, 2009; Ataş and Karadağ, 2017; Berberoğlu and Kalender, 2005; Çiftçi, 2006; Erdoğan, 2018; Köse, 1999; Özdemir, 2016; Yalçın and Tavşancıl, 2014). In Turkey, by increasing the diversity and number of students in secondary schools, it has been tried many different models in the transition to secondary school. Despite the diverse cross-level transition systems applied, academic achievement differences between school types continue to exist significantly. In the studies conducted, it is seen that the differences in academic achievement between school types begin to occur at the secondary school level, and these differences continue to increase in secondary education (MEB 2016; MEB, 2018a; ÖSYM, 2018). Therefore, academic achievement differences between school types are the result of a cumulative process, not a single educational level.

In this study, changes in student performance by school types in Turkey on PISA study is examined. In order to examine the differences in performance among school types in more detail, the distribution of students to proficiency levels, one of the most important outputs of PISA study, was used. In this context, the change in the PISA applications of student ratios with basic literacy level (the ratio of students in the second and higher level of proficiency) and advanced literacy level (the ratio of the students in the fifth and sixth level of proficiency) in each school type is examined.

The results of this study showed that in all of the applications between 2003 and 2018 when Turkey attended PISA, there are significant differences between types of school in terms of student proficiency levels. In all three fields, almost all science high school and social science high school students have reached basic proficiency levels. Even in PISA 2015, where the performance decrease was observed in other school types, there was no significant decrease in the ratio of students with basic literacy among science high school students. Findings related to science high schools and social sciences high schools show that almost all students in these high schools have basic literacy in all three fields, regardless of the structure of the transition systems.

The ratio of students with basic proficiency among Anatolian high school students showed a significant increase in mathematics and science among PISA applications and remained close to reading literacy in 2009. The findings show that after PISA 2015, when Anatolian imam hatip high school and Anatolian fine arts high school students were included in the sample, school types were collected in two groups.

The first group includes science high school, social sciences high school, and Anatolian high school with more than 70% of students having basic proficiency in all three fields in PISA 2015 implementation and afterwards. In the second group, there are vocational and technical Anatolian high schools and multi-program high schools, where the ratio of students with basic proficiency is lower. The access of students to basic literacy from these two school types showed significant and remarkable changes in both directions.

Between PISA 2015 and PISA 2018 applications, of which they are included in the sample, there has been a tendency to increase the access to the basic proficiency level of students in Anatolian imam hatip high schools and Anatolian fine arts high schools. The increase in ratios of students with basic literacy proficiency in mathematics and science in Anatolian imam hatip schools is remarkable (14% and 19.8%, respectively). Additionally, it is found that the ratio of students in Anatolian imam hatip high schools with advanced proficiency increased significantly in all literacy fields in PISA 2018. Therefore, the ratio of students in Anatolian imam hatip high schools with both basic- and advance proficiency increased significantly in all literacy fields in PISA 2018.

On the other hand, the increases in Anatolian fine arts high schools have not reached a significant level yet. In the future PISA applications, the longitudinal evaluations about the performance of students in these school types will be made after the new PISA applications.

Academic achievement differences between school types become clearer when the ratio of students at an advanced level in terms of literacy is examined. Science high schools perform considerably higher than other school types in terms of student ratios with advanced literacy. Although social sciences high schools and science high schools are similar in terms of students with basic literacy proficiency, they differ greatly in terms of students with advanced literacy proficiency. In PISA 2018, the ratio of students with advanced literacy proficiency in Anatolian high schools is higher in all three areas compared to social science high schools.

Among the students who are in multi-program Anatolian high school and vocational and technical Anatolian high schools, the ratio of students with advanced proficiency is below 1% in all three fields. It is noteworthy that the ratio of students who have advanced mathematics literacy among vocational and technical Anatolian high school students decreased from 3.8% to 0.5% in PISA 2009 application and then showed a downward trend. Among the types of schools which participated in the sampling of PISA 2015, it was observed that Anatolian fine arts high school students could not reach advanced literacy proficiency in all three areas. Another important finding is that the ratio of students with advanced literacy proficiency among Anatolian Imam High School students increased significantly in all three areas in PISA 2018.

It is an important finding that the ratio of students with basic proficiency in all three literacy fields is lower than 60% in vocational and technical Anatolian high schools and multi-program Anatolian high schools. Among the most important indicators of the achievement difference among the school types are the fact that the student ratios at advanced proficiency levels in these school types are below 1% and even in some PISA applications, no student can reach the advanced proficiency levels.

The huge achievement differences between science high schools, social sciences high schools, and other high school types strengthen the opinion that these differences are directly related to student input. With school tracking at an early age in Turkey, students are involved in a process which is quite decisive for life and career training. In this process, students tend to be grouped in school types according to their academic achievement levels and indirectly their socioeconomic levels (Özdemir, 2016; Ozer and Perc, 2020). As a result of this situation, there is a very heterogeneous distribution among school types in terms of academic achievement and student behavior. For example, high school dropout and high absenteeism ratios in vocational and technical Anatolian high schools compared to other school types affect student performance (Ozer, 2018; Ozer, 2019a).

In order to reduce the achievement differences between school types, it is necessary to support low performing school types academically, socially and financially. In the current situation, it is observed

that the opportunities transferred to schools with higher academic achievement such as science high school and social sciences high school are higher (Özdemir, 2016). In this sense, it is important to support schools with lower achievements in terms of teacher quality and financial resources, and to make positive discrimination when necessary (Ozer, 2020). Thus, the areas of development of students can be determined in low-achieving school types and intervention can be carried out in a short time.

In the context of Turkey's Education Vision 2023, numerous projects such as Turkish-Mathematics-Science Student Monitoring Study (TMF-ÖBA) (MEB, 2019b), Supporting Program in Elementary Schools (İYEP), and the steps to strengthen vocational and technical education (VET) in Turkey are conducted to minimize the academic achievement differences between school types. Within the scope of VET, increasing the collaboration between MoNE and sectors, establishing the balance of supply-demand chain on a rational base, increasing accessibility of VET via recently established online platforms, selecting high performing students (at 1% of achievement level) to VET institutions are some of the examples for steps to strengthen VET by MoNE (Ozer, 2019b; Ozer and Suna, 2019; Ozer and Suna, 2020). It is suggested to take steps that increase the academic heterogeneity within the schools and to begin these implementations with schools with high performing students. With increasing heterogeneity within schools, disadvantaged students can have the academic support they need, and peer-education can increase its positive effect on these students' learning process.

REFERENCES

Australian Council for Educational Research (2014). *Australian students' readiness for study, work and life in the digital age: ICILS 2013.* Australia: ACER Publishing.

Alacacı, C., & Erbaş, A. K. (2010). Unpacking the inequality among Turkish schools: Findings from PISA 2006. *International Journal of Educational Development, 30,* 182-192.

Albayrak, A. (2009). *PISA 2006 sınavı sonuçlarına göre Türkiye'deki öğrencilerin fen başarılarını etkileyen bazı faktörler.* (Yayımlanmamış Yüksek Lisans Tezi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü).

Ataş, D., & Karadağ, Ö. (2017). An analysis of Turkey's PISA 2015 results using two-level hierarchical linear modelling. *Journal of Language and Linguistic Studies, 13*(2), 720-727.

Berberoğlu, G., & Kalender, İ. (2005). Öğrenci başarısının yıllara, okul türlerine, bölgelere göre incelenmesi: ÖSS ve PISA analizi. *Eğitim Bilimleri ve Uygulama, 4*(7), 21-35.

Bourdieu, P. (1986) The forms of capital. In J. Richardson (Ed.) *Handbook of theory and research for the sociology of education* (New York, Greenwood), 241-258.

Bourdieu, P., & Passeron, J. C. (2010). *Reproduction in education, society and culture.* London: Sage Publications.

Brunello, G. (2004). *Stratified or comprehensive? Some economic consedirations on the design of secondary education.* CESifo DICE Rep 4:7-10.

Coleman, J. et al. (1966). *Equality of educational opportunity.* Washington D. C.: U. S. Government Printing Office.

Coleman, J., Hoffer, T., & Kilgore, S. (1982). Cognitive outcomes in public and private schools. *Sociology of Education, 55*(2-3), 65–76.

Coleman, J., & Hoffer, T. (1987). *Public and private high schools: The impact of communities*. New York: Basic Books.

Çiftçi, A. (2006). *PISA 2003 sınavı matematik alt testi sonuçlarına göre Türkiye'deki öğrencilerin başarılarını etkileyen bazı faktörlerin incelenmesi.* (Yayınlanmamış Yüksek lisans tezi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü).

Darling-Hammond, L. (2014). What can PISA tell us about US education policy?. *New England Journal of Public Policy, 26,* 1. Erişim adresi: http://scholarworks.umb.edu/nejpp/vol26/iss1/4.

Erdoğan, E. (2018). *Uluslararası öğrenci değerlendirme programında öğrencilerin sosyoekonomik özellikleri ile okuma becerileri arasındaki ilişki.* (Yayımlanmamış Yüksek Lisans Tezi. Trakya Üniversitesi Sosyal Bilimler Enstitüsü).

Eğitimde Reform Girişimi (2009). *Eğitimde eşitlik: Politika analizi ve öneriler.* ERG Raporları. Retrieved from www.egitimreformugirisimi.org/wp-content/uploads/2017/03/Egitimde_Esitlik_Politika_Analizi_ve_Oneriler_1.pdf.

Ferreira, F. H. G., Gignoux, J., & Aran, M. (2010). *Measuring inequality of opportunity with imperfect data the case of Turkey.* The World Bank Policy Research Working Paper 5204.

Greenwald, R., Hedges, L. V., & Lane, R. D. (1996). The effects of school resources on student achievement. *Review of Educational Research, 66,* 361-396.

Gür, B.S., Çelik, Z., & Özoğlu, M. (2012). Policy options for Turkey: A critique of the interpretation and utilization of PISA results in Turkey. *Journal of Education Policy, 27*(1), 1-21.

Hanushek, E.A., & Woessmann, L. (2006). Does educational tracking affect performance and equality? Differences-in-differences evidence across countries. *Economic Journal,* 116, 63-76.

Hanushek, E.A., & Woessmann, L. (2007). *The role of education quality in economic growth.* World Bank Policy Research Working Paper. 4122.

Hopfenbeck, T.N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J.A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research, 62*(3), 333-353.

Inter-American Development Bank (2012). *Assessing educational equality and equity with large-scale assessment data: Brazil as a case study*. IDB Tehcnical Notes No. IDB-TN-389. Retrieved from https://publications.iadb.org/publications/english/document/Assesing-Educational-Equality- and-Equity-with-Large-Scale-Assessment-Data-Brazil-as-a-Case-Study.pdf

International Association for the Evaluation of Educational Achievement (2010). *ICCS 2009 international report: Civic knowledge, attitudes, and engagement among lower secondary school students in 38 countries.* Amsterdam: IEA Publishings.

Karasar, N. (2005). *Bilimsel araştırma yöntemi.* Ankara: Nobel Yayın Dağıtım.

Köse, M. R. (1999). Üniversiteye giriş ve liselerimiz. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 15*, 51-60.

Lavy, V., Paserman, M., & Schlosser, A. (2011). Inside the black box of ability peer effects: Evidence from variation in the proportion of low achievers in the classroom. *The Economic Journal, 122*(559), 208-237.

Levin, B. (2003). *Approaches to equity in policy for lifelong learning.* OECD Equity in Education Thematic Review Paper. Retrieved from https://www.oecd.org/education/school/38692676.pdf.

Malik, R. S. (2018). Educational challenges in 21st century and sustainable development. *Journal of Sustainable Development Education and Research, 2*(1), 9-20.

Milli Eğitim Bakanlığı (2010). *PISA 2006 projesi: Ulusal nihai rapor.* Ankara: MEB Yayınları.

Milli Eğitim Bakanlığı (2013). *PISA 2012 ulusal ön raporu.* Ankara: MEB Yayınları.

Milli Eğitim Bakanlığı (2016). *Akademik becerilerin izlenmesi değerlendirilmesi (ABİDE) 2016: 8. sınıf raporu*. Ankara: MEB Yayınları.

Milli Eğitim Bakanlığı (2018a). *2018 Liselere geçiş sistemi (LGS): Merkezi sınavla yerleşen öğrencilerin performansı.* Eğitim Analiz ve Değerlendirme Raporları Serisi No:3. Ankara: MEB Yayınları.

Milli Eğitim Bakanlığı (2018b). *2023 eğitim vizyonu.* Ankara: MEB Yayınları.

Milli Eğitim Bakanlığı (2019a). *PISA 2018 Türkiye raporu*. Ankara: MEB Yayınları.

Milli Eğitim Bakanlığı (2019b). *Türkçe-Matematik-Fen Bilimleri Öğrenci Başarı İzleme Araştırması (TMF-ÖBA)-I: 2019 4. sınıf seviyesi.* Eğitim Analiz ve Değerlendirme Raporları Serisi No:9. Ankara: MEB Yayınları.

Mendolia, S., Paloyo, A., & Walker, I. (2018). *Heterogeneous effects of high school peers on educational outcomes.* Oxford Economic Papers. Retrieved from http://dx.doi.org/10.1093/oep/gpy008.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *School matters.* Berkeley, CA: University of California Press.

National Economic & Social Council (2012). *Understanding PISA and what it tells us about educational standards in Ireland.* NECS Secretaritat Papers No:2. Retrieved from https://www.nesc.ie/publications/nesc-secretariat-paper-02-2012-understanding-pisa-and-what-it-tells-us-about-educational-standards-in-ireland/.

Organisation for Economic Cooperation and Development (2007). *Reviews of national policies for education: Basic education in Turkey.* Paris: OECD Publishing.

Organisation for Economic Cooperation and Development (2014). *Learning for tomorrow's world: First results from PISA 2003.* Paris: OECD Publishing.

Organisation for Economic Cooperation and Development (2016). *PISA 2015 results: Excellence and equity in education – Volume I.* Paris: OECD Publishing.

Organisation for Economic Cooperation and Development (2017). *PISA 2015 technical report.* Paris: OECD Publishing.

Organisation for Economic Cooperation and Development (2019a). *PISA 2018 results: What student know and can do - Volume I.* Paris: OECD Publishing.

Organisation for Economic Cooperation and Development (2019b). *PISA 2018 results: Where all students can succeed – Volume II.* Paris: OECD Publishing.

Ozer, M. (2018). 2023 eğitim vizyonu ve mesleki ve teknik eğitimde yeni hedefler [The 2023 education vision and new goals in vocational and technical education]. *Yükseköğretim ve Bilim Dergisi, 8*(3), 425–435.

Ozer, M. (2019a). Mesleki ve Teknik eğitimde sorunların arka planı ve Türkiye'nin 2023 Eğitim Vizyonunda çözüme yönelik yol haritası [Background of problems in vocational education and training and its road map to solution in Turkey's education vision 2023]. *Yükseköğretim ve Bilim Dergisi, 9*(1), 1–11.

Ozer, M. (2019b). Reconsidering the fundamental problems of vocational education and training in Turkey and proposed solutions for restructuring. *İstanbul Üniversitesi Sosyoloji Dergisi, 39*(2), 1–19.

Ozer, M., & Suna, H.E. (2019). Future of vocational and technical education in Turkey: Solid steps taken after Education Vision 2023. *Eğitim ve İnsani Bilimler Dergisi: Teori ve Uygulama, 10*(20), 166-192.

Ozer, M. (2020). *What PISA tells us about performance of education systems? Bartın University Journal of Faculty of Education, 9*(2), 217-228.

Ozer, M., & Perc, M. (2020). Dreams and realities of school tracking and vocational education. *Palgrave Communications, 6*, 34.

Ozer, M., & Suna, H. E. (2020). The linkage between vocational education and labor market in Turkey: Employability and skill mismatch. *Kastamonu Education Journal, 28*(2), 558-569.

Önder, E., & Güçlü, N. (2014). İlköğretimde okullar arası başarı farklılıklarını azaltmaya yönelik çözüm önerileri. *Eğitim Bilimleri Dergisi, 40,* 109-132.

Ölçme, Seçme ve Yerleştirme Merkezi (2018). *2018 YKS değerlendirme raporu.* Değerlendirme Raporları Serisi No:9. Ankara: ÖSYM Yayınları.

Özdemir, C. (2016). Equity in the Turkish education system: A multilevel analysis of social background influences on the mathematics performance of 15-year-old students. *European Educational Research Journal, 15*(2), 193–217.

Rosenholtz, S. J. (1985). Effective schools: Interpreting the evidence. *American Journal of Education, 93*, 352–388.

Rutkowski, D., Rutkowski, L., & von Davier, M. (2014). A brief introduction to modern international large-scale assessment. In Rutkowski, L., von Davier, M., & Rutkowski, D. *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis.* CRC Press, Taylor & Francis, pp 3-11.

Scheerens, J. (1992). Evaluating non-cognitive aspects of education. In Vedder. P. (ed.) *Measuring the quality of education.* Amsterdam: Swet & Zeitlinger Inc.

Scheerens, J., & Creemers, B. P. M. (1989). Conceptualizing school effectiveness. *International Journal of Educational Research, 13*, 691–706.

Schumacker, R. E. (2015). *Learning statistics using R.* California: SAGE Publications.

Thomson, S. (2019). *Assessing and understanding social and emotional skills: The OECD Study on Social and Emotional Skills*. ACER Conference Paper. Retrieved from https://research.acer.edu.au/cgi/viewcontent.cgi?article=1354&context=research_conference

Turan, S., Açıkalın, A., & Şişman, M. (2007). *Bir insan olarak müdür.* Ankara: Pegem Yayıncılık.

Wang, M.C., Haertel, G.D., & Walberg, H.J. (1993). Toward a knowledge base for school learning. *Review of Educational Research, 63*, 249-294.

Woessmann, L. (2016). The importance of school systems: Evidence from international differences in student achievement. *Journal of Economic Perspectives, 30*(3), 3-32.

Yalçın, S., & Tavşancıl, E. (2014). The comparison of Turkish students' PISA achievement levels by year via data envelopment analysis. *Educational Sciences: Theory and Practice, 14*(3), 961- 968.