Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN:1309-6575 *Güz* 2020 *Cilt: 11- Sayı: 3*
Autumn 2020 *Volume: 11- Issue: 3*

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 (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 AYBEK (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.)
Fulya BARIŞ PEKMEZCİ (Bozok 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 (Eylül 2020, Sayı: 11-3)

Journal of Measurement and Evaluation in Education and Psychology (September 2020, Issue: 11-3)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(3); 219-242

How does the ICT Access and Usage Influence Student Achievement in PISA 2009 and 2012? *

Gülfem Dilek YURTTAŞ KUMLU ** Nuri DOĞAN ***

Abstract
The purpose of this study is to investigate the effects of access and usage of information and communication technologies (ICT) on Turkish students' mathematics achievement implemented in PISA 2009 and PISA 2012. A correlational research model was used in this study. In this study, the data which were obtained from the PISA 2009 and PISA 2012 mathematics achievement tests and from the information and communications technologies familiarity questionnaire (ICTFQ) in Turkey were used. In this study, three student level variables and two school variables of ICTFQ which are common indexes both in PISA 2009 and PISA 2012 were selected to compare the effect of ICT variables on PISA mathematics achievement implemented in different years. Two-level Hierarchical Linear Modeling (HLM) analysis was performed in the analysis of the data. As a result, the student level variables had a small or a trivial effect on mathematics achievement. The effect size value of the ENTUSE variable was similar in the PISA 2009 and the PISA 2012 implementation, but the effect size value of the HOMSCH variable and the ICTHOME variable on mathematics achievement in PISA 2012 was lower than in PISA 2009. The ICTSCH and the USESCH variables at the school level had a large effect on mathematics achievement in two implementations of PISA 2009 and PISA 2012. The effect size value of the ICTSCH variable on mathematics achievement in PISA 2012 was higher than in PISA 2009. The effect size value of the ICTSCH variable, having a negative relationship with mathematics achievement in PISA 2012, was lower than in PISA 2009. In this study, the explained variance ratio of mathematics achievement by the school ICT variables level was greater than by the student ICT variables level.

Key Words: Information and Communication Technologies (ICT), mathematics achievement, PISA 2009, PISA 2012, two-level hierarchical linear models.

INTRODUCTION

Today, the perspective of learning mathematics has been involved five standards which are related to conceptual understanding, problem solving, mathematical thinking and reasoning, communicating, making realistic plans for the future and applying these plans (National Council of Teachers of Mathematics-NCTM, 2000, 2014). This viewpoint is consistent with PISA (Programme for International Student Assessment) mathematics literacy defined by OECD (Organization for Economic Cooperation and Development) (2013, 2017) as "using mathematical concepts, processes, and devices to define, explain and guess reasoning mathematically." (p. 17, p. 15). However, mathematics, consisting of sequential abstractions and generalization processes of various structures and connections (Alakoç, 2003), is one of the aspects of lessons which makes learning and comprehension skills difficult for students (Akın & Cancan, 2007; Alakoç, 2003; Murphy, 2016). Technology is one of the applications that will enable students to understand mathematics and to see the usage of mathematics in real life properly (Murphy, 2016). "The information and communication technologies (ICT) include the usage of dynamic mathematics/geometry software, Excel program, manipulative geometric shapes, internet resources (web site, animation, tutorial web applications, video, etc.)" (Ural 2015, p. 94) for developing mathematical teaching. These information and communication technologies contribute to students to learn mathematical concepts easily, to concrete

* This paper was a part of thesis was produced from the first author's master thesis.
** PhD., Sinop University, Faculty of Education, Sinop-Turkey, gdyurttas@sinop.edu.tr, ORCID ID: 0000-0003-4741-2654
*** Prof. PhD., Hacettepe University, Faculty of Education, Ankara-Turkey, nurid@hacettepe.edu.tr, ORCID ID: 0000-0001-6274-2016

To cite this article:
Yurttaş-Kumlu, G. D., & Doğan, N. (2020). How does the ICT access and usage influence student achievement in PISA 2009 and 2012? *Journal of Measurement and Evaluation in Education and Psychology, 11*(3), 219-242. doi: 10.21031/epod.581379.

Received: 23.06.2019
Accepted: 21.02.2020

the concepts, to solve the problems, to think critically and creatively (Alakoç, 2003; Barkatsas, Kasimatis, & Gialamas, 2009; Jang 2009; Lazakidou & Retails 2010; McMahon 2009; Murphy, 2016; Pamuk, Çakır, Ergun, Yılmaz, & Ayas, 2013; Shaikh & Khoja, 2011; Ural, 2015; Yorgancı & Terzioğlu, 2013; Yusuf & Afolabi, 2010; Zengin, Kağızmanlı, Tatar, & İşleyen, 2013). The information and communication technologies are important for using in mathematical teaching because of these features (Ural 2015). Also, the usage of the information and communication technologies are included in the curriculum of elementary school mathematics lessons which were updated in 2013 by the Ministry of National Education in the context of Turkey (Ministry of National Education-MEB, 2013a).

Many countries have heavily invested in ICT infrastructure to adopt implementing ICT-related policies (De Witte & Rogge, 2014; Skryabin, Zhang, Liu, & Zhang, 2015). The reason for adopting ICT-related policies for usage of ICT in education is to improve students' 21st-century competencies (Anderson, 2008; Kim, Kil, & Shin, 2014; Scheuermann & Pedró, 2009). Due to the importance of the integration of ICT into education, the OECD also conducts various studies on the usage of ICT at the international level. The goal of these studies is to evaluate the education policies of countries and to compare them with each other (Bilican-Demir & Yıldırım, 2016). PISA is one of the large-scale assessments to evaluate students' knowledge and skills at the national and international level (OECD, 2014b). Also, PISA examines the causes and factors affecting the student's achievement at national and international levels and provides scientific data for evaluating curriculum and designing appropriate educational settings (Acar, 2012; Bilican-Demir & Yıldırım, 2016).

Recently, especially the studies of the relationship between ICT and academic achievement have increased in large-scale international assessments (OECD, 2014b; Skryabin, et al., 2015; Şengül & Demir, 2018). When the studies are reviewed to determine the relationship between ICT-based learning, teaching, and achievement, it has been especially found that there is an inconclusive relationship between ICT and mathematics achievement. Also, the results of different studies are inconsistent with one another. It was concluded that there was little evidence of the impact of ICT on achievement, and limited comparability on the large-scale assessments (Balanskat, Blamire, & Kefala, 2006; Cox & Marshall, 2007; De Witte & Rogge, 2014; Skryabin et al., 2015; Trucano, 2005). Although digital technologies are claimed to be important in the 21st century, some doubts have occurred that more or better ICT means better education (Livingstone, 2012). Pandolfini (2016) concluded that the majority of the studies are related to the impact of ICT and are figured out simple outcomes on the individual level, such as only teachers or students. In recent years, the tendency has been argued that the impact of ICT is highly complicated. In order to interpret the effects of ICT in education, more information is needed about how ICT operates at different levels (such as teacher, student, school, and parent) and what levels are measured (Erstad, 2009). The ICT-related research needs to be synthesized from a holistic perspective (Sutherland, Robertson, & John, 2009).

The studies of multilevel approaches to how the impact is interrelated on different levels, and to clarify the effects of ICT usage are becoming important (Pandolfini, 2016). This study focused on different levels of students and schools for the impact of ICT on students' mathematics achievement. One data set of PISA was used in the majority of studies to determine the effect of ICT on PISA mathematics achievement. For instance, Demir and Kılıç (2009) and Güzeller and Akın (2014) used PISA 2006 dataset, Delen and Bulut (2011) assessed PISA 2009 dataset, Wittwer and Senkbeil (2008) examined PISA 2003 dataset and Petko, Cantieni and Prasse (2017) investigated PISA 2012 dataset in their studies. One reason for this can be that one of science, reading and mathematics is chosen as the major domain in each assessment, and so the focused domain varies with each PISA implementation. The major domain is assessed more; the other two domains are minor domains and assessed less thoroughly. It is important to remember that these three domains are measured in every implementation of PISA. There are fewer studies which are related to the relationship between student and school characteristics and PISA mathematics achievement implemented in different years (e.g., Karabay, Yıldırım, & Güler, 2015). It can be said, according to our knowledge, that there are insufficient studies in literature on examining how the student and school level of ICT variables affect PISA mathematics achievement implemented in 2009 and in 2012.

This study focused on examining the effect of ICT variables on students' mathematics achievement in both PISA 2009 and PISA 2012 and comparing the predictive level of ICT variables on students' PISA mathematics achievement implemented in 2009 and in 2012. In PISA 2009, just five of seven scaled indexes ICT-related aspects for the information and communication technologies familiarity questionnaire (ICTFQ) were used in this study. In PISA 2012 ICT familiarity questionnaire, just five of eight scaled indexes ICT-related aspects were used in this study. The ICT variables are grouped into student level and school level in this study. The student level ICT variables are the ICT availability at home (ICTHOME), the ICT use for entertainment (ENTUSE), and the ICT use at home for school-related tasks (HOMSCH). The school level ICT variables are the ICT availability at school (ICTSCH) and the ICT use at school (USESCH). These three student level variables and two school variables of the ICTFQ, which are common in both PISA 2009 and PISA 2012, were selected in this study to compare the effect of ICT variables on PISA mathematics achievement implemented in 2009 and 2012. These student level and school level ICT variables are the common variables in both PISA 2009 and PISA 2012 ICTF questionnaire (OECD, 2012; OECD, 2014c). The reason for the selection of these variables is to compare two implementations of PISA which are PISA 2009 and PSA 2012.

This study will contribute to the following gaps in the literature: (a) ICT is constantly evolving, and its impact is difficult to isolate from the environment (Youssef & Dahmani, 2008). This research may contribute to the literature to clarify the impact of the level of access and usage of ICT on mathematics achievement. (b) As far as we investigate, there is a dearth of studies in the literature on comparing the explained variance ratio in mathematics achievement caused by ICT variables in two different implementations of PISA. In this study, the explained variance ratio in mathematics achievement in 2009 and 2012 caused by ICT variables was compared. The disclosure variance ratio could be given an idea about the effective usage of ICT in mathematics education by years because of changing the usage of ICT continuously over the years. (c) In this research, hierarchical linear models have been established. Considering the structure of the PISA dataset, it can be said that since the hierarchical models have calibrated the estimated standard error better, it started to become important to interpret the findings with less errors in order to reach more accurate results. (d) While the major domain was mathematics in PISA 2012, the domain of reading was given greater emphasis on PISA 2009. This study will provide an opportunity to interpret how the effect of ICT variables on mathematics achievement changes depending on the domain. Thus, this study aims to present a holistic perspective on the effect of ICT on mathematics achievement.

Purpose of the Study

This research aimed to investigate the impact of access and usage of ICT at both student variables and school variables on Turkish students' mathematics achievement in PISA 2009 and PISA 2012. The problem of this study is to examine the ratio of variance explained in mathematics achievement caused by the access and usage of ICT in PISA 2009 and PISA 2012 implementations. The research questions of this study are as follows:

1. What is the explained variance ratio in mathematics achievement caused by the difference among students and between schools according to PISA 2009 and 2012 data in Turkey?
2. What is the explained variance ratio in mathematics achievement caused by the variables regarding the access and usage of ICT at student level according to PISA 2009 and 2012 data in Turkey?
3. What is the ratio of variance explained in mathematics achievement caused by the variables related to ICT both at school level and at student level according to PISA 2009 and 2012 data in Turkey?

METHOD

This study was established on the correlational model. This research method is used to examine whether a relationship among two or more variables. The purposes of correlation model is to explore the phenomena and to make predictions by identifying relationships among variables (Fraenkel, Wallen, & Hyun, 2011).

Sample

The sample of this research consisted of a student group at the age of 15 having participated in PISA 2009 and PISA 2012 (MEB, 2010, 2013b). The sample design was a two-stage stratified sample design according to the PISA. The first-step sampling units involved in schools having 15-year-old students. The second-step sampling units included students within sampled schools. The sample consisted of 4996 students who participated in the PISA 2009 survey (OECD, 2012) and 4848 students who participated in the PISA 2012 survey (OECD, 2014b).

Data Collection Instruments

The data obtained from the mathematics achievement of students in PISA 2009 and PISA 2012, and the common indexes in the ICTFQ in PISA 2009 and 2012 were used in this study. The mathematics achievements of students in PISA 2009 and 2012 were calculated by using the generalized form of the Rasch model (OECD, 2014a). PISA mathematics performance was reported as five plausible variables (PVs) calculated using the one-parameter (Rasch) model for dichotomous items for each student in the sample. The PVs are random and draw from the marginal posterior distribution in PISA. PV1MATH, PV2MATH, PV3MATH, PV4MATH, and PV5MATH are the variables for mathematical literacy. Since the correlation between these plausible values is high, the PV1MATH randomly selected was used in this study. The value of the reliability of PISA 2009 mathematics domain is .90 (OECD, 2012), and the reliability value for PISA 2012 mathematics domain is .92 for Turkey (OECD, 2014c).

The ICT familiarity questionnaire was administered in both PISA 2009 and PISA 2012 (OECD, 2012, 2014c). The ICT variables are grouped into student level and school level in this study. The student level ICT variables are the ICT availability at home (ICTHOME), the ICT use for entertainment (ENTUSE), and the ICT use at home for school-related tasks (HOMSCH). The school level ICT variables are the ICT availability at school (ICTSCH) and the ICT use at school (USESCH).

In PISA 2009, seven scaled indexes ICT-related aspects were computed for this questionnaire, and five of them were used in this study. The labels of these student level ICT-related indexes are the ICT availability at home (ICTHOME and Cronbach α = .81), the ICT use for entertainment (ENTUSE and Cronbach α = .91) and the ICT use at home for school related tasks (HOMSCH and Cronbach α = .84). The labels of these school level ICT-related indexes are the ICT availability at school (ICTSCH and Cronbach α = .74) and the ICT use at school (USESCH and Cronbach α = .89) (OECD, 2012). ICTHOME variable had eight items in PISA 2009. The eight items provide information on ICT availability of a desktop computer, portable laptop or notebook, internet connection, video games console, cell phone, Mp3/Mp4 player, iPod or similar, printer and USB stick at home. This variable had three response categories which were *Yes, and I use it*, *Yes, but I don't use it* and *No*. ENTUSE variable included eight items. These items give information on the use of ICT and Internet for entertainment such as playing one-player games, playing collaborative online games, using e-mail, chatting online, browsing the internet for fun, downloading music, films, games or software from the Internet, publishing and maintaining a personal website or blog, participating in online forums, virtual communities or spaces. This variable had four response categories varying from *Never or hardly ever*, *Once or twice a month*, *Once or twice a week* to *Every day or almost every day*. The response categories for HOMSCH variable were same as the response categories of the ENTUSE variable. The five items of HOMSCH variable inform on the use of ICT for school related tasks. To browse the Internet for schoolwork, to use e-mail for communication with other students about schoolwork, to use e-mail for communication with teachers and submission of homework or other schoolwork, to

download, to upload or to browse material from your school's website (e.g., time table or course materials), to check the school's website for announcements, e.g., absence of teachers are the items of HOMSCH variable. ICTSCH variable had five items. The items were related to the availability of a desktop computer, portable laptop or notebook, internet connection, printer, and USB (memory) stick at school. The response categories for this variable were same as the response categories of the ICTHOME variable. USESCH variable had nine items, such as chatting online, using e-mail at school, browsing the Internet for schoolwork, downloading, uploading, or browsing material from the school's website, posting your work on the school's website, playing simulations at school, etc. These USESCH variable items provide information on student involvement in ICT related tasks at school. The response categories for this variable were same as the response categories of the ENTUSE variable.

Eight scaled indexes ICT-related aspects were computed utilizing the information which was obtained from PISA 2012 ICT familiarity questionnaire, and five of them were used in this study. The labels of these student level ICT-related indexes are the ICT availability at home (ICTHOME and Cronbach α = .78), the ICT use for entertainment (ENTUSE and Cronbach α = .90) and the ICT use at home for school related tasks (HOMSCH and Cronbach α = .86). The labels of these school level ICT-related indexes are the ICT availability at school (ICTSCH and Cronbach α = .75) and the ICT use at school (USESCH and Cronbach α = .89). In PISA 2012, the indexes of the ICTHOME, the ICTSCH and the ENTUSE were revised from 2009, and new items were added. The indexes of the HOMSCH and the USESCH were revised from 2009 (OECD, 2014c). For PISA 2012, ICTHOME variable had eleven items. These items were revised from 2009, and new items were added. The revised items are such as tablet computer, cell phone (without Internet Access), cell phone (with Internet Access), eBook reader. ENTUSE variable had ten items. Some of them were revised from 2009, and new items were added. The examples of the revised items of the ENTUSE variable are reading news on the Internet, obtaining practical information from Internet, uploading your own created contents for sharing. This variable had five response categories varying from *Never or hardly ever*, *Once or twice a month*, *Once or twice a week Almost every day* to *Every day*. HOMSCH variable for PISA 2012 included seven items. The items of this variable were revised from 2009. Five response categories for this variable were same as the response categories of the ENTUSE variable. Compared to PISA 2009, two new items, which were tablet computer and eBook reader, were added in the ICTSCH variable for PISA 2012, and the other items were revised from 2009. This variable had seven items and three response categories for this variable were same as the response categories of the ICTHOME variable. The items of USESCH variable were modified from 2009. This variable had nine items and five response categories for this variable were same as the response categories of the ENTUSE variable.

These three student level variables and the two school variables of ICT familiarity questionnaire are common indexes both in PISA 2009 and PISA 2012, and these variables were selected in this study to compare the effect of ICT variables on PISA mathematics achievement implemented in different years. For the construct validity of these scales, psychometric techniques such as correlations, confirmatory factor analyses, and Item Response Theory (IRT) scaling were used.

Most questionnaire items were scaled using IRT scaling methodology in PISA. One Parameter (Rasch) model was used for the dichotomous items (1, 0), and the partial credit model was used for items with multiple score categories (e.g., Likert type items). In order to obtain student scores, weighted likelihood estimation was primarily used by estimating international item parameters from the calibration sampling. Weighted likelihood estimations were transformed into an international metrics with an OECD average of 0 and 1 OECD standard deviation of 1, and indexes were obtained (OECD, 2012, 2014a). The data set ware taken from the website of OECD (2018a, 2018b). The data of Turkey were used from the file named INT_STQ09_DEC11 for the PISA 2009 data and from the file named INT_STU12_DEC03 for PISA 2012 data.

Data Analysis

Two level Hierarchical Linear Modelling (HLM) analysis was used (Raudenbush & Bryk, 2002). Since PISA dataset has a hierarchical structure, the student variables were dealt with at level 1, and the school variables were dealt with at level 2. HLM analysis has some assumptions. These were examined separately for PISA 2009 data and PISA 2012 data. One of these assumptions is related to missing value and outliers. Since the rate of missing value is low, missing value methods were utilized in HLM program for the assignment of missing value. Considering the size of sampling, no analysis was performed related to outliers. In order to determine the multicollinearity which is one of the HLM assumptions, the correlation coefficient value between the predictor variables in level 1 (student) and level 2 (school) is estimated. The correlation matrix for the first and second level variables is given in Table 1 (see Appendix).

The correlation coefficient values between student level variables ranged from .30 to .62. The correlation coefficient values between school level variables ranged from .23 to .35. These values were calculated as less than .70 in Table 1. In order to minimize the high correlation between level 1 and level 2 variables, the data are centered in the analysis (Raudenbush & Bryk, 2002). If the intercept variance represents the between group variance in the outcome measure, the data are centered around the group mean. In grand mean centered models, the intercept variance defines the between group variance in the outcome variable adjusted for the level 1 variables (Hofmann & Gavin, 1998). Hence the level 1 variables were centered around the group mean, while the second level variables were centered around the grand mean in this study. In another assumption of HLM, the normality of the errors at the student level and at the school level were analyzed. Histogram and likelihood graphics were obtained for this (P-P plot or Q-Q plot), and these graphics were found to compose 45-degree lines. Thus, the assumption of errors normality of at both levels were met. For the homogeneity of student level variances, H statistics was calculated, and p value was found to be significant. Considering the assumption of independence of errors, intra-school errors in PISA 2009 mathematics achievement were found to be independent of the student level variables (p_{ENTUSE} = 0.444 > .05; $p_{ICTHOME}$ = .418 > .05; p_{HOMSCH} = .825 > .05). Also, the assumption of independence of errors was ensured for PISA 2012 mathematics achievement (p_{ENTUSE} = .253 > .05; $p_{ICTHOME}$ = .133 > .05; p_{HOMSCH} = .211 > .05).

In order to examine the effects of ICT factors at both student and school levels on mathematics achievement, four models were established for both the implementations of PISA 2009 and PISA 2012. Model 1 is called the One-Way Variance Analysis Random Effects Model (also known as Null model). This model was established to answer the first research question. The equation for this model is as Equation 1, Equation 2 and Equation 3.

Level -1 (Student level) Model:

$$(Y_{ij}|M_{2009}/M_{2012}) = \beta_{0j} + r_{ij} \quad (1)$$

Level -2 (School level) Model:

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (2)$$

Combined Model:

$$(Y_{ij}|/M_{2009}/M_{2012}) = \gamma_{00} + u_{0j} + r_{ij} \quad (3)$$

Model 2 is called Random Coefficients Regression Model. This model involves a covariate at student level with a random effect which has different effects on the school level variables. This model was established in accordance with the second research question. The student level variables are allowed to be distributed randomly between schools, but the outcome variables at school level are not added to the model. The equation for this model is as Equation 4, Equation 5 and Equation 6.

Level - 1 (Student level) model:

$$(Y_{ij}|M_{2009}/M_{2012}) = \beta_{0j} + \beta_{1j} * (ENTUSE_{ij}) + \beta_{2j} * (HOMSCH_{ij}) + \beta_{3j} * (ICTHOME_{ij}) + r_{ij} \quad (4)$$

Level - 2 (School level) model:

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (5)$$
$$\beta_{1j} = \gamma_{10} + u_{1j}$$
$$\beta_{2j} = \gamma_{20} + u_{2j}$$
$$\beta_{3j} = \gamma_{30} + u_{3j}$$

Combined Model:

$$(Y_{ij}|M_{2009}/M_{2012}) = \gamma_{00} + \gamma_{10} * (ENTUSE_{\text{ij}}) + \gamma_{20} * (HOMSCH_{\text{ij}}) + \gamma_{30} * (ICTHOME_{\text{ij}}) + u_{0\text{j}} + u_{1\text{j}} * (ENTUSE_{\text{ij}}) + u_{2\text{j}} * (HOMSCH_{\text{ij}}) + u_{3\text{j}} * (ICTHOME_{\text{ij}}) + r_{\text{ij}} \quad (6)$$

In this model, β_{0j} stands for mean outcome variable, β_{1j}, β_{2j}, and β_{3j} stand for slope or the effects of predictors, r_{ij} coefficient stands for the random effect for i student clustered in j school, u_{0j} stands for error coefficients.

Model 3 is called Intercept and Slopes as Outcomes Model. This model was established in accordance with the third research question. The equation for this model is as Equation 7, Equation 8 and Equation 9.

Level - 1 (Student level) model:

$$(Y_{ij}|M_{2009}/M_{2012}) = \beta_{0j} + \beta_{1j} * (ENTUSE_{ij}) + \beta_{2j} * (HOMSCH_{ij}) + \beta_{3j} * (ICTHOME_{ij}) + r_{ij} \quad (7)$$

Level - 2 (School level) model:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} * (ICTSCH_{\text{ij}}) + \gamma_{02} * (USESCH_{\text{ij}}) + u_{0j} \quad (8)$$

Combined model:

$$(Y_{ij}|M_{2009}/M_{2012}) = \gamma_{00} + \gamma_{01} * (ICTSCH_{\text{ij}}) + \gamma_{02} * (USESCH_{\text{ij}}) + \gamma_{10} * (ENTUSE_{\text{ij}}) + \gamma_{20} * (HOMSCH_{\text{ij}}) + \gamma_{30} * (ICTHOME_{\text{ij}}) + u_{0\text{j}} + r_{\text{ij}} \quad (9)$$

RESULTS

Within the scope of the aim of the study, the results were obtained from Random Effects Model of One-Way Variance Model developed based on PISA 2009 mathematics achievement and PISA 2012 mathematics achievement to answer the first research question are given in Table 2 (see Appendix).

When Table 2 is examined, it is seen that average school mean mathematics achievement of PISA 2009 was statistically different from zero (t = 73.36, p < .001). Considering the mean and variance, the mean mathematics achievement of PISA 2009 varied between 424.48 and 447.76 by a possibility of 95% (436.12 ± 1.96(5.94)). For PISA 2012 data set, average school mean mathematics achievement was statistically different from zero (t = 77.04, p < .001). In addition to that, the mean mathematics achievement of PISA 2012 shifted from 428.71 to 451.09 within 95% confidence interval. Table 3 is related to the information on the last estimation of the random effects in the model (see Appendix).

When Table 3 is reviewed, considering the general average in Turkey, the variance of school means (inter-school variability) was estimated to be 5795.96 for PISA 2009. The variance of the student's mathematics achievement scores was estimated to be 3502.58 within the framework of the school average (intra-school variability) at the student level (level 1). The value range for the school averages shifted from 286.9 to 585.33 by a possibility of 95% ($436.12 \pm 1.96*\sqrt{5795.96}$). The variance of school means (inter-school variability) was estimated to be 5327.39 for PISA 2012. The variance of the student's mathematics achievement scores was estimated to be 3158.00 within the framework of the school average at the student level for PISA 2012. With 95% confidence, the school averages range from 296.85 to 582.95.

These results showed that there is a broad range of variance in mathematics achievement levels between the schools. In order to determine the explained variance ratio of students' mathematics achievement scores in PISA 2009 and PISA 2012, the interclass correlation coefficient and the intraclass correlation coefficient were calculated, and the calculations are given in Table 4. The intraclass correlations are related to the difference between students, and the interclass correlations are regarding the difference between schools (see Appendix).

Table 4 presented that the difference between the mathematics achievement scores of the students was found to be 62% in both PISA 2009 and PISA 2012. The remaining 38% of the variability in mathematics achievement was within the schools. It refers that mean mathematics achievement of schools differs heterogeneously between schools. These coefficient values show that there is an explained variance between schools. Therefore, the analysis was continued, including variables at student and school levels. The student-level variables were added to reduce the variance within schools, and the school-level variables were added to explain between-school variance.

The second research question is related to the explained variance ratio at the student level ICT variables in students' mathematics achievement scores PISA 2009 and PISA 2012. In order to examine this research question, three variables which are the ICT availability at home (ICTHOME), the ICT use for entertainment (ENTUSE), the ICT use at home for school-related tasks (HOMSCH) were added in the model. This model includes in level-1 variables. The findings regarding Random Coefficients Regression Model are given in Table 5 (see Appendix).

Considering each of the predictor variables at student level, which affect mathematics achievement, other variables were held fixed except one to determine its impact in Table 5. The relationship between the ICT use for entertainment (ENTUSE) and PISA 2009 mathematics achievement was positive, and this relationship was statistically significant ($M_{ENTUSE\gamma 10}$ = 3.85, *SE* = 0.92, $p < .05$). The ICT use at home for school-related tasks (HOMSCH) decreased PISA 2009 mathematics achievement, and this decline was statistically significant ($M_{HOMSCH\gamma 20}$ = -8.77, *SE* = 0.99, $p < .05$). The relationship between the ICT availability (e.g. laptop, computer, printer, USB, internet connection) at home (ICTHOME) and PISA 2009 mathematics achievement was positive, and this relationship was statistically significant ($M_{ICTHOME\gamma 30}$ = 6.39, *SE* = 0.94, $p < .05$). In order to compute the effect size of each student level variable which has a significant effect on mathematics achievement, each beta coefficient was divided by the pooled within-school standard deviation. The pooled within-school standard deviation is computed by taking the square root of σ^2 in Null Model (von Secker & Lissitz, 1999). Effect size is a standard deviation (SD) unit that allows comparison of outcomes with different measurements. It describes changes in the dependent variable when other independent variables are held fixed. Thus, it can be represented as the SD change in the dependent variable connected to 1SD change in an independent variable. If the value of effect size is computed as smaller than .1 SD, the effect is trivial. If the effect size value is between .1 SD and .3 SD, the effect is small. If the effect size value is between .3 SD and .5 SD, the effect is moderate. If the effect size value is computed as larger than .5 SD, this effect is large (Rosenthal & Rosnow, 2008; von Secker & Lissitz, 1999). When Table 3 was examined, the standard deviation was calculated as 59.2 ($\sqrt{3502.58}$) for within-school. The beta coefficient value for the ENTUSE variable was 3.85 in Table 5. The effect size value of the ENTUSE variable was calculated as .07 SD. It means that an increase of 1 SD in the variable of ENTUSE causes an increase of .07 SD in the students' mean mathematics achievement. The effect size value was calculated as .15 SD for the HOMSCH variable, and as .11 SD for the ICTHOME variable. The effect size of the HOMSCH variable indicates that an increase of 1 SD in the HOMESCH variable results in a decrease of .15 SD in the students' mean mathematics achievement. The effect size of the ICTHOME variable interprets as the .11 SD increase in the students' mean mathematics achievement linked to 1 SD increase in the ICTHOME variable. Considering the effect sizes, the HOMSCH and the ICTHOME variables had small effects, and the ENTUSE had a trivial effect on student's mathematics achievement in PISA 2009.

The ICT use for entertainment (ENTUSE) increased their PISA 2012 mathematics achievement, so this increment was statistically significant ($M_{ENTUSE\gamma 10}$ = 4.04, *SE* = 0.76, $p < .05$). The relationship between the ICT use at home for school-related tasks (HOMSCH) and PISA 2012 mathematics

achievement was negative, but this relationship was not statistically significant ($M_{HOMSCH\gamma 20}$ = -1.60, *SE* = 0.97, $p > .05$). The relationship between the ICT availability (e.g., laptop, computer, printer, USB, internet connection) at home (ICTHOME) and PISA 2012 mathematics achievement was also positive, and this relationship was statistically significant ($M_{ICTHOME\gamma 30}$ = 2.71, *SE* = 0.84, $p < .05$). When Table 3 was examined, the standard deviation was calculated as 56.1 ($\sqrt{3158.00}$). The value of effect size was calculated as .07 SD for the ENTUSE variable and as .05 SD for the ICTHOME variable. The effect size of the ENTUSE variable indicates that an increase of 1 SD in the ENTUSE variable results in an increase of .07 SD in the students' mean mathematics achievement. The effect size of the ICTHOME means that an increase of 1 SD in the variable of ICTHOME causes an increase of .05 SD in the students' mean mathematics achievement. When the effect size value of each variable was reviewed, each of the predictive variables had a trivial effect on students' mathematics achievement in PISA 2012.

The random effect of predictive variables which were caused by the variance between schools in students' PISA mathematics achievements is given in Table 6 (see Appendix).

When Table 6 is reviewed, the variance of the mathematics achievement scores of the schools was estimated to be 5807.83 in PISA 2009 and 5329.93 in PISA 2012, after the student level variables were added to the model. In order to determine the explained variance ratio in 2009 mathematics achievement caused by the difference within schools, the data obtained from the One-Way Variance Analysis and the data obtained in Table 6 were used. The explained variance ratio in PISA 2009 mathematics achievement at the student level is calculated as 0.027 [(3502.58 - 3405.48) / (3502.58)]. According to this result, there is a decrease of 2.7% in the explained variance ratio with the addition of the student level variables to the model in PISA 2009. In other words, the proportion of 2.7% of students' individual differences in PISA 2009 mathematics achievement results from the student level ICT variables added to the model (the ICT availability at home, the use of ICT for entertainment, the use of ICT at home for school-related task). Considering the Null model, 38% of the total variance in PISA 2009 mathematics achievement was caused by the differences between students. Thus, only 1.03% (38% * 2.7%) of the total variance of the student level ICT variables explained the difference of PISA 2009 mathematics achievement.

The variance ratio in PISA 2012 mathematics achievement explained by the student level ICT variables was calculated as 0.012. Accordingly, the explained variance ratio will decrease nearly by 1.2% after the student level variables are added to the model. In other words, the percent of 1.2 of the variability in students' PISA 2012 mathematics achievement is caused by the student level ICT variables added to the model (r^2 = .012). Considering the Null model, 38% of the total variance in PISA 2012 mathematics achievement was caused by the differences between students, only 0.45% (38% * 1.2%) of the total variance of the student level ICT variables explained the difference of PISA 2012 mathematics achievement.

Intercept and Slopes as Outcomes Model was tested to answer the third research question of the study. The model is obtained by the inclusion to the analysis all of the ICT variables which were determined to have a significant effect on the mathematics achievement at student and school level in PISA 2009 and PISA 2012. The findings regard the Intercept and Slopes as Outcomes Model are given in Table 7 (see Appendix).

In table 7, it is seen that PISA 2009 mean mathematics achievement and PISA 2012 mean mathematics achievement was statistically different from zero ($\gamma_{00} = 435.69$, $p < .001$ for PISA 2009; $\gamma_{00} = 438.30$, $p < .001$ for PISA 2012). When the variable of the ICT use at school (USESCH) was holding fixed, it was determined that the variable of the ICT availability at school (ICTSCH) had a significant effect on mathematics achievement in PISA 2009. When the variable of the ICT availability at school (ICTSCH) was holding fixed, the ICT use at school (USESCH) variable reduced PISA 2009 average mathematics achievement. Holding fixed the variables which are the ICT availability at home (ICTHOME) and the ICT use at home for school-related tasks (HOMSCH), the variable of the ICT use for entertainment (ENTUSE) increased PISA 2009 average mathematics achievement. When the variables of the ICT availability at home (ICTHOME) and the ICT use for entertainment (ENTUSE)

were holding fixed, the variable of the ICT use at home for school-related tasks (HOMSCH) decreased PISA 2009 average mathematics achievement. Holding fixed the variables of the ICT use for entertainment (ENTUSE) and the ICT use at home for school-related tasks (HOMSCH), the ICT availability at home (ICTHOME) increased PISA 2009 average mathematics achievement. The variables with the highest impact value in PISA 2009 mathematics achievement are the ICTSCH and USESCH variables. These variables are the school level variables. It is expected that 1 SD increase in the ICTSCH variable will increase .69 SD in the students' mean mathematics achievement while 1 SD increase in the USESCH variable will decrease 1 SD in the students' mean mathematics achievement in PISA 2009. When the student level variables reviewed, their effect size were not greater than the school level variables.

When the variable of the ICT use at school (USESCH) was holding fixed, the variable of the ICT availability at school (ICTSCH) increased PISA 2012 average mathematics achievement. When the variable of the ICT availability at school (ICTSCH) was holding fixed, the ICT use at school (USESCH) decreased PISA 2012 average mathematics achievement. When the variables which are the ICT availability at home (ICTHOME) and the ICT use at home for school-related tasks (HOMSCH) were holding fixed, the ICT use for entertainment (ENTUSE) increased PISA 2012 average mathematics achievement. When the variables of the ICT availability at home (ICTHOME) and the ICT use for entertainment (ENTUSE) were holding fixed, the ICT use at home for school-related tasks (HOMSCH) reduced PISA 2012 average mathematics achievement. Holding fixed the variables which are the ICT use for entertainment (ENTUSE) and the ICT use at home for school-related tasks (HOMSCH), the variable of the ICT availability at home increased PISA 2012 average mathematics achievement. The variables with the highest impact value in PISA 2012 mathematics achievement is the ICTSCH and USESCH variables. It is expected that 1 SD increase in the ICTSCH variable will increase .83 SD in the students' mean mathematics achievement while 1 SD increase in the USESCH variable will decrease .78 SD in the students' mean mathematics achievement in PISA 2012. When the student level variables reviewed, their effect sizes were not greater than the school level variables.

When Table 7 was examined in general, it was seen that the ICT variables at school level caused an excessive amount of increase and decrease in average mathematics achievement defined as outcome variable. However, the student level ICT variables caused a low amount of increase and decrease in average mathematics achievement. Table 8 comprises the random effect of predictive variables caused by the variance among students and schools of mathematics achievement (see Appendix).

The data obtained from Table 8 and the data obtained from Random Coefficients Regression Analysis were used to calculate the explained variance ratio in 2009 mathematics achievement caused by the student and school levels. According to the calculation, 27% of the variance in the between-school difference in mean PISA 2009 mathematics achievement was explained by the school level variables. Also, χ^2 = 5599.33 was calculated, and p value was found to be statistically significant, so it can be said that there is still an unexplained variance between schools. The effect size value was calculated as .69 for the ICTSCH variable, and as -1 for the USESCH variable. The value of effect size was calculated as .06 for the ENTUSE variable, as -.14 for the HOMSCH variable, and as .08 for the ICTHOME variable. When the effect sizes were reviewed, it was seen that the ICTSCH and the USESCH variables had a large effect, the HOMSCH had a small effect, and the ENTUSE and the ICTHOME had a trivial effect on student's mathematics achievement in PISA 2009.

For PISA 2012 mathematics achievement the variance ratio was calculated as 31% [(5327.39 - 3656.48) / 5329.93]. The variables which are the ICT availability at school and the ICT use at school explained 31% of the variance in the between-school difference in mean PISA 2012 mathematics achievement. In addition, χ^2 = 5901.47 was calculated, and p value was found to be statistically significant, so it can be said that there is still an unexplained variance between schools. When Table 3 was examined, the standard deviation was calculated as 72.9 ($\sqrt{5327.39}$). The effect size of the ICTSCH variable was calculated as .83. The effect size was calculated as -.78 for the USESCH variable. The effect size was calculated as .07 for the ENTUSE variable, and as .05 for the ICTHOME variable. When the effect sizes were examined, it was seen that the ICTSCH and the USESCH

variables had a large effect, the ENTUSE and the ICTHOME had a trivial effect on student's mathematics achievement in PISA 2012.

Four different models were established for HLM analyses in the study. Likelihood ratio test was calculated to determine whether the established the model 4 was better likelihood than the other models or not. For this reason, firstly, the difference of deviance statistics values of each model divides by the degree of freedom. The obtained value is compared to the critical chi-square value. The model is statistically significant if this value is greater than the critical value (critical $\chi^2 = 5.99$ for $p = .05$). The results of the likelihood ratio test using deviance statistics in each outcome variable to determine whether the Model 4 fits significantly better are given in Table 9 (see Appendix). When the results of the Likelihood ratio test for both PISA 2009 mathematics achievement and PISA 2012 mathematics achievement were examined, it could be said that the Model 4 fits significantly better.

DISCUSSION and CONCLUSION

In the study, the ICT variables predicting mathematics achievement at the student level and the school level were examined. When the student level ICT variables are reviewed, one of the variables at the student level is the ICT use for entertainment. There are studies in the literature similar to the consequence of this study in which there is a positive and significant relationship between the ICT use for entertainment and PISA mathematics achievement (e.g., Bilican-Demir & Yıldırım, 2016; Demir, Kılıç, & Ünal, 2010; Dumais, 2009; Hu, Gong, Lai, & Leung, 2018; Petko et al., 2017; Skryabin et al., 2015). It is emphasized that the usage of computers for entertainment such as playing games on computer which is thought by parents as a waste of time is important in the cognitive development of students (Becker, 2000; Hamlen, 2011; Li & Atkins, 2004) and in visual intelligence development (Subrahmanyam, Greenfield, Kraut, & Gross, 2001), which can positively affect achievement. Also, entertainment can help overcoming their stress and anxiety and thus, it can enable them to focus on their learning; besides, it can contribute to students' effective and critical thinking (Wittwer & Senkbeil, 2008; Ziya, Doğan, & Kelecioğlu, 2010). However, there are also studies about that the internet usage for entertainment is a negative and significant predictor of mathematics achievement in the literature (e.g., Cheema & Hang, 2013; Güzeller & Akın, 2014). The reason for this result can be explained by the fact that excessive ICT use for entertainment neglects students' responsibilities for school (Cheema & Hang, 2013; Luu & Freeman, 2011). If students' usage of ICT is not controlled and monitored, it will cause negative social and psychological effects such as addiction to game playing (Grüsser, Thalemann, & Griffiths, 2006). Moreover, the reason why there are inconsistent results related to the effect of ICT use for entertainment on mathematics achievement in the literature can be explained by the fact that the ICT use for entertainment causes different effects on different mathematics topics (Biagi & Loi, 2013). Further studies about the influences of the ICT activities for entertainment on students' academic outcomes and the causes of these influences are still needed.

Another variable dealt with at the student level is the ICT use at home for school-related tasks. In the study, it was found that the relationship between the ICT use at home for school-related tasks and PISA 2009 mathematics achievement is negative and significant. However, that relationship of it with PISA 2012 mathematics achievement is negative but not significant. There are studies with similar results in the literature (e.g., Hu et al., 2018). However, there are several studies that the use of ICT has a positive effect on learning outcome (e.g., Kubiatko & Vlckova, 2010; O'Neil, Wainess, & Baker, 2005; Skryabin et al., 2015). The students' ICT use for school-related tasks mostly includes homework. Turkish students frequently have difficulty in mathematics homework (Güven & Demirçelik, 2013; MEB, 2011). Thus, students may develop negative prejudices and attitudes towards mathematics lessons and homework (Yenilmez & Dereli, 2009). This case can negatively affect achievement. Besides, the students' spending much time on ICT activities not related to their school-related tasks (Zhang & Liu, 2016) and their lack of knowledge how to use ICT for accomplishing school-related tasks (Kubiatko & Vlckova, 2010; Petko et al., 2017) are among the factors that affect achievement negatively.

The other variable dealt with at the student level is about the ICT availability at home (ICTHOME), and it was concluded that the relationship between this variable and PISA mathematics achievement in 2009 and 2012 is positive and significant in this study. This result is consistent with the results of some studies in the literature (e.g., Delen & Bulut, 2011; Demir & Kılıç, 2009; Erdoğdu & Erdoğdu, 2015; Özer & Anıl, 2011). Taking into consideration to this result, it can be mentioned that the students can reach more information from several sources regarding the topics (Kubiatko & Vlckova, 2010). Also, the average percentage of internet access at home has increased over the years (OECD.Stat, 2018). Yet, Aypay (2010), Bilican-Demir and Yıldırım (2016), and Wittwer and Senkbeil (2008) couldn't find a significant relationship between the student's ICT opportunity and achievement in their studies. Hu et al. (2018) found that ICT availability at home is negatively associated with student's academic success. The reason for this inconsistency in literature can be explained by the fact that while the ICT availability at home gives many opportunities in education, the ineffective usage of ICT for education can affect his/her education negatively (Hu et al., 2018; Lei & Zhao, 2007). In brief, achievement is affected by how and for what purpose the availability of ICT is used at home (İlgün-Dibek, Yalçın, & Yavuz, 2016).

One of the variables dealt with at school level in the study is the ICT availability at school (ICTSCH), and a positive and significant relationship was found between this variable and PISA mathematics achievement in 2009 and 2012. In literature, there are studies having reached similar results (Delen & Bulut, 2011; Hu et al., 2018; Olkun & Altun, 2003; Özer & Anıl, 2011). The students in schools with ICT facilities can have access to more information using several sources regarding lessons (Kubiatko & Vlckova, 2010). The schools in Turkey are also well enough with regard to ICT devices (Seferoğlu, 2015). Another variable at school level is ICT use at school (USESCH). And, the consequence of its negative and significant relationship with PISA mathematics achievement in 2009 and 2012. Bilican-Demir and Yıldırım (2016), Cheema and Hang (2013) and Petko et al. (2017) found similar findings using PISA data and Skryabin et al. (2015) reached similar results using TIMMS dataset. This may be due to the lack of restrictions on access to websites in schools (Kubiatko & Vlckova, 2010). Another reason can be the students' unfamiliarity with ICT use in lessons (İlgün-Dibek et al., 2016). One of the other reasons is that the teacher's proficiency in ICT and their information in teaching methods can be lacking and insufficient (Baki, Yalçınkaya, Özpınar, & Uzun, 2009; Pandolfini, 2016). Because, if the students' learning targets with ICT are not certain, the teaching value of ICT is low (Kubiatko & Vlckova, 2010), and it gets harder to reach the targeted achievement. The applicability of the FATİH project in Turkey is discussed in this context, because the number of teachers using the ICT in lessons is very low, and they generally use word processor and presentation programs actively (Demiraslan & Usluel, 2005; Kayaduman, Sırakaya, & Seferoğlu, 2011).

In the study, it is noticed that the results regarding the relationship between ICT variables at student level and school level and PISA mathematics achievement are consistent with the results of some studies but contradict with some other studies in the literature. One of the reasons for this can be methodological restrictions and differences (Cox & Marshall, 2007; De Witte & Rogge, 2014). The different data analysis techniques were used in studies with PISA dataset or one of the other large-scale assessments. Also, the results of this study were compared with the results of studies using PISA dataset of the different countries in literature, and some of the results were determined to be consistent and some others to be inconsistent with them. This case could be caused by the fact that each country has its own educational policies and applications regarding ICT use, and these ICT applications and these ICT skills may be different in each country (Heinz, 2016; Skryabin et al., 2015).

The variables dealt with both at student level and at school level in the study can be categorized as ICT availability and ICT use. At both levels, it was concluded that ICT availability increases achievement, but ICT use is not effective in increasing achievement. Thus, the technological richness of a house or a school does not mean that using these technologies effectively. Effective technology usage is connected to the knowledge, the ability, and the experiences of the parents at homes and of the administrators and the teachers at schools (Hu et al., 2018; Lei & Zhao, 2007; Seferoğlu, 2015).

One of the other results of this study is that the explained variance ratio in mathematics achievement caused by the ICT variables at school level was greater than by the ICT variables at student level. This

situation can be affected by the factors such as the principals' awareness of the ICT applications, the school culture, the cooperation regarding how ICT is used in schools, the teachers' ICT proficiency, the teacher education on teaching methods (Pandolfini, 2016) and the pedagogical developments (Ertmer & Ottenbreit-Leftwich, 2010).

This study also examined the comparison of mathematics achievement between PISA 2009 and PISA 2012. Mathematics is the major domain in PISA 2012, but this domain was minor in PISA 2009. Therefore, the effect of ICT on mathematics achievement was compared with whether it depends on the focused domain. Comparing the results regarding mathematics achievement of PISA 2009 and PISA 2012, it was concluded that whether the major domain is mathematics, in other words, mathematics achievement test is long or short did not make a serious difference in mathematics achievement.

When the effect sizes of the student level variables on mathematics achievement were compared with two implementations of PISA which are PISA 2009 and PSA 2012, the ENTUSE had a trivial effect on student's mathematics achievement in both PISA 2009 and PISA 2012. While the relationship between the HOMSCH variable and PISA 2009 mathematics achievement was negative and statistically significant, the relationship between the HOMSCH variable and PISA 2012 mathematics achievement was negative and not statistically significant. The ICTHOME variable had a small effect on PISA 2009 mathematics achievement, but this variable had a trivial effect on PISA 2012 mathematics achievement. The effect size value of ICTHOME variable on mathematics achievement in the PISA 2012 implementation was lower than in the PISA 2009 implementation. The reason of the trivial and the small effect of student level variables may be the students' competence and awareness of the effective ICT use (Grüsser et al., 2006) and the parents' views of the ICT use (Becker, 2000; Hamlen, 2011; Li & Atkins, 2004).

When the effect sizes of the school level variables on mathematics achievement were compared with two implementations of PISA which are PISA 2009 and PSA 2012, The ICTSCH variable and the USESCH variable at the school level had a large effect on mathematics achievement in both PISA 2009 and PISA 2012. The reason for the large effect of the ICTSCH variable at the school level can be explained by the perspective that a good learning environment has an effect on the students' achievement (Youssef & Dahmani, 2008). The ICTSCH variable had a positive effect on mathematics achievement in both PISA 2009 and PISA 2012. The effect size value of the ICTSCH variable on mathematics achievement in the PISA 2012 implementation was higher than in the PISA 2009 implementation. The relationship between the USESCH variable and mathematics achievement in PISA 2009 and PISA 2012 was negative and statistically significant. The result of the negative relationship may be due to the teachers' quality and characteristics of the usage of ICT (Youssef & Dahmani, 2008). The effect size value of USESCH variable on mathematics achievement in the PISA 2012 implementation was lower than in the PISA 2009 implementation. The effect size value of the USESCH variable reduced in PISA 2012, but there has been a negative relationship between the USESCH variable and mathematics achievement. The reason for this negative relationship may be related to many barriers such as lack of confidence and competence and access to resources encountered (Bingimlas, 2009). In other words, the school principals' and the teachers' perceptions and their usage of ICT have not changed seriously over the years. In brief, the higher impact variables on mathematics achievement in both PISA 2009 and PISA 2012 were the ICTSCH variable and the USESCH variable which are the school level variables. The student level variables had the lowest impact on mathematics achievement in both PISA 2009 and PISA 2012.

It was found that the ICT variables both at school level and at student levels explained 27% of PISA 2009 mathematics achievement variance, while these variables explained 31% of PISA 2012 mathematics achievement variance. So, it was noticed that there was a slight increase in the explained variance ratio from 2009 to 2012. Yet the explained variance ratio at student level was calculated as 2.7% in PISA 2009, and this ratio was accounted for 1.2% for PISA 2012. When the student level variables were compared by years, the effect of the ICT variables at student level had reduced from 2009 to 2012. The reason of the small amount of variance increase obtained from the study can be

explained by the slight increase of the ICT use awareness of the families, the teachers, and the administrators who shape the students' ICT use at home or at school. If students have several ICT availabilities, these opportunities offer a great number of sources and access to information for students' learning. However, it should be remembered that the usage and the purpose of ICT affect the students' learning (İlgün-Dibek et al., 2016).

Having a negative relationship between ICT use at home for school-related tasks and mathematics achievement actually poses a problem. This problem can be solved by changing the content of the school-related tasks. For instance, the school-related mathematical tasks may include entertaining components that help students to develop a love for mathematics. Besides, students can be consciously directed to use online materials for school-related tasks and for accomplishing their homework. Also, there are important responsibilities at home for families. One of them is the families' monitoring. Another responsibility is controlling the students' ICT use materials at home and teaching their children how to use online materials consciously.

The negative relationship between ICT use at school and mathematics achievement is another problem. In order to eliminate this problem, ICT use for entertainment can be integrated into lessons. For instance, games can be utilized to be successful in mathematics lessons at schools. For effective ICT applications, the teachers' ICT proficiency is important. Therefore, the teachers should be encouraged to participate in in-service training for developing their ICT proficiencies. Besides, there is a need for projects related to increasing the teachers' effective ICT use and the families' awareness of ICT use. Students' socio-economic background, age and gender, and learning expectations are important factors that affect ICT use and achievement (Balanskat, Bannister, Hertz, Sigillò, & Vuorikari, 2013). However, these variables were not included in the model in this study. This is one of the limitations of this study. As a suggestion to this limitation, some researches in which the variables related to the student's characteristics, the learning environment, and the school features are added in the model can be done. The other limitation of this study is to use two level Hierarchical Linear Modelling. Several studies can be offered for different multi-levels (e.g., three level models) related to investigating the effect of ICT on achievement by adding these variables into the model. The data in this study is limited to only one country. The studies related to comparing the effect of ICT usage on achievement between different countries are suggested to be performed.

REFERENCES

Acar, T. (2012). 2009 yılı uluslararası öğrenci başarılarını değerlendirme programı'nda Türk öğrencilerin başarılarını etkileyen faktörler. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 3*(2), 309-314. http://dergipark.org.tr/tr/pub/epod/issue/5802/77221 adresinden edinilmiştir.

Akın, Y., & Cancan, M. (2007). Matematik öğretiminde problem çözümüne yönelik öğrenci görüşleri analizi. *Atatürk Üniversitesi Kazım Karabekir Eğitim Fakültesi Dergisi,* (16), 374-390. http://dergipark.org.tr/tr/pub/ataunikkefd/issue/2777/37247 adresinden edinilmiştir.

Alakoç, Z. (2003). Matematik öğretiminde teknolojik modern öğretim yaklaşımları. *TOJET: The Turkish Online Journal of Educational Technology, 2*(1), 43-49. http://tojet.net/articles/v2i1/217.pdf adresinden edinilmiştir.

Anderson, R. E. (2008). Implications of the information and knowledge society for education. In J. Voogt & G. Knezek (Eds*.), International handbook of information technology in primary and secondary education* (pp. 5-22). Boston, MA: Springer.

Aypay, A. (2010). Information and communication technology (ICT) usage and achievement of Turkish students in PISA 2006. *Turkish Online Journal of Educational Technology (TOJET), 9*(2), 116-124. Retrieved from https://files.eric.ed.gov/fulltext/EJ898009.pdf

Baki, A., Yalçınkaya, H. A., Özpınar, İ., & Uzun, S. Ç. (2009). İlköğretim matematik öğretmenleri ve öğretmen adaylarının öğretim teknolojilerine bakış açılarının karşılaştırılması. *Turkish Journal of Computer and Mathematics Education, 1*(1), 67-85. http://dergipark.org.tr/tr/pub/turkbilmat/issue/21560/231419 adresinden edinilmiştir.

Balanskat, A., Bannister, D., Hertz, B., Sigillò, E., & Vuorikari, R. (2013). Overview and analysis of 1: 1 learning initiatives in Europe. In S. Bocconi, A. Balanskat, P. Kampylis, & Y. Punie (Eds.), *Overview and analysis of 1: 1 learning initiatives in Europe* (pp. 1-166). Spain, Luxembourg: European Commission. doi: 10.2791/20333.

Balanskat, A., Blamire, R., & Kefala, S. (2006). The ICT impact report: A review of studies of ICT impact on schools in Europe. *European Schoolnet, 1*, 1-71. Retrieved from http://colcct.colfinder.org/sites/default/files/ict_impact_report_0.pdf

Barkatsas, A. T., Kasimatis, K., & Gialamas, V. (2009). Learning secondary mathematics with technology: Exploring the complex interrelationship between students' attitudes, engagement, gender and achievement. *Computers & Education, 52*(3), 562-570. doi: 10.1016/j.compedu.2008.11.001

Becker, H. J. (2000). Pedagogical motivations for student computer use that lead to student engagement. *Educational Technology, 40*(5), 5-17. Retrieved from https://www.jstor.org/stable/pdf/44428608.pdf

Biagi, F., & Loi, M. (2013). Measuring ICT use and learning outcomes: Evidence from recent econometric studies. *European Journal of Education, 48*(1), 28-42. doi: 10.1111/ejed.12016

Bilican-Demir, S., & Yıldırım, Ö. (2016). Okulda ve okul dışında bilgi ve iletişim teknolojilerinin kullanımının öğrencilerin PISA 2012 performansıyla ilişkisinin incelenmesi. *Kastamonu Eğitim Dergisi - Kastamonu Education Journal, 24*(1), 251-262. http://dergipark.org.tr/tr/pub/kefdergi/issue/22606/241619 adresinden edinilmiştir.

Bingimlas, K. A. (2009). Barriers to the successful integration of ICT in teaching and learning environments: A review of the literature. *Eurasia Journal of Mathematics, Science & Technology Education, 5*(3), 235-245. doi: 10.12973/ejmste/75275

Cheema, J. R., & Hang, B. (2013). Quantity and quality of computer use and academic achievement: Evidence from a large-scale international test program. *International Journal of Education and Development using Information and Communication Technology, 9*(2), 95-106. Retrieved from http://ijedict.dec.uwi.edu/viewissue.php?id=35

Cox, M. J., & Marshall, G. M. (2007). Effects of ICT: Do we know what we should know? *Education and Information Technologies, 12*(2), 59-70. doi: 10.1007/s10639-007-9032-x

De Witte, K., & Rogge, N. (2014). Does ICT matter for effectiveness and efficiency in mathematics education? *Computers & Education, 75*, 173-184. doi: 10.1016/j.compedu.2014.02.012

Delen, E., & Bulut, O. (2011). The relationship between students' exposure to technology and their achievement in science and math. *TOJET, 10*(3), 311-317. Retrieved from http://www.tojet.net/articles/v10i3/10336.pdf

Demir, İ., & Kılıç, S. (2009). Effects of computer use on students' mathematics achievement in Turkey. *Procedia Social and Behavioral Sciences, 1*(1), 1802-1804. doi: 10.1016/j.sbspro.2009.01.319

Demir, İ., Kılıç, S., & Ünal, H. (2010). Effects of students' and schools' characteristics on mathematics achievement: findings from PISA 2006. *Procedia-Social and Behavioral Sciences, 2*(2), 3099-3103. doi: 10.1016/j.sbspro.2010.03.472

Demiraslan, Y., & Usluel, Y. K. (2005). Bilgi ve iletişim teknolojilerinin öğrenme öğretme sürecine entegrasyonunda öğretmenlerin durumu. *The Turkish Online Journal of Educational Technology, 4*(3), 109-114. http://www.tojet.net/articles/v4i3/4315.pdf adresinden edinilmiştir.

Dumais, S. A. (2009). Cohort and gender differences in extracurricular participation: The relationship between activities, math achievement, and college expectations. *Sociological Spectrum, 29*(1), 72-100. doi: 10.1080/02732170802480543

Erdoğdu, F., & Erdoğdu, E. (2015). The impact of access to ICT, student background and school/home environment on academic success of students in Turkey: An international comparative analysis. *Computers & Education, 82*, 26-49. doi: 10.1016/j.compedu.2014.10.023

Erstad, O. (2009). Addressing the complexity of impact. A multilevel approach towards ICT in education. In F. Scheuermann & F. Pedró (Eds.), *Assessing the effects of ICT in education Indicators, criteria and benchmarks for international comparisons* (pp. 21-40). Luxembourg: Publications Office of the European Union.

Ertmer, P. A., & Ottenbreit-Leftwich, A. T. (2010). Teacher technology change: How knowledge, confidence, beliefs, and culture intersect. *Journal of Research on Technology in Education, 42*(3), 255-284. doi: 10.1080/15391523.2010.10782551

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2011). *How to design and evaluate research in education* (8th ed.). New York, NY: McGraw-Hill Humanities/Social Sciences/Languages.

Grüsser, S. M., Thalemann, R., & Griffiths, M. D. (2006). Excessive computer game playing: Evidence for addiction and aggression? *CyberPsychology and Behavior, 10*(2), 290-292. doi: 10.1089/cpb.2006.9956

Güven, S., & Demirçelik, D. A. (2013). 6. 7. ve 8. sınıf öğrencilerin performans ödevleri hakkındaki görüşleri ve bu ödevi hazırlamaya yönelik etik algıları. *Uluslararası Avrasya Sosyal Bilimler Dergisi-International Journal of Eurasia Social Sciences, 4*(13), 83-104. http://www.ijoess.com/Makaleler/271634930_sevim%20g%C3%BCven-performan%20%C3%B6devlerine%20y%C3%B6nelik.pdf adresinden edinilmiştir.

Güzeller, C. O., & Akın, A. (2014). Relationship between ICT variables and mathematics achievement based on PISA 2006 database: International evıdence. *TOJET, 13*(1), 184-192. Retrieved from https://files.eric.ed.gov/fulltext/EJ1018171.pdf

Hamlen, K. R. (2011). Children's choices and strategies in video games. *Computers in Human Behavior, 27*(1), 532-539. doi: 10.1016/j.chb.2010.10.001

Heinz, J. (2016). Digital skills and the influence of students' socio-economic background: An exploratory study in German elementary schools. *Italian Journal of Sociology of Education, 8*(2), 186-212. doi: 10.14658/pupj-ijse-2016-2-9

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management, 24*(5), 623-641. doi: 10.1177/014920639802400504

Hu, X., Gong, Y., Lai, C., & Leung, F. K. (2018). The relationship between ICT and student literacy in mathematics, reading, and science across 44 countries: A multilevel analysis. *Computers & Education, 125*, 1-13. doi: 10.1016/j.compedu.2018.05.021

İlgün-Dibek, M., Yalçın, S., & Yavuz, H. Ç. (2016). Matematik okuryazarlığı ile bilgi ve iletişim teknolojileri kullanım becerileri arasındaki ilişki: PISA 2012. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi (KEFAD)-Ahi Evran University Journal of Kırşehir Education Faculty, 17*(3), 39-58. http://kefad.ahievran.edu.tr/Kefad/ArchiveIssues/PDF/ad952109-a151-e711-80ef-00224d68272d adresinden edinilmiştir.

Jang, S. J. (2009). Exploration of secondary students' creativity by integrating web-based technology into an innovative science curriculum. *Computers & Education, 52*(1), 247-255. doi: 10.1016/j.compedu.2008.08.002

Karabay, E., Yıldırım, A., & Güler, G. (2015). Yıllara göre PISA matematik okuryazarlığının öğrenci ve okul özellikleri ile ilişkisinin aşamalı doğrusal modeller ile analizi. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi, 1*(36), 137-151. http://dergipark.org.tr/tr/pub/maeuefd/issue/19409/206317 adresinden edinilmiştir.

Kayaduman, H., Sırakaya, M., & Seferoğlu, S. S. (2011, Şubat). *Eğitimde FATİH projesinin öğretmenlerin yeterlik durumları açısından incelenmesi.* XIII. Akademik Bilişim Konferansı- XIII. Conference of Academic Informatics, Malatya, Türkiye.

Kim, H. S., Kil, H. J., & Shin, A. (2014). An analysis of variables affecting the ICT literacy level of Korean elementary school students. *Computers & Education, 77*, 29-38. doi: 10.1016/j.compedu.2014.04.009

Kubiatko, M., & Vlckova, K. (2010). The relationship between ICT use and science knowledge for Czech students: A secondary analysıs of PISA 2006. *International Journal of Science and Mathematics Education, 8*(3), 523-543. doi: 10.1007/s10763-010-9195-6

Lazakidou, G., & Retalis, S. (2010). Using computer supported collaborative learning strategies for helping students acquire self-regulated problem-solving skills in mathematics. *Computers & Education, 54*(1), 3-13. doi: 10.1016/j.compedu.2009.02.020

Lei, J., & Zhao, Y. (2007). Technology uses and student achievement: A longitudinal study. *Computers & Education, 49*(2), 284-296. doi: 10.1016/j.compedu.2005.06.013

Li, X., & Atkins, M. S. (2004). Early childhood computer experience and cognitive and motor development. *Pediatrics, 113*(6), 1715-1722. doi: 10.1542/peds.113.6.1715

Livingstone, S. (2012). Critical reflections on the benefits of ICT in education. *Oxford Review of Education, 38*(1), 9-24. doi: 10.1080/03054985.2011.577938

Luu, K., & Freeman, J. G. (2011). An analysis of the relationship between information and communication technology (ICT) and scientific literacy in Canada and Australia. *Computers & Education, 56*(4), 1072-1082. doi: 10.1016/j.compedu.2010.11.008

McMahon, G. (2009). Critical thinking and ICT integration in a Western Australian secondary school. *Educational Technology & Society, 12*(4), 269-281. Retrieved from https://www.j-ets.net/collection/published-issues/12_4

Milli Eğitim Bakanlığı. (2010). *PISA 2009 Ulusal ön raporu.* Ankara: Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı.

Milli Eğitim Bakanlığı. (2011). *İlköğretim okullarındaki (1-5. Sınıf) ödev uygulamalarının değerlendirilmesi araştırması.* Ankara: Eğitim Araştırma ve Geliştirme Dairesi Yayınları.

Milli Eğitim Bakanlığı. (2013a). *Ortaokul öğretim matematik dersi (5, 6, 7 ve 8. sınıflar) öğretim programı.* Ankara: MEB Yayınları.

Milli Eğitim Bakanlığı. (2013b). *PISA 2012 Ulusal ön raporu.* Ankara: Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü.

Murphy, D. (2016). A literature review: The effect of implementing technology in a high school mathematics classroom. *International Journal of Research in Education and Science (IJRES), 2*(2), 295-299. Retrieved from https://www.ijres.net/index.php/ijres/article/view/109/73

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics

National Council of Teachers of Mathematics. (2014). *Principles to actions: Ensuring mathematical success for all*. Reston, VA: National Council of Teachers of Mathematics

O'Neil, H. F., Wainess, R., & Baker, E. L. (2005). Classification of learning outcomes: Evidence from the computer games literature. *The Curriculum Journal, 16*(5), 455-474. doi: 10.1080/09585170500384529

OECD.Stat (2018). *Student-teacher ratio and average class size.* Retrieved from: https://stats.oecd.org/Index.aspx?DataSetCode=EAG_PERS_RATIO#.

Olkun, S., & Altun, A. (2003). İlköğretim öğrencilerinin bilgisayar deneyimleri ile uzamsal düşünme ve geometri başarıları arasındaki ilişki. *The Turkish Online Journal of Educational Technology, 2*(4), 86-91. http://www.tojet.net/volumes/v2i4.pdf#page=86 adresinden edinilmiştir.

Organisation for Economic Co-operation and Development. (2012). *PISA 2009 technical report.* Paris: OECD Publishing. doi: 10.1787/9789264167872-en

Organisation for Economic Co-operation and Development. (2013). *PISA 2012 assessment and analytical framework. Mathematics, reading, science, problem solving and financial literacy*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development. (2014a). *PISA 2012 results: What students know and can do (Volume I, Revised Edition, February 2014): Student performance in mathematics, reading and science.* Paris: OECD Publishing. Retrieved from https://www.oecd.org/pisa/keyfindings/pisa-2012-results-volume-I.pdf

Organisation for Economic Co-operation and Development. (2014b). *PISA 2012 technical report.* Paris: OECD Publishing. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Organisation for Economic Co-operation and Development. (2014c). *Scaling procedures and construct validation of context questionnaire data.* Paris: OECD Publishing. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA%202012%20Technical%20Report_Chapter%2016.pdf

Organisation for Economic Co-operation and Development. (2017). *PISA 2015 assessment and analytical framework. Science, reading, mathematic, financial literacy and collaborative problem solving, revised edition.* Paris: OECD Publishing. doi: 10.1787/9789264281820-en

Organisation for Economic Co-operation and Development. (2018a). *PISA database.* Retrieved from http://www.oecd.org/pisa/data/pisa2009database-downloadabledata.htm

Organisation for Economic Co-operation and Development. (2018b). *PISA database.* Retrieved from http://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm

Özer, Y., & Anıl, D. (2011). Öğrencilerin fen ve matematik başarılarını etkileyen faktörlerin yapısal eşitlik modeli ile incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi- H. U. Journal of Education,* (41), 313-324. http://www.efdergi.hacettepe.edu.tr/shw_artcl-702.html adresinden edinilmiştir.

Pamuk, S., Çakır, R., Ergun, M., Yılmaz, H. B., & Ayas, C. (2013). Öğretmen ve öğrenci bakış açısıyla tablet pc ve etkileşimli tahta kullanımı: FATİH projesi değerlendirmesi. *Kuram ve Uygulamada Eğitim Bilimleri - Educational Sciences: Theory & Practice, 13*(3), 1799-1822.

Pandolfini, V. (2016). Exploring the impact of ICTs in education: Controversies and challenges. *Italian Journal of Sociology of Education, 8*(2), 28-53. doi: 10.14658/pupj-ijse-2016-2-3

Petko, D., Cantieni, A., & Prasse, D. (2017). Perceived quality of educational technology matters: A secondary analysis of students' ICT use, ICT-related attitudes, and PISA 2012 test scores. *Journal of Educational Computing Research, 54*(8), 1070-1091. doi: 10.1177/0735633116649373

Raudenbush, S. W., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd edition). London: Sage.

Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis* (3rd ed.). New York, NY: McGraw-Hill

Scheuermann, F., & Pedró, F. (Eds.). (2009). *Assessing the effects of ICT in education: Indicators, criteria and benchmarks for international comparisons.* Luxembourg: Publications Office of the European Union.

Seferoğlu, S. S. (2015). Okullarda teknoloji kullanımı ve uygulamalar: Gözlemler, sorunlar ve çözüm önerileri. *Artı Eğitim, 123*, 90-91. http://www.egitimtercihi.com/okulgazetesi/17207-okullarda-teknoloji-kullan-m-ve-uygulamalar.html adresinden edinilmiştir.

Shaikh, Z., A., & Khoja, S. A. (2011). Role of ICT in shaping the future of Pakistani higher education system. *The Turkish Online Journal of Educational Technology, 10*(1), 149-161.

Skryabin, M., Zhang, J., Liu, L., & Zhang, D. (2015). How the ICT development level and usage influence student achievement in reading, mathematics, and science? *Computers & Education, 85*, 49-58. doi: 10.1016/j.compedu.2015.02.004

Subrahmanyam, K., Greenfield, P., Kraut, R., & Gross, E. (2001). The impact of computer use on children's and adolescents' development. *Journal of Applied Developmental Psychology, 22*(1), 7-30. doi: 10.1016/S0193-3973(00)00063-0

Sutherland, R., Robertson, S., & John, P. (2009). *Improving classroom learning with ICT*. London: Routledge.

Şengül, M., & Demir, E. (2018). Farklı ülkelerdeki öğrencilerin bilgi-iletişim teknolojilerine aşinalıklarının çeşitli değişkenlere göre sınıflama doğruluklarının incelenmesi. *Electronic Journal of Social Sciences, 17*(68), 386-1409. doi: 10.17755/esosder.345757

Trucano, M. (2005). *Knowledge maps: ICT in education.* Washington, DC: infoDev/World Bank.

Ural, A. (2015). Ortaokul matematik öğretmenlerinin bilgi iletişim teknolojisi ve psikomotor beceri kullanımlarının incelenmesi. *Turkish Journal of Computer and Mathematics Education, 6*(1), 93-116. doi: 10.16949/turcomat.18249

von Secker, C. E., & Lissitz, R. W. (1999). Estimating the impact of instructional practices on student achievement in science. *Journal of Research in Science Teaching, 36*(10), 1110-1126. doi: 10.1002/(SICI)1098-2736(199912)36:10<1110::AID-TEA4>3.0.CO;2-T

Wittwer, J., & Senkbeil, M. (2008). Is students' computer use at home related to their mathematical performance at school? *Computers & Education, 50*(4), 1558-2571. doi: 10.1016/j.compedu.2007.03.001

Yenilmez, K., & Dereli, A. (2009). İlköğretim okullarında matematiğe karşı olumsuz önyargı oluşturan etkenler. *e-Journal of New World Sciences Academy Education Sciences, 4*(1), 25-33. http://dergipark.org.tr/tr/pub/nwsaedu/issue/19829/212470 adresinden edinilmiştir.

Yorgancı, S., & Terzioğlu, Ö. (2013). Matematik öğretiminde akıllı tahta kullanımının başarıya ve matematiğe karşı tutuma etkisi. *Kastamonu Eğitim Dergisi - Kastamonu Education Journal, 21*(3), 919-930. http://dergipark.org.tr/tr/pub/kefdergi/issue/22605/241582 adresinden edinilmiştir.

Youssef A. B., & Dahmani M. (2008). The impact of ICT on student performance in higher education: Direct effects, indirect effects, and organizational change. *Revista de Universidad y Sociedad del Conocimiento, 5*(1), 45-56. doi: 10.7238/rusc.v5i1.321

Yusuf, M. O., & Afolabi, A. O. (2010). Effects of computer assisted instruction (CAI) on secondary school students' performance in biology. *The Turkish Online Journal of Educational Technology, 9*(1), 62-69. Retrieved from https://files.eric.ed.gov/fulltext/EJ875764.pdf

Zengin, Y., Kağızmanlı, T. B., Tatar, E., & İşleyen, T. (2013). Bilgisayar destekli matematik öğretimi dersinde dinamik matematik yazılımının kullanımı. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 10*(23), 167-180. http://dergipark.org.tr/tr/pub/mkusbed/issue/19550/208256 adresinden edinilmiştir.

Zhang, D., & Liu, L. (2016). How does ICT use influence students' achievements in math and science over time? Evidence from PISA 2000 to 2012. *Eurasia Journal of Mathematics, Science & Technology Education, 12*(9), 2431-2449. doi: 10.12973/eurasia.2016.1297a

Ziya, E., Doğan, N., & Kelecioğlu, H. (2010). What is the predict level of which computer using skills measured in PISA for achievement in math. *The Turkish Online Journal of Educational Technology,* 9(4), 185-191. Retrieved from https://files.eric.ed.gov/fulltext/EJ908084.pdf

Bilgi ve İletişim Teknolojilerine Erişim Düzeyi ve Kullanımı PISA 2009 ve 2012 Öğrenci Başarısını Nasıl Etkiler?

Giriş

Matematik öğretme ve öğrenme sürecinde bilgisayarların kullanımının önemi yıldan yıla artış göstermekte ve bilgi ve iletişim teknolojilerinin (BİT) matematik başarısını olumlu yönde etkileyeceği düşüncesi ile BİT'e ilişkin ciddi miktarlarda yatırımlar yapılmaktadır (Anderson, 2008; Kim, Kil, & Shin, 2014; Scheuermann & Pedró, 2009). Yapılan yatırımların ve sonuçların hem ulusal hem de uluslararası boyutta PISA (Uluslararası Eğitim Değerlendirme Testi) ve TIMMS (Uluslararası Matematik ve Fen Eğilimleri Araştırması) gibi uygulamalar ile değerlendirilmesine ve BİT ile akademik başarı arasındaki ilişkiye yönelik çalışmalar hız kazanmaya başlamıştır (OECD, 2014b; Skryabin, Zhang, Liu, & Zhang, 2015; Şengül & Demir, 2018). BİT'e dayalı öğretim ve öğrenme ile başarı arasındaki ilişkiyi belirlemek amacıyla yapılan çalışmalardan kesin bir sonucun elde edilemediği ve bu çalışmaların sonuçlarının birbiri ile tutarsız olduğu görülmüştür (Balanskat, Blamire, & Kefala, 2006; Cox & Marshall, 2007; De Witte & Rogge, 2014; Skryabin ve diğerleri, 2015; Trucano, 2005). Ayrıca bu tür araştırmalar, genellikle bireysel ve basit düzeydedir. BİT'in

başarıyı nasıl etkilediğine ve başarıda hangi BİT değişkenlerinin rol oynadığına yönelik çok düzeyli yaklaşımların yer aldığı çalışmalar ise oldukça azdır (Pandolfini, 2016). Ek olarak, bu tür çalışmalarda genellikle PISA uygulamasının tek yılına odaklanılmıştır (örneğin, Demir & Kılıç, 2009; Güzeller & Akın, 2014; Petko, Cantieni & Prasse, 2017). BİT değişkenlerinin öğrencinin matematik başarısını açıklama düzeyini farklı yıllarda uygulanan PISA verilerine göre karşılaştıran bir çalışmaya rastlanılmamıştır. Bunun bir nedeni PISA'da farklı yıllarda odaklanılan alanın değişmesi olabilir ancak az soruyla da olsa tüm alanların her yıl ölçüldüğü de bir gerçektir. Bu çalışmada da PISA 2009 ve 2012 uygulamalarında öğrencilerin matematik başarılarının bilgi ve iletişim teknolojilerine erişim ve kullanım düzeyleri açısından değerlendirilmesi amaçlanmaktadır. Her PISA uygulamasında okuma, fen ve matematik okuryazarlığından birine odaklanılmaktadır. PISA 2012 uygulamasında matematik okuryazarlığına odaklanılırken, PISA 2009'da okuma okuryazarlığa odaklanılmıştır. Böylece odaklanılan alana bağlı olarak, BİT değişkenlerinin matematik başarısını açıklama oranı belirlenebilecektir. Bu bağlamda, bu çalışma ile PISA 2009 ve 2012 sonuçlarına göre, Türkiye'deki öğrencilerin bilgi ve iletişim teknolojilerine erişim ve kullanım düzeylerinin matematik başarısını açıklama oranının belirlenmesi amaçlanmaktadır. Çalışmanın amacı doğrultusunda araştırma soruları ise şunlardır:

1. PISA 2009 ve 2012 Türkiye verisine göre, matematik başarısındaki değişkenliğin okullar arasındaki farklılıklar ve öğrenciler arasındaki farklılıklar tarafından açıklanma oranı nedir?
2. PISA 2009 ve 2012 Türkiye verisine göre, matematik başarısındaki değişkenliğin öğrenci düzeyinde ele alınan bilgi ve iletişim teknolojilerine erişim ve kullanımı ile ilgili değişkenler tarafından açıklanma oranı nedir?
3. PISA 2009 ve 2012 Türkiye verisine göre, matematik başarısındaki değişkenliğin okul düzeyinde ele alınan bilgi ve iletişim teknolojilerine erişim ve kullanımı ile ilgili değişkenler tarafından açıklanma oranı nedir?
4. PISA 2009 ve 2012 Türkiye verisine göre, matematik başarısındaki değişkenliğin hem öğrenci düzeyindeki hem de okul düzeyindeki BİT'e ilişkin değişkenler tarafından açıklanma oranı nedir?

Bu çalışmanın, alan yazına çeşitli açılardan katkı sağlayacağı düşünülmektedir. Bu katkılar: (a) BİT'in erişim ve kullanım düzeyinin matematik başarısı üzerindeki etkisinin açıklığa kavuşabilmesidir. (b) Alan yazında, BİT değişkenlerinin matematik başarısındaki varyans açıklama oranının farklı yıllar açısından karşılaştıran çalışmaların eksik olduğu görülmektedir. Bu çalışmada öğrencilerin matematik başarısında açıklanan varyans oranının belirlenmesinde etkili olan BİT değişkenleri farklı yıllar açısından araştırılmıştır. Açıklanan varyans oranı, BİT'in matematik eğitiminde etkili kullanımına dair bir fikir verilebilir. (c) Bu araştırmada hiyerarşik doğrusal modeller oluşturulmuştur. PISA verisinin yapısı dikkate alındığında, hiyerarşik modellerin tahmini standart hatayı daha iyi kalibre ettiği için, daha doğru sonuçlara ulaşmak ve bulguları daha az hatayla yorumlamak açısından önemli olduğu söylenebilir. (d) PISA 2012 matematik alanına odaklanırken, PISA 2009 okuma alanına odaklanmıştır. Bu çalışma, BİT değişkenlerinin matematik başarısı üzerindeki etkisinin alana bağlı olarak değişip değişmediğini yorumlama fırsatı da sağlayacaktır. Bu nedenle, bu çalışmanın BİT'in matematik başarısı üzerindeki etkisine ilişkin bütüncül bir bakış açısı sunması bağlamında önemli olduğu düşünülmektedir.

Yöntem

Bu araştırmada ilişkisel araştırma modeli kullanılmıştır. Araştırmanın PISA 2009 uygulamasının örneklemi 56 il ve okul türlerine göre tabakalandırılması sonucu toplam 170 okuldan 4996 öğrenciden, PISA 2012 uygulamasının örneklemi ise 57 il ve okul türlerine göre tabakalandırılması sonucu 170 okuldan toplam 4848 öğrenciden oluşmaktadır. Araştırmada Türkiye'de uygulanan PISA 2009 ve PISA 2012 matematik başarı testinden ve her iki uygulamada öğrencilerin bilgi ve iletişim teknolojilerine yatkınlık (BİTY) anketindeki ortak indekslerden elde edilen veriler kullanılmıştır. BİTY anketindeki BİT'in evde bulunması (ICTHOME), BİT'in eğlence amaçlı kullanımı (ENTUSE)

ve BİT'in okul görevlerini yerine getirmek için evde kullanımı (HOMSCH), BİT'in okulda bulunması (ICTSCH), BİT'in okulda kullanılması (USESCH) indeksleri hem PISA 2009 hem de PISA 2012 uygulamasında yer alan ortak BİTY indeksleridir.

Araştırmada kullanılan PISA verilerinin hiyerarşik bir yapısı olduğu için veri analizinde iki düzeyli Hiyerarşik Lineer Modelleme (HLM) analizi kullanılmıştır. Modelin birinci düzeyinde öğrenci, ikinci düzeyinde okul değişkenleri ele alınmıştır. Ele alınan PISA verilerinin HLM analizi için varsayımları incelendiğinde, veri setindeki kayıp veri oranı düşük olduğu için kayıp verilerin atanmasında HLM programındaki kayıp veri yöntemlerinden faydalanılmıştır. Örneklem büyüklüğü dikkate alındığında, uç değerlerin atılmasına yönelik herhangi bir işlem yapılmamıştır. HLM'nin varsayımlarından çoklu bağlantı sorununun olup olmadığının belirlenmesine ilişkin birinci düzeyde (öğrenci) ve ikinci düzeyde (okul) yer alan bağımsız değişkenler arasındaki korelasyon katsayı değerleri hesaplanmıştır ve bu değerlerin 0.70'in altında olduğu saptanmıştır. Araştırmada birinci düzey değişkenleri grup ortalaması etrafında merkezileştirilirken; ikinci düzey değişkenleri genel ortalama etrafında merkezileştirilmiştir. HLM'in diğer bir varsayımında öğrenci düzeyindeki hataların ve okul düzeyindeki hataların dağılımının normalliği incelenmiştir. Bunun için histogram ve olasılık grafikleri (P-P plot veya Q-Q plot) elde edilmiştir ve bu grafiklerin 45 derecelik bir doğru oluşturduğu gözlemlenmiştir. Dolayısıyla her iki düzeydeki hataların normallik sayıltısı sağlanmıştır. Öğrenci düzeyi varyansların homojenliği için H istatistiği hesaplanmış ve p değeri manidar bulunmuştur. Bağımsızlık sayıltısı incelendiğinde de PISA 2009 matematik değişkeninde ve PISA 2012 matematik değişkeninde okul-içi hataların öğrenci düzeyindeki değişkenlerden bağımsız olduğu bulunmuştur.

Araştırmanın amacı doğrultusunda üç model kurulmuştur. Bu modeller sırasıyla tek yönlü varyans analizi rastgele etkiler modeli (boş model ya da yokluk modeli olarak da adlandırılmaktadır), rastgele katsayılar regresyon modeli ve kesişim ve eğim katsayılarının bağlı olduğu modeldir. Tek yönlü varyans analizi rastgele etkiler modeline birinci düzeye ve ikinci düzeye ait herhangi bir değişken eklenmemiştir ve birleştirilmiş model Eşitlik 1'de verilmiştir.

$$(Y_{ij}|/M_{2009}/M_{2012}) = \gamma_{00} + u_{0j} + r_{ij} \quad (1)$$

Rastgele katsayılar regresyon modeline öğrenci düzeyindeki matematik başarısında BİT değişkenlerinden kaynaklanan kısmını açıklamak için BİT'e evde ulaşabilirlik (ICTHOME), BİT'in eğlence amaçlı kullanılması (ENTUSE), BİT'in okul görevlerini yerine getirmek için kullanımı (HOMSCH) olmak üzere toplam üç değişken eklenmiştir ancak ikinci düzeye ait herhangi bir değişken eklenmemiştir ve birleştirilmiş model Eşitlik 2'de verilmiştir.

$$(Y_{ij}|M_{2009}/M_{2012}) = \gamma_{00} + \gamma_{10} * (ENTUSE_{\mathrm{ij}}) + \gamma_{20} * (HOMSCH_{\mathrm{ij}}) + \gamma_{30} * (ICTHOME_{\mathrm{ij}}) + u_{0\mathrm{j}} + u_{1\mathrm{j}} * (ENTUSE_{\mathrm{ij}}) + u_{2\mathrm{j}} * (HOMSCH_{\mathrm{ij}}) + u_{3\mathrm{j}} * (ICTHOME_{\mathrm{ij}}) + r_{\mathrm{ij}} \quad (2)$$

Kesişim ve eğim katsayılarının bağlı olduğu model, Türkiye'de öğrencilerin PISA 2009 matematik ve 2012 matematik başarısı ile ilişkili olan BİT'e yönelik öğrenci özelliklerinin, okulun BİT'e yönelik hangi özellikleri ile ilişkili olduğunu belirlemeye yöneliktir. Bu modele öğrenci düzeyindeki üç değişken ve okul düzeyindeki iki değişken eklenmiştir ve birleştirilmiş model Eşitlik 3'te verilmiştir.

$$(Y_{ij}|M_{2009}/M_{2012}) = \gamma_{00} + \gamma_{01} * (ICTSCH_{\mathrm{ij}}) + \gamma_{02} * (USESCH_{\mathrm{ij}}) + \gamma_{10} * (ENTUSE_{\mathrm{ij}}) + \gamma_{20} * (HOMSCH_{\mathrm{ij}}) + \gamma_{30} * (ICTHOME_{\mathrm{ij}}) + u_{0\mathrm{j}} + r_{\mathrm{ij}} \quad (3)$$

Sonuç ve Tartışma

Araştırmada öğrenci düzeyinde ele alınan değişkenlerden BİT'in eğlence amaçlı kullanımı ile PISA matematik başarısı arasında pozitif ve manidar bir ilişkinin olduğu saptanmıştır. Bilgisayarda oyun oynama gibi bilgisayarın eğlence amaçlı aktiviteler için kullanımı aileler tarafından zaman kaybı olduğu düşünülse de bu tür aktivitelerin aslında öğrencilerin bilişsel gelişiminde (Becker, 2000; Hamlen, 2011; Li & Atkins, 2004) ve görsel zekayı geliştirmede (Subrahmanyam, Greenfield, Kraut, & Gross, 2000) önemli olduğunu unutmamak gerekir ve bu durum başarıyı olumlu yönde de etkileyebilir. Öğrencinin okul görevlerini yerine getirmek amaçlı BİT kullanımı ile PISA 2009 matematik başarısı arasındaki ilişkinin negatif ve manidar olması sonucu, okul görevlerini yerine

getirmek amaçlı BİT kullanımının daha çok ödev içermesi ve öğrencilerin de genelde matematik ödevlerinde zorlanmaları (Güven & Demirçelik, 2013; MEB, 2011) ve bu durumun hem matematik dersine hem de ödevlere karşı olumsuz tutumlar oluşturması ile açıklanabilir (Yenilmez & Dereli 2009). Öğrencinin evde ve okulda BİT'e dayalı materyallere sahip olması ile PISA matematik başarısı arasındaki ilişkinin pozitif ve manidar olduğu sonucu, öğrencinin konu ile ilgili çeşitli kaynaklardan daha fazla bilgiye erişebilmeleri ile açıklanabilir (Kubiatko & Vlckova, 2010). Okulda BİT'in kullanımı ile PISA matematik başarısı arasında negatif ve manidar bir ilişkinin olması, okulların eğitim ile ilgili olan web sayfalarına erişimine izin vermemesi (Kubiatko & Vlckova, 2010), öğrencilerin derslerde BİT kullanımına aşina olmamaları (İlgün-Dibek, Yalçın, & Yavuz, 2016) ya da öğretmenlerin BİT yeterlikleri ve öğretim yöntemlerine ilişkin bilgilerinin eksik ya da yetersiz olması ile açıklanabilir (Baki, Yalçınkaya, Özpınar, & Uzun, 2009; Pandolfini, 2016).

Araştırmada hem öğrenci düzeyinde hem de okul düzeyinde ele alınan değişkenler BİT olanaklarına sahip olma ve bunların kullanımı şeklinde gruplandırıldığında, her iki düzeyde de BİT olanaklarına sahip olmanın başarıyı arttırdığı ancak BİT kullanımının başarıyı artırmada etkili olmadığı sonucuna ulaşılmıştır. Ayrıca araştırmada öğrenci düzeyinde ve okul düzeyindeki BİT değişkenleri ile PISA matematik başarısı arasındaki ilişkiye yönelik elde edilen sonuçların, alanyazındaki bazı çalışmalarla tutarlılık gösterirken, bazıları ile tutarlılık göstermediği görülmüştür. Bunun nedenleri metodolojik sınırlamalar (Cox & Marshall, 2007; De Witte & Rogge, 2014) ya da her ülkenin kendine özgü BİT kullanımına ilişkin eğitim politikalarının ve uygulamalarının olması ile açıklanabilir (Heinz, 2016; Skryabin ve diğerleri, 2015).

Öğrenci düzeyindeki ve okul düzeyindeki BİT değişkenlerinin başarıyı açıklama oranları karşılaştırıldığında, okul düzeyindeki BİT değişkenlerinin başarıyı açıklama oranının, öğrenci düzeyindeki BİT değişkenlerine göre daha fazla olduğu bulunmuştur. Bu bulgu, okul seviyesindeki müdürlerin BİT uygulamalarındaki farkındalıkları, okul kültürü, BİT'in okullarda nasıl kullanıldığı ile ilgili işbirliği, öğretmenlerin BİT yeterlikleri ve öğretim yöntemlerine ilişkin öğretmen eğitimi gibi faktörlerden kaynaklanabilir (Ertmer & Ottenbreit-Leftwich, 2010; Pandolfini, 2016).

Matematik okuryazarlığına PISA 2012'de odaklanırken, PISA 2009'da odaklanılmamıştır. BİT'in matematik başarısı üzerindeki etkisi, matematik alanına odaklanıldığı ve odaklanılmadığı yıllar açısından karşılaştırıldığında, öğrenci düzeyindeki BİT değişkenleri ile PISA matematik başarısı arasındaki ilişkinin değişmediği belirlenmiştir. Sadece öğrencinin okul görevlerini yerine getirmek amaçlı BİT kullanımı ile PISA 2009 matematik başarısı arasındaki ilişki manidarken, PISA 2012 için bu ilişki manidar bulunmamıştır. Okul düzeyindeki BİT değişkenlerinden USESCH değişkeni ile PISA 2009 ve PISA 2012 matematik başarısı arasındaki ilişki ayrı ayrı incelendiğinde de bu ilişkinin değişmediği saptanmıştır. Bu olumsuz ilişkilerin yıllara göre değişmemesinin nedeni, öğretmenlerin ya da okul yöneticilerinin güven ve yeterlilik eksikliği ve kaynaklara erişim ile ilgili karşılaşılan çeşitli engellerle ilgili olabilir. Hem öğrenci düzeyindeki hem de okul düzeyindeki değişkenlerin PISA 2009 matematik başarısı için etki büyüklükleri incelendiğinde, okul değişkenlerinden ICTSCH ve USESCH değişkenlerinin büyük etkiye, öğrenci düzeyi değişkenlerinden HOMSCH değişkeninin küçük etkiye ve ICTHOME ve ENTUSE değişkenlerinin ise önemsiz bir etkiye sahip olduğu bulunmuştur. PISA 2012 için okul düzeyi değişkenlerinin matematik başarısı üzerindeki etkisinin büyük olduğu, öğrenci düzeyi değişkenlerinin ise matematik başarısı üzerindeki etkisinin önemsiz olduğu saptanmıştır. Öğrenci düzeyindeki değişkenlerin başarı üzerindeki etkisinin önemsiz ve küçük olmasının nedeni, öğrencilerin BIT'in etkin kullanımındaki yetkinliği ve farkındalığı (Grüsser, Thalemann, & Griffiths, 2006) ve ebeveynlerin BİT kullanımına ilişkin görüşleri ile ilgili olabilir (Becker, 2000; Hamlen, 2011; Li & Atkins, 2004). ICTSCH değişkeninin okul düzeyinde etkisinin büyük olmasının nedeni, iyi bir öğrenme ortamının öğrencilerin başarısını olumlu etkilediği bakış açısı ile açıklanabilir (Youssef & Dahmani, 2008). USECH değişkeninin öğrencinin matematik başarısı üzerindeki etkisinin büyük olmasının nedeni de öğretmenlerin BİT'in kullanımıyla ilgili yeterliklerinden ve niteliklerinden kaynaklanabilir (Youssef & Dahmani, 2008). PISA 2009 ve PISA 2012 matematik başarısına ilişkin sonuçların karşılaştırılmasında, sınavın matematik odaklı olup olmamasının, başka bir ifade ile matematik başarı testinin uzun ya da kısa olmasının, ciddi bir fark oluşturmadığı da söylenebilir. Hem öğrenci hem de okul düzeyindeki BİT değişkenlerinin, PISA 2009 matematik başarısındaki

değişkenliği açıklama oranı %27 iken, PISA 2012 matematik başarısındaki değişkenliği açıklama oranı %31 olarak bulunmuştur. Açıklama varyansındaki artışın az miktarda olduğu görülmektedir. Az miktardaki varyans artışının nedeni ise, öğrencinin evde ve okulda BİT kullanımını şekillendiren ailelerin, öğretmenlerin ve yöneticilerin BİT'in kullanımına ilişkin farkındalıklarının az da olsa artması ile açıklanabilir.

Araştırma sonuçlarından öğrencilerin okul görevlerini yerine getirmek amacıyla evde BİT'i kullanmaları ile matematik başarısı arasında negatif bir ilişkinin olması, bir sorun olarak karşımıza çıkmaktadır. Bu sorunun çözümü için öğretmenler, öğrencilere matematiği sevmelerine yardımcı olabilecekleri ve eğlence içerikli öğelerin matematik ödevlerinde kullanabilmelerini sağlayacak şekilde ödevlerin içeriği değiştirilebilirler. Ayrıca öğrenciler de ödevlerini yaparken çevrimiçi materyalleri okul görevlerinde kullanımı açısından yönlendirilmelerine gerek duyulmaktadır. Bu durumda hem öğretmenlere hem de evde ailelere önemli sorumluluklar düşmektedir. Evde ailelerin, çocuklarını BİT kullanma şekilleri açısından izlemeleri ve çocuklarını çevrim içi kaynak kullanımı konusunda bilinçlendirmeleri gerekmektedir.

Okulda BİT'in kullanımı ile matematik başarısı arasındaki negatif ilişki, diğer bir sorundur. Bu sorunu giderebilmek için, eğlence amaçlı BİT kullanımı derslere dahil edilebilir. Okulda matematik dersinde başarıyı artırmaya yönelik oyunlar seçilebilir. Ayrıca öğretmenlerin BİT'e ilişkin yeterliliklerini geliştirmeleri de önem kazanmaktadır. Dolayısıyla öğretmenlerin BİT'e ilişkin yeterliklerini geliştirmeleri için hizmet içi eğitimlere katılmaları teşvik edilmelidir. Ayrıca öğretmenlerin ders ortamında BİT'i etkili kullanmaya ve ailelerin de BİT kullanımına ilişkin farkındalıklarının artırılmasına yönelik projelere ihtiyaç duyulmaktadır. Bu araştırmada BİT'in kullanım şeklini ve başarısını etkileyen öğrencinin sosyo ekonomik geçmişi, yaşı ve cinsiyeti, öğrenme beklentileri gibi faktörler ele alınmamıştır. Bu değişkenler de modele eklenerek, BİT'in başarıya etkisini belirlemeye ilişkin çok düzeyli çeşitli çalışmalar yapılabilir.

Appendix. Tables Referenced in the Text

Table 1. The Correlation Matrix for the Level 1 and Level 2 Variables

Levels of Variables	Years	Predictor Variables	ICTHOME	ENTUSE	HOMSCH
The level 1 (student)	2009	ICTHOME	1		
		ENTUSE	.62	1	
		HOMSCH	.45	.63	1
	2012	ICTHOME	1		
		ENTUSE	.43	1	
		HOMSCH	.30	.53	1
Levels of variables	Years	Predictor variables	ICTSCH	USESCH	
The level 2 (school)	2009	ICTSCH	1		
		USESCH	.35	1	
	2012	ICTSCH	1		
		USESCH	.22	1	

Table 2. Fixed Effects Estimates and One-way Variance Analysis Random Effects Model

Fixed Effects	Coefficient	Standard Error	t-ratio	df
PISA 2009 average school mean, γ_{00}	436.12	5.94	73.36*	169
PISA 2012 average school mean, γ_{00}	439.90	5.71	77.04*	169

* $p < .001$

Table 3. Estimation of Variance Components of the One-Way ANOVA Model with Random Effect

Outcome Variables	Random Effect	Standard Deviation	Variance Component	df	χ^2
PISA 2009 mathematics achievement	INTRCPT (School average), u_{0j}	76.13	5795.96	169	7039.26*
	level-1 effect, r_j	59.18	3502.58		
PISA 2012 mathematics achievement	INTRCPT (School average), u_{0j}	72.99	5327.39	169	8427.38*
	level-1 effect, r_j	56.20	3158.00		

* $p < .001$

Table 4. Interclass and Intraclass Correlation Coefficient Calculations

Mathematics Achievement Scores	Interclass and Intraclass Correlation Coefficient Calculations
PISA 2009 mathematics achievement	ρ (interclass) = $\tau_{00} / (\tau_{00} + \sigma^2)$ = 5795.96 / (5795.96 + 3502.58) = 0.62
	ρ (intraclass) = $\sigma^2 / (\sigma^2 + \tau_{00})$ = 3502.58 / (3502.58 + 5795.96) = 0.38
PISA 2012 mathematics achievement	ρ (interclass) = $\tau_{00} / (\tau_{00} + \sigma^2)$ = 5327.39 / (5327.39 + 3158.00) = 0.62
	ρ (intraclass) = $\sigma^2 / (\sigma^2 + \tau_{00})$ = 3158.00 / (3158.00 + 5327.39) = 0.38

Table 5. Estimation of Fixed Effects on Random Coefficients Model in the Student Level

Fixed Effects	Coefficient	Standard error	t-ratio	df	Effect Size
PISA 2009 mathematics achievement average, γ_{00}	436.08	5.95	73.31*	169	
Average ENTUSE effect, γ_{10}	3.85	0.92	4.17*	4510	.07
Average HOMSCH effect, γ_{20}	-8.77	0.99	-8.85*	4510	-.15
Average ICTHOME effect, γ_{30}	6.39	0.94	6.80*	4510	.11
PISA 2012 mathematics achievement average, γ_{00}	439.89	5.71	77.03*	169	
Average ENTUSE effect, γ_{10}	4.04	0.76	5.29*	4477	.07
Average HOMSCH effect, γ_{20}	-1.60	0.97	-1.65	4477	
Average ICTHOME effect, γ_{30}	2.71	0.84	3.24*	4477	.05

* $p < .001$

Table 6. Estimation of the Variance Components on Random Coefficients Regression Model in the Student Level

Outcome Variables	Random Effect	Standard Deviation	Variance Component	df	χ^2
PISA 2009 mathematics achievement	Level-2 error term, u_0	76.21	5807.83	169	7241.57*
	Level-1 error term, r_{ij}	58.36	3405.48		
PISA 2012 mathematics achievement	Level-2 error term, u_0	73.01	5329.93	169	8535.79*
	Level-1 error term, r_{ij}	55.84	3118.09		

* $p < .001$

Table 7. Fixed Effects for Mathematics Achievement in the Intercept and Slopes as Outcomes Model

Fixed Effects	Coefficient	Standard Error	t-ratio	df	Effect Size
PISA 2009 mathematics achievement average, γ_{00}	435.69	5.01	86.94*	167	
Average ICTSCH effect, γ_{01}	52.91	13.04	4.06*	167	.69
Average USESCH effect, γ_{02}	-76.32	14.64	-5.21*	167	-1.00
Average ENTUSE effect, γ_{10}	3.85	0.90	4.29*	4510	.06
Average HOMSCH effect, γ_{20}	-8.77	0.97	-9.02*	4510	-.14
Average ICTHOME effect, γ_{30}	6.39	0.91	7.04*	4510	.08
PISA 2012 mathematics achievement average, γ_{00}	438.30	4.77	91.82*	167	
Average ICTSCH effect, γ_{01}	60.76	10.34	5.88*	167	.83
Average USESCH effect, γ_{02}	-57.65	8.59	-6.71*	167	-.78
Average ENTUSE effect, γ_{10}	4.04	0.76	5.28*	4477	.07
Average HOMSCH effect, γ_{20}	-1.60	0.97	-1.65	4477	
Average ICTHOME effect, γ_{30}	2.71	0.84	3.23*	4477	.05

* $p < .001$

Table 8. Random Effects for Mathematics Achievement in the Intercept and Slopes as Outcomes

Variables	Random Effect	Standard Deviation	Variance Component	df	χ^2
PISA 2009 mathematics achievement	Level-2 error term, u_0	64.83	4203.46	167	5599.33*
	Level-1 error term, r_{ij}	58.36	3405.35		
PISA 2012 mathematics achievement	Level-2 error term, u_0	60.47	3656.48	167	5901.47*
	Level-1 error term, r_{ij}	55.85	3119.47		

* $p < .001$

Table 9. Likelihood Ratio Test Results of Outcome Variables

Variables	Compared models	Calculating of Likelihood Ratio Test and Results
PISA 2009 mathematics achievement	For goodness of fit of model 1 - model 4:	x_1^2 = (52139.20 - 51959.54) / (169 - 167) = 89.83
	For goodness of fit of model 2 - model 4:	x_2^2 = (52012.49 - 51959.54) / (169 - 167) = 26.47
	For goodness of fit of model 3 - model 4:	x_3^2 = (52086.26 - 51959.54) / (169 - 167) = 63.36
PISA 2012 mathematics achievement	For goodness of fit of model 1 - model 4:	x_1^2 = (51293.51 - 51177.37) / (169 - 167) = 58.08
	For goodness of fit of model 2 - model 4:	x_2^2 = (51236.54 - 51177.37) / (169 - 167) = 29.58
	For goodness of fit of model 3 - model 4:	x_3^2 = (51234.33 - 51177.37) / (169 - 167) = 28.48

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(3); 243-265

Psychometric Properties of Turkish Version of Aggression Questionnaire Short Form: Measurement Invariance and Differential Item Functioning across Sex and Age

Yaşar KUZUCU * Özge SARIOT ERTÜRK **

Abstract
The aim of the present study was to test the psychometric properties of the Aggression Questionnaire Short Form for adolescents and adults in Turkish. The adaptation study was conducted with 778 adolescents aged between 15-18 and 1067 adults aged between 19 and 44. The construct validity of the questionnaire was tested via Parallel Analysis, Exploratory Factor Analysis and Confirmatory Factor Analysis. Furthermore, item-total correlations, test-retest score correlation, and internal consistency (Cronbach Alpha and McDonald's Omega) were calculated as reliability analyses. The Measurement Invariance test and Differential Item Functioning in male and female, adolescent and adult samples were also conducted. The results yielded that the Turkish version of the Aggression Questionnaire Short Form is a reliable questionnaire with four-factors, and without sex and age differences, it can be used to measure aggression among Turkish adolescents and adults.

Key Words: Aggression questionnaire short form, measurement invariance, differential item functioning

INTRODUCTION

Aggression is a multidimensional construct that develops within a complex interaction of biological, psychological, social, and cultural factors (Vitoratou, Ntzoufras, Smyrnis, & Stefanis, 2009) and has received great deal of attention in mental health area (Evren, Çınar, Güleç, Çelik, & Evren, 2011; Hinshaw; 1987; Johnson, Carve, & Joormann, 2013; Podubinski, Lee, Hollander, & Daffern, 2017). A large number of theoreticians and researchers tried to explain the origin and reason of aggression and association of aggression with other behaviors (Chang, Schwartz, Dodge, & McBride-Chang, 2003; Coie & Dodge, 1998; Maslow, 1943; Moyer, 1982; Sexton et al. 2019).

Several measurement tools were developed to measure this essential issue (Buss & Perry, 1992; Orpinas & Frankowski, 2001; Kang, Lim, Suh, Gang, & Pedersen, 2020; Palmstierna & Wistedt, 1987; Raine et al. 2006). The Buss-Perry Aggression Questionnaire (BPAQ; Buss & Perry, 1992) is one of the most frequently used measurement tool in the literature to measure aggression (Adıgüzel, Özdemir & Şahin, 2019; Kühn et al. 2019; Singh, 2017). Buss-Durkee Hostility Inventory (BDHI; Buss & Durkee, 1957) is the origin of the questionnaire. Researchers constructed BPAQ as a more current instrument in terms of psychometric properties. BPAQ is a 5-point Likert scale, consists of 29 items and has four factors. These factors are physical aggression, verbal aggression, anger, and hostility. Additionally, different from the other instruments developed to measure aggression, BPAQ has validity for both adolescent (Reyna, Sanchez, Ivacevich, & Brussino, 2011) and adult samples (Vitoratou et al. 2009). Moreover, it is used with both clinical (Evren et al. 2011) and nonclinical samples (Özdemir, Vazsonyi & Çok, 2017) rather than just with clinical or nonclinical ones (Palmstierna & Wistedt, 1987). BPAQ also provides valid and reliable data from offenders (Diamond, Wang & Buffington-Vollum, 2005). In terms of factor structure, the scale explains aggression with four structures that involve different forms of active and passive aggression, rather than just proactive or reactive aggression (Raine et al. 2006). The

* Associate professor, Aydın Adnan Menderes University, Education Faculty, Aydın-Turkey, e-mail: yasarku@yahoo.com, ORCID ID: 0000-0002-8487-9993
** Research assistant, Aydın Adnan Menderes University, Faculty of Science and Arts, Psychology Department, Aydın, Turkey, e-mail: ozge.sariot@adu.edu.tr, ORCID ID: 0000-0003-4565-8300

To cite this article:
Kuzucu, Y., & Sarıot-Ertürk, Ö. (2020). Psychometric Properties of Turkish Version of Aggression Questionnaire Short Form in Adolescents and Adults. *Journal of Measurement and Evaluation in Education and Psychology, 11*(3), 243-265. doi:

Received: 01.02.2020
Accepted: 17.09.2020

psychometric properties of the BPAQ were tested with different methodologies and samples, and research results confirmed the original four-factor structure of the questionnaire (Bernstein & Gesn, 1997; García-León et al. 2002; Gerevich, Bácskai, & Czobor 2007; Harris, 1997; Reyna et al. 2011; Torregrosa et al. 2020). However, most of the studies reported better fit to original factor structure or better factor loadings when some items are omitted (Bernstein & Gesn, 1997; Gerevich et al. 2007; Harris, 1995). Additionally, researchers reported BPAQ as an inadequate measurement tool because of the explained common variance by these four factors (Bryant & Smith, 2001).

In order to develop an acceptable measurement model for the BPAQ, Bryant and Smith (2001) refined the questionnaire and proposed a 12 item version (short form) of the Aggression Questionnaire (AQ-SF). The new short form of the AQ-SF also has a four-factor structure model with the same names, physical aggression, verbal aggression, anger arousal and hostility. Each factor includes three items. Unlike the BPAQ, the AQ-SF is a 6 point Likert questionnaire (Bryant & Smith, 2001). However, most of the studies (e.g., Maxwell, 2007; Torregrosa et al. 2020) which includes AQ-SF preferred the 5 point Likert type version.

As BPAQ, the psychometric properties of the AQ-SF (12 item version of AQ) was tested with different methods and samples. The AQ-SF showed good construct validity in the offenders (Diamond & Magaletta, 2006) and mentally ill male prisoners (Diamond et al. 2005). Sex invariance of the questionnaire was also confirmed for the Argentinean adolescents (Reyna et al., 2011) and federal offenders (Diamond & Magaletta, 2006). Maxwell (2007) tested validity on the translated Chinese version AQ-SF with Chinese sample. Results indicated a good fit to the data and adequate internal reliability. The Dutch version of AQ-SF also has sufficient validity and reliability in the psychiatric patient and the student samples (Hornsveld, Muris, Kraaimaat, & Meesters, 2009).

In addition to the good psychometric properties of the AQ-SF, remarkable relations with aggression and other mental health issues were reported in the studies that used the 12-item version of the AQ-SF. The relation between aggression and collective narcissism (De Zavala, Cichocka, Eidelson, & Jayawickreme, 2009), hubristic pride (Carver, Sinclair, & Johnson; 2010) mindfulness and rumination (Borders, Earleywine, & Jajodia, 2010) were pointed out. Johnson et al. (2013) reported significant relation of anger and verbal aggression dimensions with borderline personality characteristics, anxiety symptoms and alcohol consumption.

As in varied languages the Turkish 29 item version of the BPAQ was also studied. In order to test the psychometric properties of BPAQ, studies were conducted with college students (Madran, 2012), adolescents (Önen, 2016) and male substance dependent inpatients (Evren et al. 2011). Despite their different sample profiles, all have a common result; the Turkish version of the BPAQ is a valid and reliable questionnaire to measure aggression. However, no studies have been conducted to test the psychometric properties of the AQ-SF in Turkish.

The AQ-SF was reported as acceptable to use in different cultures, sexes, clinical and nonclinical samples. The relation of aggression with both well-being and ill-being variables was pointed out when aggression was measured through the AQ-SF. Taking into account all of these, it seems essential to introduce the AQ-SF into Turkish. Therefore, this study aims to test the construct validity and reliability of the AQ-SF and to test sex and age invariance of the questionnaire in the Turkish sample.

METHOD

This study, which aims to adapt the AQ-SF into Turkish, is a descriptive study. Descriptive studies attempt to explain “what” events, objects, entities, institutions, groups, and areas are (Fraenkel, Wallen & Hyun, 2012). In this descriptive study, the validity and reliability analyses were conducted, and the psychometric properties of AQ-SF were determined. Detailed information about participants, the data collection instrument, and data analysis are presented below.

Study Group and Process

The AQ-SF was implemented to 778 students between the ages of 15 and 18 from five different high schools. The self-report measures were administered to the participants at their school. Participants were volunteers, and no personal information was assembled. The whole data was collected two times for Parallel Analysis (PA), Exploratory Factor Analysis (EFA), and Confirmatory Factor analysis (CFA). PA and EFA were conducted with 383 adolescents. In order to CFA, the data from 395 adolescents were used.

AQ-SF was also applied to the adult group. The adult group consists of overall 1067 people, undergraduate students from Aydın Adnan Menderes University, University of Ege and University of Ankara, graduated from university and participated in pedagogical formation training and trainees in the public training center. Participants were determined by convenience sampling, and they were voluntarily participating. Two different data sets were used for PA, EFA (n= 648) and CFA (n= 419). The distribution of the study groups is given in Table 1 and Table 2.

Table 1. The Distribution of the Study Group for the EFA

Adolescent (15-18 years of age)			Adult (19-35 years of age)		
Sex	F	%	Sex	F	%
Male	98	26.41	Male	220	33,95
Female	273	73.58	Female	428	66,04
Total	371	100.0	Total	648	100.0
Age	F	%	Age	F	%
15	97	25.3	19-23	510	78.70
16	77	20.1	24-30	122	18.82
17	128	33.4	31-35	16	2.46
18	81	21.1	Total	648	100.0
Total	383	100.0			

Table 2. The Distribution of the Study Group for the CFA

Adolescent (15-18 years of age)			Adult (19-44 years of age)		
Sex	F	%	Sex	F	%
Male	165	41.1	Male	130	31.63
Female	230	57.4	Female	281	68.36
Total	395	100.0	Total	411	100.0
Age	F	%	Age	F	%
15	201	50.1	19-23	291	74.44
16	124	30.9	24-30	71	18.15
17	65	16.2	31-35	14	3.58
18	6	1.5	36-44	15	3.83
Total	396	100.0	Total	391	100.0

The Adaptation Procedure

The original questionnaire was independently translated from English into Turkish by four experts in psychology. In addition to the individual transition, using the focus group technique, each item was evaluated by the same experts. The group members are composed of people who know both languages and cultures, have measurement tool development skills, and know the purpose of the translated measurement tool. Consensus was reached on a common draft by these experts. Then back-translated by bilingual psychiatry and psychology experts who are different from the experts in the translation process.

Data Collection Instruments

Aggression

Aggression was measured by using the AQ-SF. The AQ-SF containing 12 items comprised the refined four-factor measurement model. This questionnaire was developed from Buss and Perry's 29-item AQ, and it has a four-factor structure; physical aggression, verbal aggression, anger, and hostility. The physical aggression, involves nine items, factor loadings of these items changes between .44 and .84. The verbal aggression involves five items and factor loadings of these items changes between .35 and .56. The anger, consists of seven items and these items' factor loadings change between .35 and .75. Lastly, the hostility involves eight items and their factor loadings change between .37 and .70 (Buss & Perry, 1992). Although Buss and Perry (1992) did not report explained variance for the AQ, Garcia-Leon et al. (2002) supported four-factor structure of the questionnaire and reported variance explained by the whole questionnaire as 42.1 %. Cronbach Alpha values of the factors and the total score for the AQ-SF are .85, .72, .83, .77, and .89, respectively. Moreover, test-retest reliability estimates are .80, .76, .72, .72, and .80 for the four factors and total score, respectively (Buss & Perry, 1992).

Bryant and Smith (2001) explored the factor structure of the AQ. The researchers deleted items that displayed low or multiple loadings in a principal component analysis and excluded a number of reverse-scored items. This procedure yielded the AQ-SF (12 item), for which the hypothesized four-factor model produced an acceptable fit. The AQ-SF has the same factor structure with the AQ. Each dimension had three items. However, Bryan and Smith (2001) did not report factor loadings, explained variance and test retest reliability of AQ-SF. In addition to obtaining dimension scores, a total aggression score can also be calculated. Cronbach Alpha values for the dimensions of the original AQ-SF change between .70 and .83. In the original form (Buss & Perry, 1992) the questionnaire is a 5 point Likert questionnaire and Bryant and Smith (2001) adopted the questionnaire to a 6-point response tool ranging from 1 (extremely uncharacteristic of me) to 6 (extremely characteristic of me). Despite the adaptation of Bryant and Smith (2001) in the current study, the original 5-point questionnaire (1 = uncharacteristic of me, 5 = very characteristic of me) was sustained likewise previous adaptation studies (Abd-El-Fattah, 2013; Maxwell, 2007; Torregrosa et al. 2020) in order to compare the present results with earlier researches in a credible way.

Social problem solving

The Social Problem Solving Inventory-Revised Short-Form (SPSI-RSF; D'Zurilla, Nezu, & Maydeu-Olivares, 2002) was used. The scale has 25 self- administered questions that are developed to assess cognitive, emotional or behavioral reactions of individuals to real life problem-solving situations. It has five dimensions, each involves five items, comprising two problem orientations as positive and negative, and three problem-solving styles, as rational, impulsive/carelessness, and avoidance. In terms of the validity, Sorsdahl, Stein, and Myers (2017) reported the variance explained by SPSI-RSF as 57.9%. The inventory has good internal consistency (α=.84), excellent test-retest reliability, (r=.90), and good discriminant validity tested on a sample of sexual offenders (Webster, Mann, Thornton, & Wakeling, 2007). The Turkish form of the tool (Eskin & Aycan, 2009) supported original factor structure. Factor loadings for positive orientation change between .52 and .67, for negative orientation .62 and .81, for rational orientation .60 and .72, for impulsive/carelessness orientation .38 and .76, lastly, for avoidance orientation .35 and .90. CFA results for Turkish form of the inventory is also acceptable; $\chi 2$ / df 2.15, RMSEA = .04, CFI = .92. In the adaptation study, the coefficients of internal consistency and test-retest reliability differed from .62 to .92 and from .60 to .84, respectively (Eskin & Aycan, 2009). In the present study, the coefficient of internal consistency is ranged from .68 to .90 for adolescents and .69 to .80 for adults.

Trait anger

Trait Anger was assessed using the 10-item subscale of the Anger Expression Scale (Spielberger, 1985). Trait Anger and Anger Expression Style Scale (STAXI) is a self-report scale comprised of 44 items; 10 items of this 44 item scale define trait anger, 10 items define state anger, and 24 items define anger expression style (Anger control, Anger-out and Anger-in). The scale allows researchers to use each subscale independently. Trait Anger Scales (TAS) reports how angry they generally feel. The TAS correlates positively with a variety of anger and hostility measures such as the Buss-Durkee Hostility Inventory and with various state anger measures and discriminates high from low anger groups (Spielberger, 1988). The reliability study of the STAXI-2 with adult males from the general population reports alpha coefficients ranging from .73–.95 for the total scale scores and from .73–.94 for the subscales (Spielberger et al., 1985). In Turkish adaptation study (Özer, 1994), for anger control, the coefficients of internal consistency were calculated as .84. In the present study, the coefficient of internal consistency is .83 for adolescents and .87 for adults.

Data Analysis

SPSS 25.0 (SPSS Inc.), Factor Analysis 10.10 (Fernando & Lorenza-Seva, 2017), LISREL 8.80 (Jöreskog & Sorbom, 1993) and jMetrik Version 4.1.1 statistical package programs were used in the analysis. The data were analyzed using PA, EFA, and CFA techniques for the construct validity. Furthermore, item-total correlations, test-retest score correlation, internal consistency estimates of reliability (Cronbach Alpha and McDonald's Omega) were calculated. T-test was performed to test whether the items of the questionnaire distinguished between the lower and upper 27% groups. By examining the measurement invariance (MI) in female-male and adolescent-adult samples, it was tested whether the measurement tool was appropriate for the comparisons between groups. In order to test the validity of the questionnaire by item, Differential Item Functioning (DIF) tests were conducted for sex and age groups. Expert opinion was used to determine what the source of the DIF is for an item that gives DIF (Doğan & Öğretmen, 2008).

RESULTS

AQ-SF Adolescent Application

To test the psychometric properties of the measurement in adolescent, validity and reliability analyses was conducted. All analyses were explained in detail.

Pre-analyses

In order to determine whether the data showed normal distribution or not, measures of central tendency, Skewness and Kurtosis values were examined. The results about central tendencies, showed that Mean = 29.17, Median = 29, and Mode = 30. The similarity of these scores indicates the normal distribution of the data (Büyüköztürk, 2007). For aggression total score Skewness is .11, and Kurtosis is -.23 (n=778, data set for PA, EFA, and CFA). The fact that both values are between the range of -1, +1 implies that they show normal distribution.

Kaiser Meyer Olkin (KMO) coefficient was used to determine whether the data structure was appropriate for factor analysis in terms of the sample size of the application. As a result, KMO value was determined as 0.79. The fact that KMO value is high means that each variable in the questionnaire can be estimated well by the other variables (Field, 2013). Bartlett's test of Sphericity was significant (χ^2 (66, n = 383) = 1261.459 p<.001), and this value supported the factorability of the correlation matrix. Another indicator of the appropriateness of the data for factor analysis is the Anti-image Correlation Matrix. These values need to be above 0.5, and the values below this must be excluded from the analysis (Field, 2013). The diagonal values for each variable in the anti-image matrix vary between .70 and .89.

The fact that all the values of the intersection point are above 0.5 indicates that it is accurate to include all the items in the questionnaire.

The validity analysis

The factor structure for the construct validity of the questionnaire was determined by performing PA and EFA. The purpose of performing PA and EFA is to gather the variables that are related to each other and that measure the same quality together, and to reduce the number of items forming the questionnaire (Aksu, Eser, & Güzeller, 2017; Horn, 1965). CFA was performed to test whether the restricted structure defined by PA and EFA was verified as a model (Horn, 1965; Tabachnick & Fidell, 2013).

When the factor structure of the questionnaire is analyzed via PA and EFA the scree plots are also examined. As can be seen in Figure 1 the graph curve shows a sharp decrease till the fourth factor and that the curve proceeds horizontally after the fourth factor. It indicates that this finding supports the four-factor structure of the questionnaire.

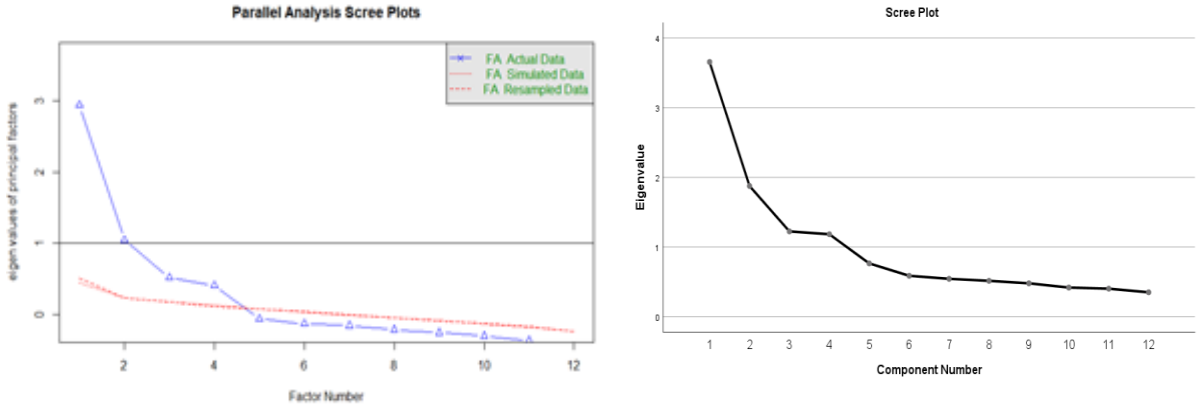


Figure 1. AQ-SF Parallel Analysis and EFA Scree Plots Graph of Adolescent Application

In PA, factor number is decided through comparing eigenvalues from real data and simulated random parallel data set that is produced based on the real data set. Factor number is accepted till the point in which the real data eigenvalue is larger than the parallel data eigenvalue (Akbaş, Karabay, Yıldırım-Seheryeli, Ayaz, & Demir, 2019). Depending on these explanations and the values mentioned in Table 3, the PA results indicated that, the adolescent application of AQ-SF has four factors.

Table 3. Eigenvalues from PA for Adolescents Application

Factors	1	2	3	4
Eigenvalues from sample correlation matrix	3.88	1.63	1.21	1.14
Average eigenvalues from parallel analysis	1.20	1.15	1.11	1.08
95th percentile eigenvalues from parallel analysis	1.25	1.19	1.14	1.10

Notes: n = 778

The result of the EFA with 12 items indicated that the items were collected in 4 sub-dimensions, with eigenvalues greater than 1. The items of each sub-dimension were examined, and it was determined that they were grouped under the factor to which they were related. To clarify the relationship among factors, the varimax rotation (the orthogonal rotation technique of Principal Component Analysis) is used. As a result of the EFA it was found that the eigenvalue of the factors from the first to the fourth were 2.12, 2.09, 1.88 and 1.83 respectively. Additionally, the variance explained by the factors from the first to the fourth were 17.74, 17.45, 15.70 and 15.25 respectively. The total variance explained by the

questionnaire was found at 66.16%. When the eigenvalues and cumulative variance percentages of the four factors were taken into consideration, it was determined that the questionnaire had four factors. The findings obtained as a result of the EFA performed for AQ-SF Adolescent Application revealed that the construct validity of the questionnaire was sufficient and factor structure was similar to the original form. The factors formed after EFA and the items collected under each factor are given in Table 4.

Table 4. Factor Loadings, Item-Total Correlations and Common Variances for Adolescent Application

Factors	PA Factor Loadings				EFA Factor Loadings				Item-Total Correlation	Common Variances
	1	2	3	4	1	2	3	4		
Physical	.80	-.25	.16	.00	.83	.19	.09	-.01	.45	.73
	.80	-.04	-.01	-.02	.82	.06	.18	.06	.46	.72
	.64	.03	.03	.02	.77	.21	.11	.11	.49	.67
Verbal	.04	.46	-.01	.04	.16	.79	.10	.03	.40	.67
	-.02	.79	-.04	-.03	.15	.76	.26	.01	.45	.67
	.01	.70	.04	-.00	.12	.65	-.03	.10	.30	.45
Anger	-.07	-.09	.70	-04	.01	-.10	.82	.08	.33	.69
	-.00	-.01	.70	05	.10	.23	.75	.14	.49	.65
	.08	.10	.63	-01	.24	.31	.65	.12	.56	.60
Hostility	-.06	.04	.05	.66	.00	.06	.06	.86	.39	.76
	.04	-.04	-.10	.94	.00	.04	.04	.84	.36	.72
	-.00	.02	.11	.51	-.01	.06	.22	.76	.38	.62

Notes: PA = Parallel Analysis, EFA= Exploratory Factor Analysis

When Table 4 is examined, the results of PA and EFA reveal that each item is clustered under a factor that is related to a value that is more than twice as much as the factor loading value that they have in other factors. This finding, which shows that the items differentiate in terms of factors, supports the construct validity of the questionnaire. As can be seen in Table 4, each factor is composed of the three items. The factor loadings of the first factor vary between .80 and .64 for PA, .83 and .77 for EFA. The factor loading of the second factor values varies between .46 and .70 for PA, .79 and .65 for EFA. The factor loadings of the third factor vary between. 70 and.63 for PA, .82 and .65 for EFA. The factor loadings of the fourth vary between .66 and .94 for PA, 86 and .76 for EFA. Following this phase, the items in each factor were examined as a whole, and a factor structure consistent with the original form of the questionnaire was observed. In order to determine whether there were significant correlations among the factors forming AQ-SF adolescent application, Pearson Correlation Analysis was performed. It was revealed that the relationship coefficients of "Physical aggression" factor with "Verbal Aggression", "Anger", and "Hostility" were found as .39, .38, and .25 respectively; and the relationship coefficient of "Verbal Aggression" with "Anger" and "Hostility" was found as .38 and .22 respectively; and lastly, the relationship coefficient between "Anger" and "Hostility" was determined as .34. The results obtained, consistent with the literature (Şahin, 2018), show a positive significant ($p \leq .001$) relationship among all the factors of the questionnaire.

First-order and second-order CFA was performed to evaluate the applicability of the four factors of AQ-SF Adolescent application. The models obtained from these analyses are given in Figure 2. Additional to the first and second-order CFA, 1- factor solution was also tested.

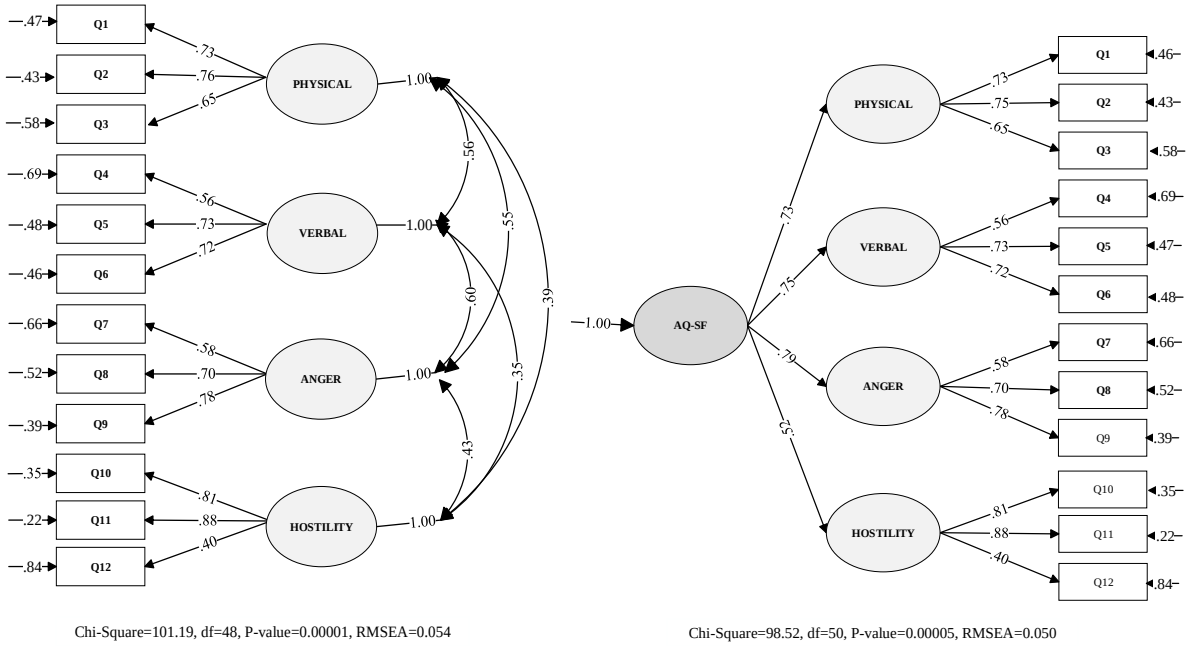


Figure 2. AQ-SF Adolescent 1st and 2nd Order CFA

First and second-order CFA were performed for four-factor structured AQ-SF adolescent application. When the CFA was evaluated, χ^2/sd ratios for the first and second-order were determined as 2 (χ^2/sd=96/48) and 1.97 (χ^2/sd=98.52/50), respectively. The fact that χ^2/sd ratios obtained as a result of first and second-order CFA are ≤2.0, correspond to a good fit. RMSEA fit index values were determined as 0.051 and 0.050 as a result of first and second-order CFA, respectively. The fact that RMSEA fit index value is below and equal to 0.05 can be interpreted as a good fit (Kline, 2015). It was determined that, among the fit index values related to the model as a result of the first and second order CFA, AGFI was 0.93, GFI was 0.96, standardized RMR fit index value was 0.059, NFI fit index value was 0.96, and CFI fit index value was 0.98. There is no statistically significant difference between first and second-order CFA (less than 3.84 chi-square difference with one degree of freedom); however, the second-order was evaluated to be superior since it is more parsimonious. When all the values related to data fit of the model are taken into consideration, it can be seen that the model formed shows a sufficient order to fit with the data.

Another CFA was performed to support the multifactorial structure of AQ-SF adolescent application; the results of first and second-order factor analyses were compared with the one-factor analysis of the questionnaire. The questionnaire was assumed to have one dimension, and it produced the following statistics: χ^2/sd ratio of the fit values used in the model comparisons was calculated as 9.41 (χ^2/sd=508.48/54, RMSEA= 0.15, GFI= 0.82, NFI= 0.79, CFI = 0.81). The results showed that the one-factor structure had poorer fit values than the multifactorial structure.

In order to determine the convergent validity of AQ-SF adolescent application, the relationship between AQ-SF scores with trait anger scores was examined with Pearson Product-Moment Correlation Analysis. The correlation of the AQ-SF with trait anger (r=.54) is moderate and statistically significant (p≤.001). Additionally, to determine the divergent validity of AQ-SF, the relationship between AQ-SF scores and social problem-solving scores was examined in the same way. Results showed a negative (r=-.30) and statistically significant (p≤.001) relationship between the two variables.

The reliability analysis

Item analysis was conducted with all adolescent data (n=778) to determine the contribution of the items in the questionnaire of the implicit structure they belong to and to measure the level of discrimination between the items with and without relevant characteristics of the structure they belong to (Erkuş, 2012).

The Cronbach Alpha and McDonald's Omega coefficients were calculated for all and each factor of the questionnaire. It is suggested that McDonald's Omega coefficient is more appropriate for multi-dimensional measures (Revelle, 2018). Cronbach Alpha and McDonald's Omega coefficients have following values for the first factor .76 and.73, for second factor .68 and .70, for the third factor .70 and .60, for the fourth factor, .74 and .64 respectively. Cronbach Alpha and McDonald's Omega were calculated as .80 and .76 for the total score.

Test-retest reliability was found as .99. Item total correlation coefficients varied between .57 and .62 for the first factor, .41 and .56 for the second factor, .49 and .54 for the third factor, .49 and .67 for the fourth factor, .30 and .56 for total.

It was also analyzed whether there was a significant difference between individuals with low scores and high scores. As a result of the *t*-test conducted to compare the responses of the individuals in the lower 27% group and the responses of the individuals in the upper 27% group to all the items in the questionnaire, the items' *t* values varied between 62.73 ($p<.001$) and 32.96 ($p<.001$) and a significant difference was found. In the analysis performed, it was found that the variances were heterogeneous. It can be seen that the reliability values of the overall and factors of the AQ-SF adolescent application are generally acceptable for social sciences.

AQ-SF Adult Application

To test the psychometric properties of the measurement in adults, validity and reliability analyses were conducted. All analyses were explained in detail.

Pre-analyses

In a similar manner with the adolescent application analysis, measures of central tendency, Skewness, and Kurtosis values were examined. The central tendency results showed that Mean = 29, Median = 29, and Mode = 29. Skewness and Kurtosis values were examined (n = 1067, data set for PA, EFA, and CFA). Skewness was found .30, and Kurtosis was found .02. As for the data of adolescence, the similarity of central tendency measures, Skewness, and Kurtosis values indicated normality for data of adults.

KMO value was determined as 0.78. It means that each variable can be estimated well by the other variable. Bartlett's test of Sphericity was significant (χ^2 (66, n = 648) = 1985.553 $p<.001$) and this value supported the factorability of the correlation matrix. Besides, the Anti-Image Correlation Matrix intersection values were also analyzed and it was found that these values varied between .68 and .89. As the values at this intersection point were above 0.5, it was determined that it was accurate to include all the items in the questionnaire.

The validity analysis

PA and EFA were conducted for adult application data, too. When the "Scree Plots" graphs were examined (Figure 3), it can be seen that the curves show a sharp decrease till the fourth factor and that the curve proceeds horizontally after the fourth factor. The results are consistent with the previous results showing that the questionnaire has a four-factor structure.

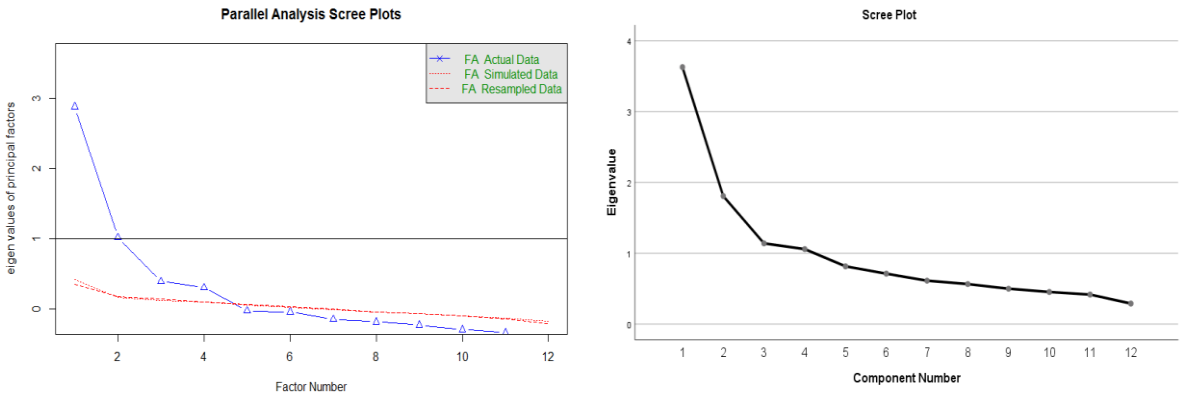


Figure 3. AQ-SF Parallel Analysis and EFA Scree Plots Graph of Adult Application

Accordingly, to the PA results, when eigenvalues from real data and stimulated parallel data were compared (see in Table 5), it indicates that consistent with the original structure, the adult application of AQ-SF has four factors.

Table 5. Eigenvalues from PA for Adult Application

Factors	1	2	3	4
Eigenvalues from sample correlation matrix	3.63	1.74	1.15	1.10
Average eigenvalues from parallel analysis	1.17	1.13	1.10	1.07
95th percentile eigenvalues from parallel analysis	1.21	1.16	1.12	1.09

Notes: n=648

The items were grouped under the factor, with eigenvalues greater than 1, to which they were related. To clarify the relationship among factors, varimax rotation (the orthogonal rotation technique of Principal Component Analysis) is used.

As a result of the EFA it was found that the eigenvalue of the factors from the first to the fourth were 2.17, 1.92, 1.86 and 1.67 respectively. Additionally, the variance explained by the factors from the first to the fourth were 18.12, 16.01, 15.25 and 13.95 respectively. The total variance explained by the questionnaire was 63.61%. The findings revealed that the construct validity of the questionnaire was sufficient and factor structure was similar with the original form. The factors formed after EFA and the items collected under each factor are given in Table 6.

Table 6. Factor Loadings, Item-Total Correlations and Common Variances for Adult Application

Factors	PA Factor Loadings				EFA Factor Loadings				Item-Total Correlation	Common Variances
	1	2	3	4	1	2	3	4		
	.68	-.19	.25	-.03	.83	.16	.06	.03	.39	.72
Physical	.82	-.01	-.07	-.21	.77	.03	.27	.09	.44	.67
	.45	.02	.12	.04	.69	.18	.09	.05	.37	.52
	-.02	.32	.04	.10	.22	.77	.14	-.02	.41	.66
Verbal	-.04	.72	-.02	-.03	.20	.75	.19	.06	.46	.65
	.05	.70	-.00	-.01	-.01	.58	.09	.22	.33	.40
	-.03	.07	.51	-.01	.17	.06	.79	.16	.41	.57
Anger	.00	-.10	.77	.02	.22	.20	.73	.12	.48	.67
	.02	.06	.67	-.03	.03	.22	.72	.06	.56	.64
	-.01	.02	.14	.57	.03	.12	.06	.89	.44	.81
Hostility	-.01	.03	-.10	.97	.07	.03	.10	.84	.42	.73
	.02	-.04	.02	.75	.06	.11	.17	.76	.44	.62

Notes: PA= Parallel Analysis, EFA= Exploratory Factor Analysis

As can be seen in Table 4, each factor is composed of the three items. The factor loadings of the first factor vary between .82 and .45 for PA, .83, and .69 for EFA. The factor loading of the second factor values varies between .72 and .32 for PA, .77, and .58 for EFA. The factor loadings of the third factor vary between .77 and .51 for PA, .79 and .72 for EFA. The factor loadings of the fourth factor vary between .97 and .57 for PA, .89 and .76 for EFA. Following this phase, the items in each factor were examined as a whole and a factor structure consistent with the original form of the questionnaire was observed. In order to determine whether there were significant correlations among the factors forming AQ-SF adult application, Pearson Product-Moment Correlation Analysis was performed. It was revealed that the relationship of "Physical Aggression" factor with "Verbal Aggression", "Anger", and "Hostility" was found as .38, .38, .21, respectively; the relationship of "Verbal Aggression" with "Anger" and "Hostility" was found as .38 and .24 respectively, and lastly, the relationship between "Anger" and "Hostility" was determined as .29. The results obtained, consistent with the literature (Şahin, 2018), show a positive significant relationship among all the sub-dimensions of the questionnaire $p \leq .001$.

First and second-order CFA were performed to determine whether the 12-item, 4-factor structure of the questionnaire achieved after EFA performed for AQ-SF adult application would be verified. The models obtained from these analyses are given in Figure 4. One-factor solution was also tested.

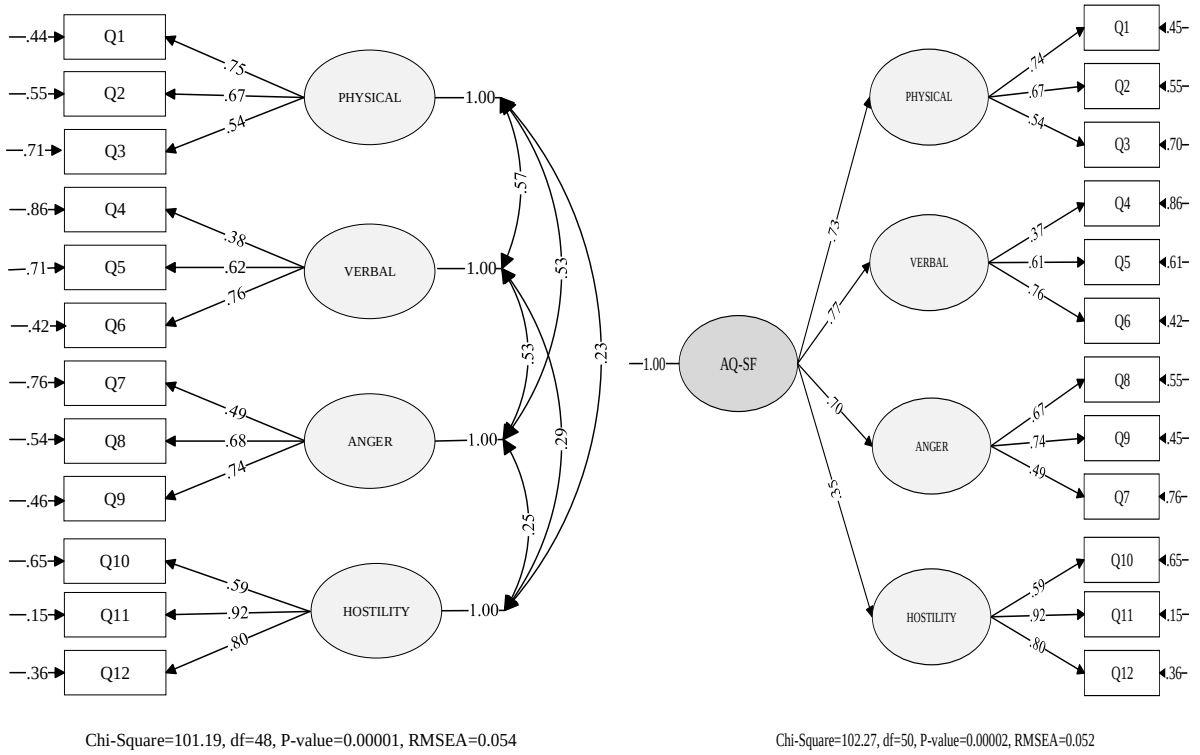


Figure 4. AQ-SF Adult 1st and 2nd Order CFA

First and second-order CFA were performed for AQ-SF adult application. When the result of CFA was evaluated, χ^2/sd ratios for the first and second-order were determined as 2.11 (χ^2/sd=101.19/48) and 2.05 (χ^2/sd=102.27/50), respectively. RMSEA fit index values were as determined as 0.054 and 0.052 as a result of first and second second-order CFA, respectively. It was determined that, among the fit index values related to the model as a result of the first and second-order CFA, AGFI was 0.93, GFI was 0.96, standardized RMR fit index value was 0.063, NFI fit index value was 0.94, and CFI fit index value was 0.97. When all the values related to data fit of the model are taken into consideration, it can be seen that the model has good fit indices.

An additional CFA was performed to support the multifactorial structure of AQ-SF adult application; the results of the first and second-order factor analyses were compared with the one-factor analysis of the questionnaire. The questionnaire was assumed unidimensional and it produced following statistics: χ^2/sd ratio of the fit values used in the model comparisons was calculated as 11.41 (χ^2/sd=616.26/54, RMSEA= 0.17, GFI= 0.79, NFI= 0.66, CFI = 0.67). Consistent with the model comparison in the adolescent group, the second-order CFA was considered to be superior since it has higher degrees of freedom, i.e., having more parsimony. The results also showed that the one-factor structure had poorer fit values than the multifactorial structure.

In order to determine the convergent validity of AQ-SF adult application, the relationship between trait anger scores and AQ-SF scores from the adult application was examined with Pearson Correlation Analysis, and it was found that there is a positive (r=.56) and statistically significant ($p\leq.001$) relationship between the two variables. Additionally, to determine the divergent validity of AQ-SF adult application, the relationship between social problem solving and AQ-SF scores from the adult application was examined, and aggression has a statistically significant relationship with social problem solving (r =-.31, $p\leq.001$).

The reliability analysis

The reliability analysis of each factor and overall of the AQ-SF adult application was also conducted. Cronbach Alpha and McDonald's Omega coefficients have the following values for first factor .70 and.68, for second factor .60 and .60, for the third factor .68 and .62, for the fourth factor, .80 and .65 respectively. Cronbach Alpha and McDonald's Omega were calculated as .78 and .72 for all questionnaire.

Test-retest reliability was found as .98. Item total correlation coefficients varied between .45 and .56 for the first factor, .29 and .48 for the second factor, .44 and .53 for the third factor, 56 and .74 for the fourth factor, .33 and .56 for the total.

Item analysis was performed to compare the responses of the individuals with low scores and high scores As a result of the *t*-test performed for this purpose, *t* values of the items varied between 8.16 (p<.001) and 2.83 (p<.001), and there was a significant difference. It can be seen that the reliability values of the overall and sub-dimensions of the AQ-SF adult application are generally acceptable values for social sciences.

Measurement Invariance for Sex and Age

For the questionnaire to show this it measures in the same manner for two subgroups MI is tested (Vandenberg & Lance, 2000). In the MI process, the aim is to test the factor structure of the questionnaire for different groups and to reach to a similar factor structure for compared groups. MI is frequently checked via multi-group confirmatory factor analysis (MG-CFA) (Jöreskog & Sörbom, 1993; Meredith, 1993). Additionally, these models are based on the increasingly restrictive assumptions regarding to the relations between the observed variables and the latent factor(s). These hierarchical models are named structural invariance, metric invariance, strong invariance, and strict invariance respectively through the least strict one to the most. For structural invariance an equal factor structure (i.e., constraining the number of factor(s) and the pattern of fixed and free loadings) across groups is required. When this requirement is met, it means respondents from various groups employ the same conceptual framework when responding (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000). Metric invariance requires invariant factor loadings across groups. This would have accepted that the content of the factors is the same across groups and that relationships between variables can justifiably be compared across groups (Iurino & Saucier, 2020; Milfont & Fischer, 2015). The third step, strong invariance necessitates equivalent intercepts (for continuous variables) or equivalent thresholds (for ordinal variables), invariant intercepts across groups and it suggests that means across groups can be compared Gustavsson, Eriksson, Hilding, Gunnarsson, & Östensson, 2008; Iurino & Saucier, 2020). In the most rigid model, a strict invariance implies equivalent residual variances and indicates that the systematic measurement error is invariant across groups (Iurino & Saucier, 2020; Meredith, 1993). Among these models, in a hierarchical way, to meet the requirements of a model means to meet the requirements of the previous model(s).

Additionally, in the decision of how well MI models fit the data, several model indexes are used. Chi-square (χ^2), the root mean square of error of approximation (RMSEA), comparative fit index (CFI), non-normed fit index (NNFI), are some of them (Emerson, Guhn, & Gadermann; 2017; Guo et al., 2017). In the acceptable MI conditions, it is expected that differences between indexes (RMSEA, CFI, NNFI) of ensuing models should be equal or smaller than -0.01, ΔRMSEA, ΔCFI, ΔNNFI $\leq$ -0.01 (Guo et al., 2017; Wu, Li, & Zumbo, 2007) and χ^2 show insignificant change from previous model Guo et al., 2017).

In this manner, the present study tested whether participants from different groups having the same aggression level will have the same scores from AQ-SF or not through MI. In other words, to determine whether the properties of the questionnaire are invariant among males and females, MI was examined in terms of sex. In addition to this sex comparison, the questionnaire was tested in different age groups. To test the MI of the factor structure of the questionnaire was being measured for the sex groups (male, female) and age (adolescents and adults), MG-CFA was used. For this purpose, four hierarchical models

were tested respectively: structural invariance, metric invariance, strong invariance, and strict invariance.

Moreover, in this study, it was examined whether the invariance conditions of ΔRMSEA, ΔCFI, ΔNNFI ≤ -0.01 for MG-CFA study files which are compatible with the data were obtained. The fact that ΔRMSEA, ΔCFI, and ΔNNFI values obtained as a result of the comparison of the two models are equal to -.01 or below can be used as the evidence that the MI is achieved (Wu, Li, & Zumbo, 2007).

The findings regarding the invariance steps tested are present in Table 7. "The Structural Invariance Model" in the table represents the factor loads, regression constant, and the error variances free model; "The Weak Invariance Model" in the table represents the factor loads constant, regression constants, and error variances free model; "The Strong Invariance Model" in the table represents the factor loads, regression constants, and error variance free model; and "The Strict Invariance Model" in the table represents the factor loads, regression constants, and error variances constant model.

Table 7. Fit Statistics Regarding MI

Steps	$\chi 2$	Df	RMSEA (CI)	ΔRMSEA	CFI	ΔCFI	NNFI	ΔNNFI
Sex								
SI	162.70	108	0.036 (0.02; 0.04)		0.98		0.98	
MI	232.91	120	0.049 (0.04; 0.05)	-0.013	0.97	0.01	0.97	0.01
SgI	246.35	126	0.049 (0.04; 0.05)	0.000	0.97	0.00	0.97	0.00
StI	246.61	126	0.050 (0.04; 0.05)	0.001	0.97	0.00	0.97	0.00
Age								
SI	282.19	108	0.064 (0.05; 0.07)		0.96		0.95	
MI	333.32	120	0.067 (0.05; 0.07)	0.003	0.95	0.01	0.95	0.00
SgI	338.54	126	0.065 (0.05; 0.07)	0.002	0.95	0.00	0.95	0.00
StI	380.06	126	0.071 (0.06; 0.08)	-0.006	0.94	-0.01	0.94	0.01

Notes: n= 782 (for sex), 792 (for age) CI= Confidence Interval, SI= Structural Invariance, MI= Metric Invariance, SgI= Strong Invariance, StI= Strict Invariance

As can be seen in Table 7, the fit indexes obtained as a result of multi-group RMSEA, CFI, NNFI and ΔRMSEA, ΔCFI, ΔNNFI values obtained as a result of the CFI difference test can be interpreted for each step as follows. According to the results, it is seen that the structural invariance is provided, and this finding shows that the measured structures use the same conceptual perspectives in responding to the questionnaire items of the adolescents and adults; males and females. The result regarding the metric invariance indicates that the factor structures of the variables taken in the model are the same in the adolescent and adult; male and female groups. It is confirmed that strong invariance is provided, and the constant number in the regression equations formed for the items is invariant between the groups. In the last stage, considering the ΔRMSEA, ΔCFI, ΔNNFI values calculated with the fit indexes, it is accepted that the error terms regarding the items forming the measurement tool are invariant between the comparison groups. Hierarchical analysis results, factor structure, and pattern of the questionnaire, factor loads, regression constants, and error variances are seen to be invariant for the adolescent and adult; male and female groups.

Differential Item Functioning for Sex and Age

In order to provide evidence for the validity of the items included in the measurement tools used in the study, it was examined whether each item showed bias according to the sex and age variables. In this context, it has been examined how the responses given to the items according to sex and age variables with the help of logistic functions by using the Mantel-Haenszel technique, which is based on the Item Response Theory. The change in the likelihood that individuals with the same level of ability will

respond correctly to an item is based on two reasons item bias or differences of actual knowledge, skill, etc. Determining whether items give DIF is a more commonly used technique, as it is seen as a more objective approach to bias (Doğan & Öğretmen, 2008).

DIF results for sex

As a result of the determination of males as focus groups and females as reference groups; the comparison variable is accepted as the score obtained from the questionnaire's each item. The chi-square values, significance values, and statistics showing the level of DIF obtained as a result of the analysis are presented in Table 8.

Table 8. DIF Results for Sex

Item	χ^2	Error	CI Lower	CI Upper	Class
Q1	4.03*	0.16	0.01	0.31	AA
Q2	74.63***	0.66	0.50	0.81	CC+
Q3	6.22**	0.20	0.04	0.36	BB+
Q4	0.18	-0.03	-0.17	0.11	AA
Q5	5.16**	0.20	0.03	0.36	AA
Q6	12.73***	0.27	0.11	0.42	BB+
Q7	6.56**	-0.28	-0.48	-0.09	BB-
Q8	12.03***	-0.28	-0.45	-0.11	BB-
Q9	2.54	-0.12	-0.28	0.04	AA
Q10	4.61*	-0.20	-0.39	-0.02	BB-
Q11	2.14	-0.14	-0.31	0.04	AA
Q12	2.49	-0.15	-0.33	0.02	AA

Notes: n= 1825, *= $p < .05$, **= $p < .01$, ***= $p < .001$, CI= Confidence Interval

When Table 8 is analyzed, it is seen that the χ^2 values obtained for all the items except Q2 coded item among the items in the measurement tool are not statistically significant in the determined degree of freedom. In other words, in the AQ-SF it was found that six items showed negligible (AA) DIF, six items showed medium (BB) DIF and one item showed high (CC) DIF (Güzeller, Eser & Aksu, 2018). This result explains that the 12 items in the measurement tool do not work in favor of female or male participants and the results obtained from the measurement tool didn't differ for both groups. However, it was determined that the Q2 coded item in the measurement tool showed DIF in favor of the focus group at the CC + (high) level. In order to say that an item produces biased results for or against one of the subgroups in the study universe, it should show at least C (high) DIF (Koyuncu, Aksu, & Kelecioğlu, 2018). Therefore, it is necessary to examine whether the item is biased according to the sex variable. The characteristic curve obtained for the second item determined to show a high level of DIF is shown in Figure 5.

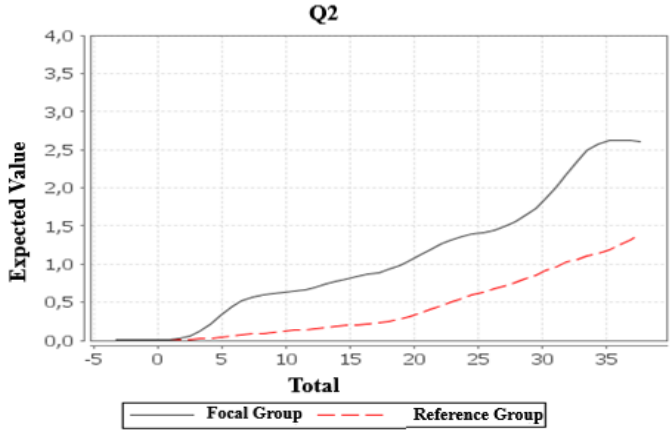


Figure 5. Item Characteristic Curve for the Related Item

When Figure 5 is examined, it is determined that the scores obtained from this item show DIF in favor of male participants who are determined as the focus group at all ability levels. In other words, Q2 measures aggression differently for males from aggression for females. Item impact means that respondents in different groups answer one item correctly express the real differences in their probabilities. This difference is explained by the knowledge or experience that one of the groups has (Gök, Kelecioğlu, Doğan & 2010). Item impact is also evident when examinees from different groups have differing probabilities of responding correctly to (or endorsing) an item because there are true differences between the groups in the underlying ability being measured by the item (Zumbo, 1999).

DIF results for age

It was analyzed whether each item differs depending on the age variable. As a result of the determination of the fewer adolescents as focus groups and adults as reference groups, the comparison variable is accepted as the score obtained for each questionnaire item. The chi-square values, significance values, and statistics showing the level of DIF obtained as a result of the analysis are presented in Table 9.

Table 9. DIF Results for Age

Item	χ^2	Error	CI Lower	CI Upper	Class
Q1	7.13**	0.11	0.02	0.19	AA
Q2	1.76	0.05	-0.03	0.13	AA
Q3	30.85***	0.23	0.15	0.32	BB+
Q4	0.99	-0.04	-0.11	0.04	AA
Q5	0.22	0.02	-0.07	0.10	AA
Q6	14.53***	0.16	0.07	0.24	AA
Q7	3.47	-0.10	-0.20	0.00	AA
Q8	3.45	0.08	-0.01	0.17	AA
Q9	3.88*	-0.09	-0.16	-0.01	AA
Q10	1.63*	-0.06	-0.16	0.04	AA
Q11	29.00***	-0.24	-0.33	-0.15	BB-
Q12	6.35**	-0.12	-0.21	-0.02	AA

Notes: n= 1825, *= $p < .05$, **= $p < .01$, ***= $p < .001$, CI= Confidence Interval

Table 9 indicated that the χ^2 values obtained for all the items in the measurement tool are not statistically significant in the determined degree of freedom. In the AQ-SF, it was found that 10 items showed negligible (AA) DIF and two items showed medium (BB) DIF (Güzeller, Eser & Aksu, 2018). This result explains that the 12 items in the measurement tool do not work in favor of female or male students and the results obtained from the measurement tool didn't differ for both groups.

Accordingly, when the results obtained regarding the reliability and validity of the measurement tool were analyzed as a whole, it was determined that the aggressive characteristics of the adolescents and adults were measured with a valid and reliable measurement tool.

DISCUSSION and CONCLUSION

This study aims to make the adaptation of the Aggression Questionnaire Short Form-in Turkish with adolescent and adult samples. In order to test the construct validity of the questionnaire, PA was conducted. The four factor structure of the questionnaire was confirmed via PA, which was defined as the best way to determine factor numbers to retain (Ledesma & Valero-Mora, 2007). This analysis has been indicated consistently accurate in determining the threshold for significant components, variable loadings, and analytical statistics when decomposing a correlation matrix (Franklin, Gibson, Robertson, Pohlmann, & Fralish, 1995). Moreover, the factor structure of the questionnaire was tested through EFA. EFA findings indicated that the questionnaire has a four-factor structure of adolescent and adult samples similar to the original form of the questionnaire (Bryant & Smith, 2001). Additionally, the results of the CFA, which were conducted for both adolescents and adults confirmed the four-factor structure of the questionnaire. These results also parallel the findings of Braynt & Smith (2001) that about the CFA for

the original form of the questionnaire. The four factors structure of the questionnaire was also approved via CFA in the study, which includes Spanish (Morales-Vives, Codorniu-Raga, & Andreu Vigil-Colet, 2005), Egyptian, Omani (Abd-El-Fattah, 2013), Dutch (Hornsveld et al. 2009) adolescents. In studies conducted with adults by Maxwell (2007) and Vitoratou et al. (2009), CFA results indicated four factor structure. McKay, Perry, and Harway (2016) tested both unidimensional and four-factor models of AQ-SF and reported limited evidence for unidimensional models beside four-factor model supported results. Different from the studies which support four-factor structure of AQ-SF via CFA, Kožený, Tišanská, & Csémy (2017) reported one component, Reyna et al. (2011) indicated two-component structure for AQ-SF.

For validity analysis, convergent and divergent validity of AQ-SF was examined. The moderate and significant correlation of AQ-SF scores with trait anger and social problem-solving scores in adolescent and adult applications confirmed the construct validity of AQ-SF. A significant and moderate correlation between AQ scores and trait anger level was reported by Wang et al. (2018). Similarly, Kuzucu (2016) reported a significant correlation between AQ-SF scores and social problem-solving scores. These results are not only evidence for convergent and divergent validity of the AQ-SF, but also show the correlation of questionnaire both with ill-being and well-being variables.

In terms of reliability, internal consistency and test-retest reliability scores were calculated. While the Cronbach Alpha scores in the present study are acceptable similar to the original form (Bryant & Smith, 2001), the test-retest reliability scores are higher than the original form of the questionnaire (Buss and Perry, 1992) and most of the previous studies (Harris, 1997; Surís, Borman, Lind, & Kashner, 2007; Webster et al. 2014). The differences were found between the responses of the individuals with low and high scores in adolescent and adult groups.

To test invariant measurement models of the AQ-SF between different sex and age groups, MI of the questionnaire was also tested in terms of sex and age. In the present study, there is sex invariance for measurement through AQ-SF between males and females. It is consistent with the other findings in the literature. Sex differences about the type and magnitude of aggressive behaviors seem as common results of the studies (Björkqvist, Österman, & Lagerspetz, 1994; Eron, Huesmann, Dubow, Romanoff, & Yarmel, 1987). The invariance of sex was also mentioned by Bryant & Smith (2001). Moreover, among Greek adults (Vitoratou et al. 2009) and federal offenders (Diamond & Magaletta, 2006), sex invariance was reported. Different from the sex invariance results of the present study, partial sex MI of AQ-SF for Argentinean (Reyna et al. 2011), Egyptian (Abd-El-Fattah, 2013) adolescents, and adolescents from Singapour (Ang, 2007) and Liverpool (McKay et al. 2016) was reported. The previous studies tested and showed MI of the questionnaire also with several samples from similar demographic backgrounds (Ang, 2007; Bryant & Smith, 2001; Vitoratou et al., 2009).

There is an age invariance for measurement through AQ-SF between adolescents and adults. In literature, adolescents are reported no more aggressive than adults. Adults are not less hostile than adolescents, but they use different and more latent means of aggression (Björkqvist et al., 1994). Torregrosa et al. (2020) showed age invariance between 8-9 and 10-11 aged children. Moreover, longitudinal studies emphasized the continuity of aggressive behaviors through adolescence to adulthood (Eron et al. 1987; Huesmann, Eron, Lefkowitz, & Walder, 1984, Huesmann, Eron, & Dubow, 2002). The present findings confirmed the invariant measurement of aggression between adolescents and adults via AQ-SF. However, to our knowledge, there is no study in which age invariance was tested for AQ-SF among adolescents and adults.

The DIF analysis for sex showed that the item of AQ-SF coded as Q2 'There are people who pushed me so far that we came to blows' measure aggression in a biased way between the sexes. With the aim of explaining whether this difference is item bias or true difference, expert opinion was obtained. The expert group interview conducted with the consideration of it is a physical aggression related item and they focused that it measures physical aggression in favor of males. In conclusion, this difference should be accepted as the real difference due to biological reasons; as a result, males are more likely to respond to this item. Similar to the DIF results and experts' opinions about Q2, it was reported that males are more physically aggressive than females related to the testosterone level (Björkqvist, 2018). Despite the

focused age group, Lansford et al., (2012) reported more physical aggression among boys than girls, consistently across nine different countries.

The DIF analysis for age supported that there is no bias in the AQ-SF items for adolescents and adults. In addition to the power of the questionnaire in terms of factorial structural that MI for age results showed, DIF results reinforced this power by items for different age groups. All items of the questionnaire measure aggression in an unbiased way for age. This result has support in the literature. With the evidence from longitudinal studies (Eron et al. 1987; Huesmann et al., 1984; Huesmann, Eron, & Dubow, 2002) it is known that aggression has persisted from adolescence to adulthood. Moreover, aggression is a topic that is investigated in the life span approach. Several studies were conducted with different aged group participants, from toddlerhood to old-adulthood (Liu, Lewis, & Evans, 2013). This wide range of studies of aggression, both in terms of time and age could explain the power of AS-QF about giving reliable measurements for different ages.

Despite the contributions to literature, this study has limitations. The results for the AQ-SF were not compared with BPAQ (29 item version). In the current study, participants came from a nonclinical sample. In further studies, MI for clinical and nonclinical samples can be tested. In addition to the cross-sectional data set, testing sex and age invariance in aggression with longitudinal data is another suggestion for the researchers. All results for validity and reliability tests confirmed four factors and 12 items structure of the questionnaire. The findings also presented that the AQ-SF is a valid and reliable questionnaire, and it can be used for male, female, adolescent, and adult populations.

REFERENCES

Abd-El-Fattah, S. M. (2013). A cross-cultural examination of the Aggression Questionnaire–Short Form among Egyptian and Omani Adolescents. *Journal of Personality Assessment, 95*(5), 539-548.

Adıgüzel, V., Özdemir, N., & Şahin, Ş. K. (2019). Childhood traumas in euthymic bipolar disorder patients in Eastern Turkey and its relations with suicide risk and aggression. *Nordic Journal of Psychiatry, 73*(8), 490-496.

Akbaş, U., Karabay, E., Yıldırım-Seheryeli, M., Ayaz, A., ve Demir, Ö. O. (2019). Türkiye ölçme araçları dizininde yer alan açımlayıcı faktör analizi çalışmalarının paralel analiz sonuçları ile karşılaştırılması. *Kuramsal Eğitimbilim Dergisi, 12*(3), 1095-1123.

Aksu, G., Eser, M. T., & Güzeller, C. O. (2017). *Açımlayıcı ve doğrulayıcı faktör analizi ile yapısal eşitlik modeli uygulamaları.* Detay Yayıncılık.

Ang, R. P. (2007). Factor structure of the 12-item aggression questionnaire: Further evidence from Asian adolescent samples. *Journal of Adolescence, 30*(4), 671-685.

Bernstein, I. H., & Gesn, P. R. (1997). On the dimensionality of the Buss/Perry Aggression Questionnaire. *Behaviour Research and Therapy, 35*(6), 563-568.

Björkqvist, K., Österman, K., & Lagerspetz, K. M. (1994). Sex differences in covert aggression among adults. *Aggressive Behavior, 20*(1), 27-33.

Björkqvist, K. (2018). Gender differences in aggression. *Current Opinion in Psychology, 19*, 39-42.

Borders, A., Earleywine, M., & Jajodia, A. (2010). Could mindfulness decrease anger, hostility, and aggression by decreasing rumination? *Aggressive Behavior: Official Journal of the International Society for Research on Aggression, 36*(1), 28-44.

Bryant, F. B., & Smith, B. D. (2001). Refining the architecture of aggression: A measurement model for the Buss–Perry Aggression Questionnaire. *Journal of Research in Personality, 35*(2), 138-167.

Buss, A. H., & Durkee, A. (1957). An inventory for assessing different kinds of hostility. *Journal of Consulting Psychology, 21*(4), 343.

Buss, A. H., & Perry, M. (1992). The Aggression Questionnaire. *Journal of Personality and Social Psychology, 63*(3), 452.

Büyüköztürk, Ş. (2007). *Sosyal bilimler için veri analizi el kitabı*. Ankara: Pegem A Yayıncılık.

Carver, C. S., Sinclair, S., & Johnson, S. L. (2010). Authentic and hubristic pride: Differential relations to aspects of goal regulation, affect, and self-control. *Journal of Research in Personality, 44*(6), 698-703.

Chang, L., Schwartz, D., Dodge, K. A., & McBride-Chang, C. (2003). Harsh parenting in relation to child emotion regulation and aggression. *Journal of Family Psychology, 17*(4), 598.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodnessof-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255.

Coie, J. D., & Dodge, K. A. (1998). Aggression and antisocial behavior. In W. Damon & N. Eisenberg (Ed.), *Handbook of child psychology: Social, Emotional, and Personality Development* (pp. 779-862). Hoboken, NJ, US: John Wiley & Sons Inc.

D'Zurilla, T. J., Nezu, A. M. & Maydeu-Olivares, A. (2002). *Manual for the Social Problem Solving Inventory-revised (SPSI-R).* North Tonawanda: Multi-Health Systems.

De Zavala, A. G., Cichocka, A., Eidelson, R., & Jayawickreme, N. (2009). Collective narcissism and its social consequences. *Journal of Personality and Social Psychology, 97*(6), 1074.

Diamond, P. M., & Magaletta, P. R. (2006). The short-form Buss-Perry Aggression Questionnaire (BPAQ-SF) a validation study with federal offenders. *Assessment, 13*(3), 227-240.

Diamond, P. M., Wang, E. W., & Buffington-Vollum, J. (2005). Factor structure of the Buss-Perry Aggression Questionnaire (BPAQ) with mentally ill male prisoners. *Criminal Justice and Behavior, 32*(5), 546-564.

Doğan, N., & Öğretmen, T. (2008). Degisen madde fonksiyonunu belirlemede mantel - Haenszel, ki-kare ve lojistik regresyon tekniklerinin karsilastirilmasi. *Egitim ve Bilim, 33*(148), 100–112.

Emerson, S. D., Guhn, M., & Gadermann, A. M. (2017). Measurement invariance of the Satisfaction with Life Scale: reviewing three decades of research. *Quality of Life Research, 26*(9), 2251-2264.

Erkuş, A. (2012). *Psikolojide ölçme ve ölçek geliştirme.* Ankara: Pegem Akademi Yayınları.

Eron, L. D., Huesmann, L. R., Dubow, E., Romanoff, R., & Yarmel, P. W. (1987). Aggression and its correlates over 22 years. *Childhood Aggression and Violence*, 249–262.

Eskin, M., & Aycan, Z. (2009). The adaptation of the revised social problem solving inventory into Turkish: A reliability and validity analysis. *Turkish Journal of Psychology, 12*(23), 11–13.

Evren, C., Çınar, Ö., Güleç, H., Çelik, S., & Evren, B. (2011). The validity and reliability of the Turkish version of the Buss-Perry's Aggression Questionnaire in male substance dependent inpatients. *The Journal of Psychiatry and Neurological Sciences, 24*(4), 283-295.

Ferrando, P. J., & Lorenzo-Seva, U. (2017). 10 años del programa FACTOR: Una revisión crítica de sus orígenes, desarrollo y líneas futuras. *Psicothema, 29*(2), 236–240.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics.* Sage.

Fraenkel, J.R., Wallen, N.E., & Hyun, H.H. (2012). How to design and evaluate research in education (Eight Edition). New York: McGraw-Hill

Franklin, S. B., Gibson, D. J., Robertson, P. A., Pohlmann, J. T., & Fralish, J. S. (1995). Parallel Analysis: a method for determining significant principal components. *Journal of Vegetation Science, 6*(1), 99–106.

Gök, B., Kelecioglu, H., & Dogan, N. (2010). The Comparison of Mantel-Haenszel and Logistic Regression Techniques in Determining the Differential Item Functioning. *Egitim ve Bilim-Education and Science, 35*(156), 3–16.

García-León, A., Reyes, G. A., Vila, J., Pérez, N., Robles, H., & Ramos, M. M. (2002). The Aggression Questionnaire: A validation study in student samples. *The Spanish Journal of Psychology, 5*(1), 45-53.

Gerevich, J., Bácskai, E., & Czobor, P. (2007). The generalizability of the Buss–Perry Aggression Questionnaire. *International Journal of Methods in Psychiatric Research, 16*(3), 124-136.

Guo, B., Kaylor-Hughes, C., Garland, A., Nixon, N., Sweeney, T., Simpson, S., ... & Morriss, R. (2017). Factor structure and longitudinal measurement invariance of PHQ-9 for specialist mental health care patients with persistent major depressive disorder: Exploratory Structural Equation Modelling. *Journal of Affective Disorders, 219*, 1-8.

Gustavsson, J. P., Eriksson, A. K., Hilding, A., Gunnarsson, M., & Östensson, C. G. (2008). Measurement invariance of personality traits from a five-factor model perspective: multi-group confirmatory factor analyses of the HP5 inventory. *Scandinavian Journal of Psychology, 49*(5), 459–467.

Harris, J. A. (1995). Confirmatory factor analysis of the Aggression Questionnaire. *Behaviour Research and Therapy, 33*(8), 991-993.

Harris, J. A. (1997). A further evaluation of the aggression questionnaire: Issues of validity and reliability. *Behaviour Research and Therapy, 35*(11), 1047-1053.

Hinshaw, S. P. (1987). On the distinction between attentional deficits/hyperactivity and conduct problems/aggression in child psychopathology. *Psychological Bulletin, 101*(3), 443.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185.

Hornsveld, R. H., Muris, P., Kraaimaat, F. W., & Meesters, C. (2009). Psychometric properties of the aggression questionnaire in Dutch violent forensic psychiatric patients and secondary vocational students. *Assessment, 16*(2), 181-192.

Huesmann, L. R., Eron, L. D., & Dubow, E. F. (2002). Childhood predictors of adult criminality: are all risk factors reflected in childhood aggressiveness?. *Criminal Behaviour and Mental Health, 12*(3), 185-208.

Huesmann, L. R., Eron, L. D., Lefkowitz, M. M., & Walder, L. O. (1984). Stability of aggression over time and generations. *Developmental Psychology, 20*(6), 1120–1134.

Iurino, K., & Saucier, G. (2020). Testing measurement invariance of the Moral Foundations Questionnaire across 27 countries. *Assessment, 27*(2), 365-372.

Johnson, S. L., Carver, C. S., & Joormann, J. (2013). Impulsive responses to emotion as a trans diagnostic vulnerability to internalizing and externalizing symptoms. *Journal of Affective Disorders, 150*(3), 872-878.

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

Kang, J. H., Lim, C. H., Suh, Y. I., Gang, A. C., & Pedersen, P. M. (2020). Establishing a web-based measurement of aggression (WTCRTT): Examining the validity of a modified Taylor's competitive reaction time test. *International Journal of Applied Sports Sciences, 32*(1), 37-48.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford Publications.

Koyuncu, İ., Aksu, G., & Kelecioğlu, H. (2018). Mantel-Haenszel, Lojistik Regresyon ve Olabilirlik Oranı Değişen Madde Fonksiyonu İnceleme Yöntemlerinin Farklı Yazılımlar Kullanılarak Karşılaştırılması. *Elementary Education Online, 17*(2), 909–925.

Kožený, J., Tišanská, L., & Csémy, L. (2017). A Rasch Analysis of The Buss-Perry Aggression Questionnaire-Short Form: An evidence from Czech adolescents' sample. *Ceskoslovenska Psychologie, 61*(3), 257-266.

Kuzucu, Y. (2016). Do anger control and social problem-solving mediate relationships between difficulties in emotion regulation and aggression in adolescents? *Educational Sciences: Theory & Practice, 16*(3).

Kühn, S., Kugler, D. T., Schmalen, K., Weichenberger, M., Witt, C., & Gallinat, J. (2019). Does playing violent video games cause aggression? A longitudinal intervention study. *Molecular Psychiatry, 24*(8), 1220-1234.

Lansford, J. E., Skinner, A. T., Sorbring, E., Giunta, L. D., Deater-Deckard, K., Dodge, K. A., ... & Uribe Tirado, L. M. (2012). Boys' and girls' relational and physical aggression in nine countries. *Aggressive Behavior, 38*(4), 298-308.

Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research & Evaluation, 12*(2), 1-11.

Liu, J., Lewis, G., & Evans, L. (2013). Understanding aggressive behaviour across the lifespan. *Journal of psychiatric and mental health nursing, 20*(2), 156-168.

Madran, H. A. D. (2012). Buss-Perry Saldırganlık Ölçeği'nin Türkçe formunun geçerlik ve güvenirlik çalışması. *Türk Psikoloji Dergisi, 24(2),* 1-6.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review, 50*(4), 370.

Maxwell, J. P. (2007). Development and preliminary validation of a Chinese version of the Buss–Perry Aggression Questionnaire in a population of Hong Kong Chinese. *Journal of Personality Assessment, 88*(3), 284-294.

McKay, M. T., Perry, J. L., & Harvey, S. A. (2016). The factorial validity and reliability of three versions of the Aggression Questionnaire using Confirmatory Factor Analysis and Exploratory Structural Equation Modelling. *Personality and Individual Differences, 90*, 12-15.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543.

Meyer, J. P. (2018). *JMETRİK ile Ölçme Uygulamaları.* (Çev. C. O. Güzeller, M. T. Eser, G. Aksu). Maya Akademi (Orijinal yayın tarihi 2014).

Milfont, T. L., & Fischer, R. (2015). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3*(1), 111–130.

Morales-Vives, F., Codorniu-Raga, M. J., & Vigil-Colet, A. (2005). Características psicométricas de las versiones reducidas del cuestionario de agresividad de Buss y Perry. *Psicothema, 17*(1), 96-100.

Moyer, K. E. (1982). *The origins of aggression in child nurturance* (pp. 243-260). Springer, Boston, MA.

Orpinas, P., & Frankowski, R. (2001). The Aggression Scale: A self-report measure of aggressive behavior for young adolescents. *The Journal of Early Adolescence, 21*(1), 50-67.

Önen, E. (2016). Saldırganlık Ölçeği'nin Psikometrik Niteliklerinin Türk Ergenleri İçin İncelenmesi. *Türk Psikolojik Danışma ve Rehberlik Dergisi, 4*(32).

Özer, A. K. (1994). Sürekli öfke ve öfke ifade tarzı ölçekleri ön çalışması. *Türk Psikoloji Dergisi, 9(31),* 26-35.

Özdemir, Y., Vazsonyi, A. T., & Cok, F. (2017). Parenting processes, self-esteem, and aggression: A mediation model. *European Journal of Developmental Psychology, 14*(5), 509-532.

Palmstierna, T., & Wistedt, B. (1987). Staff observation aggression scale, SOAS: presentation and evaluation. *Acta Psychiatrica Scandinavica, 76*(6), 657-663.

Podubinski, T., Lee, S., Hollander, Y., & Daffern, M. (2017). Patient characteristics associated with aggression in mental health units. *Psychiatry Research, 250*, 141-145.

Raine, A., Dodge, K., Loeber, R., Gatzke-Kopp, L., Lynam, D., Reynolds, C., ... & Liu, J. (2006). The reactive–proactive aggression questionnaire: Differential correlates of reactive and proactive aggression in

adolescent boys. Aggressive Behavior: *Official Journal of the International Society for Research on Aggression*, *32*(2), 159-171.

Reyna, C., Sanchez, A., Ivacevich, M. G. L., & Brussino, S. (2011). The Buss-Perry Aggression Questionnaire: construct validity and gender invariance among Argentinean adolescents. *International Journal of Psychological Research*, *4*(2), 30-37.

Revelle, W. (2018). Psych: procedures for personality and psychological research. Northwestern University, Evanston.

Sexton, M. B., Davis, A. K., Buchholz, K. R., Winters, J. J., Rauch, S. A. M., Yzquibell, M., . . . Chermack, S. T. (2019). Veterans with recent substance use and aggression: PTSD, substance use, and social network behaviors. *Psychological Trauma: Theory, Research, Practice, and Policy*, *11*(4), 424-433.

Singh, S. (2017). Understanding aggression among youth in the context of mindfulness. *Indian Journal of Health and Wellbeing*, *8*(11), 1377-1379.

Sorsdahl, K., Stein, D. J., & Myers, B. (2017). Psychometric properties of the Social Problem Solving Inventory-Revised Short-Form in a South African population. *International Journal of Psychology*, *52*(2), 154-162.

Spielberger, C. D. (1985). The experience and expression of anger: Construction and validation of an anger expression scale. *Anger and hostility in cardiovascular and behavioral disorders*, 5-30.

Spielberger, C. D. (1988). *Manual for the State-Trait Anger Expression Scale (STAX)*. 1988. Odessa, FL: Psychological Assessment Resources.

Spielberger, C. D, Johnson E.H, Russell SF, Crane RJ, Jacobs GA, & Worden TJ. (1985). The experience and expression of anger: construction and validation of an anger expression scale. In: Chesney MA, Rosenman R.H, eds. Anger and Hostility in Cardiovascular and Behavioral Disorders. New York, NY: Hemisphere, 5-30.

Şahin, C. (2018). *Bireyi tanıma teknikleri: Psikolojik testler, test dışı teknikler.* Ankara: Pegem Akademi.

Surís, A., Borman, P. D., Lind, L., & Kashner, T. M. (2007). Aggression, impulsivity, and health functioning in a veteran population: equivalency and test–retest reliability of computerized and paper-and-pencil administrations. *Computers in Human Behavior*, *2*3(1), 97-110.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistic* (6th ed.). USA: Pearson.

Torregrosa, M. S., Gómez-Núñez, M. I., Inglés, C. J., Ruiz-Esteban, C., Sanmartín, R., & García-Fernández, J. M. (2020). Buss and Perry Aggression Questionnaire-Short Form in Spanish Children. *Journal of Psychopathology and Behavioral Assessment*, 1-16.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70.

Vitoratou, S., Ntzoufras, I., Smyrnis, N., & Stefanis, N. C. (2009). Factorial composition of the Aggression Questionnaire: A multi-sample study in Greek adults. *Psychiatry Research*, *168*(1), 32–39.

Wang, X., Yang, L., Yang, J., Gao, L., Zhao, F., Xie, X., & Lei, L. (2018). Trait anger and aggression: A moderated mediation model of anger rumination and moral disengagement. *Personality and Individual Differences*, *125*, 44-49.

Webster, G. D., DeWall, C. N., Pond Jr, R. S., Deckman, T., Jonason, P. K., Le, B. M., ... & Smith, C. V. (2014). The brief aggression questionnaire: Psychometric and behavioral evidence for an efficient measure of trait aggression. *Aggressive Behavior*, *40*(2), 120-139.

Webster, S. D., Mann, R. E., Thornton, D., & Wakeling, H. C. (2007). Further validation of the Short Self-esteem Scale with sexual offenders. *Legal and Criminological Psychology*,*12*(2), 207-216.

Wu, A. D., Li, Z. & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation*, *12(1)*, 1-26.

Zumbo, B. (1999). A handbook on the theory and methods of differential item functioning (DIF). *Ottawa: National Defense Headquarters*, *June*, 1–57.

Saldırganlık Ölçeği Kısa Formu Türkçe Versiyonunun Psikometrik Özelliklerinin İncelenmesi: Cinsiyet ve Yaş için Ölçme Eşdeğerliği ve Değişen Madde Fonksiyonu

Giriş

Saldırganlık; biyolojik, psikolojik, sosyal ve kültürel faktörlerin bir arada etkili olduğu çok boyutlu bir yapıda gelişmektedir (Vitoratou, Ntzoufras, Smyrnis, ve Stefanis, 2009). Birçok kuramcı ve uygulamacı saldırganlık davranışını ve ilişkili olduğu diğer davranışları açıklamaya çalışmaktadır (Chang, Schwartz, Dodge, ve McBride-Chang, 2003; Sexton ve diğerleri, 2019).

Saldırganlığı ölçmek için çeşitli ölçme araçları geliştirilmiş olup (Buss ve Perry, 1992; Kang, Lim, Suh, Gang, ve Pedersen, 2020; Orpinas ve Frankowski, 2001; Palmstierna ve Wistedt, 1987; Raine ve diğerleri, 2006), bunlar arasında en sık kullanılan ölçme aracı Buss ve Perry (1992) tarafından geliştirilen Buss-Perry Saldırganlık Ölçeğidir. Ölçek fiziksel saldırganlık, sözel saldırganlık, öfke ve düşmanlık alt boyutlarından oluşan 29 maddelik bir ölçektir. Ölçeğin psikometrik özellikleri farklı yöntem ve örneklemlerle test edilmiş ve sonuçlar dört faktörlü yapıyı doğrulamıştır (Bernstein ve Gesn, 1997; García-León ve diğerleri, 2002; Gerevich, Bácskai, ve Czobor 2007; Harris, 1997; Reyna et al. 2011). Bununla birlikte çalışmaların birçoğunda bazı maddeler ölçekten çıkarıldığında daha iyi uyum değerleri ve faktör yükleri elde edilmiştir (Bernstein ve Gesn, 1997; Gerevich ve diğerleri, 2007; Harris, 1995;).

Bryant ve Smith (2001) daha rafine bir ölçme aracı yaratmak için Buss ve Perry Saldırganlık ölçeğinin en iyi çalışan 12 maddesini belirlemiştir. Bu yeni kısa form dört faktörün her birinde üç madde olacak şekilde kısaltılmıştır. Kısa form hem Arjantin, Hollanda ve Çin gibi farklı ülkelerde (Reyna et al, 2011; Maxvell (2007) hem de tutuklular ve ruh sağlığı bozukluğu olanlar gibi farklı örneklemlerde (Buffington-Vollum, 2005; Diamond ve Magaletta, 2006; Diamond, Wang) test edilmiştir, sonuçlar ölçeğin iç tutarlığının ve uyum değerlerinin yüksek olduğunu göstermektedir.

Saldırganlık Ölçeği'nin kısa formunun yeterli psikometrik özelliklere sahip olmasının yanı sıra, ölçek diğer ruh sağlığı değişkenleri ile de yüksek ilişki göstermektedir. Yapılan çalışmalar saldırganlığın kolektif narsizim (De Zavala, Cichocka, Eidelson, ve Jayawickreme, 2009), ruminasyon (Borders, Earleywine, ve Jajodia, 2010), kibirlilik (Carver, Sinclair, ve Johnson; 2010), anksiyete ve alkol kullanımıyla (Johnson, Carver, ve Joormann, 2013) ilişkili olduğunu ortaya koymuştur.

Türkiye'de de Buss-Perry Saldırganlık Ölçeği'nin 29 maddelik formu lise, üniversite öğrencileri ve madde bağımlıları ile yapılan araştırmalarda kullanılmıştır. Elde edilen araştırma bulguları ölçeğin geçerli ve güvenilir bir araç olduğunu göstermektedir (Evren, Çınar, Güleç, Çelik, ve Evren, 2011; Madran, 2012; Önen, 2016). Bununla birlikte ölçeğin kısa formunun psikometrik özellikleri Türkiye'de çalışılmamıştır

Saldırganlık ölçeğinin kısa formu farklı kültürlerde, cinsiyette, klinik ve klinik olmayan gruplarda kullanılabilmektedir. Ölçek hem psikolojik iyi oluş hem de psikolojik sorunlarla korelasyon göstermektedir. Ölçeğin bu özellikleri dikkate alındığında kısa formunun dilimize kazandırılması önem taşımaktadır. Bu çalışmanın amacı, Saldırganlık Ölçeği Kısa Formu'nun Türkçedeki psikometrik özelliklerini ergen ve yetişkinler ile kadınlar ve erkekler için test etmektir.

Yöntem

Çalışma 15-18 yaşları arasındaki 778 ergen ve 19-44 yaşları arasındaki 1067 yetişkin katılımcı ile gerçekleştirilmiştir. Ergen çalışması için ölçeğin kısa formu beş farklı lisede öğrenim gören toplamda 778 öğrenciye uygulanmıştır. İlk uygulamada 383 öğrenciyle çalışılmış, elde edilen veri üzerinde Paralel Analiz (PA) ve Açımlayıcı Faktör Analizi (AFA) yapılmıştır. İkinci uygulamada 395 öğrenciye ulaşılmış ve bu veriler üzerinde Doğrulayıcı Faktör Analizi (DFA) yapılmıştır. Yetişkin uygulamaları

için toplamda 1067 yetişkinle çalışılmıştır. Katılımcılar Aydın Adnan Menderes Üniversitesi, Ege Üniversitesi ve Ankara Üniversitesinde pedagojik formasyon eğitimi alan kişiler ile halk eğitim merkezlerinde kurslara katılan kursiyerlerden oluşmaktadır. Veriler yetişkinlerden iki farklı uygulamayla toplanmıştır. İlk uygulamada 648 katılımcıya erişilmiş, bu katılımcılardan alınan veriler PA ve AFA için kullanılmıştır. İkinci uygulamada toplanan 391 veri ile ise DFA yapılmıştır.

Veri analizleri için SPSS 25.0 (SPSS Inc.), Factor Analysis 10.10 (Ferrnando ve Lorenza-Seva, 2017), Lisrel 8.80 (Jöreskog ve Sorbom, 1993) ve jMetrik Version 4.1.1 istatistik paket programları kullanılmıştır. Ölçeğin yapı geçerliği PA, AFA, DFA aracılığıyla test edilmiştir. Ölçeğin yakınsak geçerliği için sürekli öfkeyle, ıraksak geçerlik için ise sosyal problem çözmeyle ilişkisine bakılmıştır. Güvenirlik analizleri kapsamında, madde-toplam korelasyonu, test tekrar test güvenirliği, Cronbach Alpha ve McDonald Omega iç tutarlık değerleri hesaplanmıştır. Ek olarak ölçek maddelerinin en yüksek ve düşük %27 grupta ayrışması t-test ile test edilmiştir. Ölçeği ölçüm değişmezliği erkek-kadın ve ergen-yetişkin örneklemlerle test edilmiştir.

Sonuç ve Tartışma

Sonuçlar, Saldırganlık Ölçeği Kısa Formu'nun Türkçede dört faktörden oluşan, güvenirliğe sahip, cinsiyetler arası ölçüm farkı olmayan, ergen ve yetişkinler için saldırganlık ölçümünde kullanılabilecek bir ölçüm aracı olduğunu göstermiştir.

Ölçeğin yapı geçerliğini test etmek için paralel analiz kullanılmıştır. Faktör belirlemede en önemli yöntemlerden birisi olarak görülen PA (Horn, 1965) sonuçları, dört faktörlü yapının doğrulandığını göstermektedir. Ölçek çalışmalarında, yapı ve alt yapıların nasıl ve kaç tane olacağını belirleyebilmek için birden çok yönteme başvurulması önerilmektedir (Erkuş, 2012). Bu doğrultuda çalışma kapsamında PA'nın yanı sıra ölçeğin faktör yapısı AFA ile de incelenmiştir. Elde edilen sonuçlar hem ergen hem de yetişkin gruplarda dört faktörlü orijinal yapıyı doğrulamaktadır. Yamaç eğrisi grafikleri de ergen ve yetişkin gruplar için ölçeğin dört faktörlü yapıya sahip olduğunu desteklemektedir. Ergen ve yetişkin gruplarda uygulanan PA ve AFA sonuçları maddelerin ilgili faktörün altında yüksek değerle gruplandığını göstermektedir.

PA ve AFA'dan sonra, DFA kullanılmıştır. Birinci ve ikinci düzey DFA sonuçları da ölçeğin dört faktörlü yapıda iyi uyum değerlerine sahip olduğunu göstermektedir. Ölçek birinci ve ikinci düzey DFA' ya ek olarak, tek boyutlu DFA ile de test edilmiştir. Sonuçlar dört faktörlü yapısının daha iyi uyum değerlerine sahip olduğunu ortaya koymuştur.

Ölçeğin faktörleri arasındaki korelasyon ergen ve yetişkin örneklemde incelenmiş orta düzey yakın yada orta düzeyde korelasyon gösterdikleri bulunmuştur. Ölçeğin yakınsak ve ıraksak geçerliği ergen ve yetişkin grupta test edilmiştir. Saldırganlığın sürekli öfkeyle beklenen yönde pozitif ve orta düzeyde anlamlı ilişkiye sahip olduğu bulunmuştur. Iraksak geçerlik ise sosyal problem çözme ile ilişkisine bakılmış ve ölçeğin ıraksak geçerliğe sahip olduğu belirlenmiştir.

Ergen ve yetişkin grup için güvenirlik analizleri kapsamında hesaplanan madde toplam korelasyonu değerlerinin 0,30'un üzerinde olduğu görülmüştür. Ölçeğin iç tutarlığı için incelenen Cronbach Alpha ve McDonald Omega değerlerinin hem ergen hem de yetişkin grup için yeterli düzeyde olduğu bulunmuştur. Ölçeğin iç tutarlık katsayısının hesaplanmasının dışında güvenirliği değerlendirmek için ölçeğin test-tekrar test güvenirliğine bakılmıştır. Ergen ve yetişkin gruplar için ölçeğin test-tekrar test güvenirliğine sahip olduğu görülmüştür.

Bu çalışmada Buss ve Perry Saldırganlık Ölçeği Kısa Formunun ergen ve yetişkin örneklemi için yeterli düzeyde psikometrik özelliklere sahip olduğu belirlenmiştir. Elde edilen sonuçlar birbirleriyle ve literatürle tutarlılık göstermektedir (Braynt ve Smith, 2001; Hornsveld ve diğerleri. 2009; Maxwell, 2007; Morales-Vives, Codorniu-Raga, ve Andreu Vigil-Colet, 2005). Ölçeğin kısa formu üzerinden test edilen faktör yapısının, ölçeğin uzun formundaki gibi güçlü bir faktör yapısına sahip olduğunu kanıtlamaktadır. Sonuçlar Saldırganlık Ölçeği Kısa Formu'nun geçerli ve güvenilir bir ölçek olduğunu, erkek, kadın, ergen ve yetişkin gruplar için kullanılabileceğini göstermiştir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(3); 266-282

A Short Note on Obtaining Item Parameter Estimates of IRT Models with Bayesian Estimation in Mplus

Sedat ŞEN * Allan COHEN ** Seock-ho KIM ***

Abstract
Parameter estimation of Item Response Theory (IRT) models can be applied using both Bayesian and non-Bayesian methods. Although maximum likelihood estimation (MLE), a non-Bayesian method, has predominated since the 1970s, there is an increasing use of Bayesian methods, due to their capability for estimating complex models and for their implementation in commercially available software. In view of the recent increase in the popularity of these methods, a comparison between model parameter estimates from the two types of methods would be useful for practitioners. In this study, we compare MLE and Bayesian estimation, two popular methods for obtaining parameter estimates for dichotomous IRT models, using the MLE and Bayes estimator options as implemented in the Mplus software package. Results indicated Bayesian and MLE estimates differed only slightly, clearly demonstrating the consistency between estimates from the two methods. Further, Bayes estimator option in Mplus can be a viable and relatively easy to use tool for calibrations of IRT models.

Key Words: Item response theory, dichotomous models, Bayesian estimation, Mplus.

INTRODUCTION

Item response theory (IRT) models have been used for testing over the last half-century (van der Linden & Hambleton, 2013). Parameter estimation is considered one of the important processes of IRT modeling. Estimates of IRT model parameters have typically been done using methods such as maximum likelihood estimation (MLE; Bock & Aitkin, 1981) and Markov chain Monte Carlo estimation (MCMC; Patz & Junker, 1999a). MLE methods are based on a frequentist approach, and MCMC is a Bayesian method. MLE-based estimation methods have been widely used in IRT modeling since the development of software such as BILOG (Zimowski, Muraki, Mislevy, & Bock, 2003), MULTILOG (Thissen, 1991) and PARSCALE (Muraki & Bock, 1996). Implementations of MCMC methods for estimation of IRT models began to be reported in the early 1990s (e.g., Albert, 1992; Albert & Chib, 1993). MLE generally requires large samples to produce reliable results (e.g., Asparouhov & Muthén, 2010a; Meuleman & Billiet, 2009), a condition not necessarily required by Bayesian methods.

In this regard, there are often advantages with Bayesian methods that can overcome some of the problems associated with MLE methods. Lee and Song (2004) note that Bayesian methods can provide asymptotically distribution-free estimates, as well as more accurate results with smaller samples with non-normal ability distribution (see also Gao, & Chen, 2005). Fox (2010) suggests that, unlike MLE, Bayesian methods enable the use of additional information for estimating model parameters in addition to providing smaller standard errors than those of marginal maximum likelihood estimate, when reasonable prior information is available. Thus, MCMC implementations, such as Metropolis-Hastings and Gibbs sampling, became increasingly popular for IRT modeling after 1990s, particularly for

* Assoc. Prof. Dr., Harran University, Faculty of Education, Educational Sciences, Şanlıurfa-Turkey, sedatsen@harran.edu.tr, ORCID ID: orcid.org/0000-0001-6962-4960
** Prof. Dr., University of Georgia, College of Education, Educational Psychology Department, Athens-Georgia-USA, acohen@uga.edu, ORCID ID: orcid.org/0000-0002-8776-9378
*** Prof. Dr., University of Georgia, College of Education, Educational Psychology Department, Athens-Georgia-USA, shkim@uga.edu, ORCID ID: orcid.org/0000-0002-2353-7826

To cite this article:
Şen, S., Cohen, A., & Kim, S.(2020). *A Short Note on Obtaining Item Parameter Estimates of IRT Models with Bayesian Estimation in Mplus, 11*(3), 266-282. doi: 10.21031/epod.693719.

Received: 24.02.2020
Accepted: 15.08.2020

complex models (Rupp, Dey & Zumbo, 2004). Albert (1992), and Albert and Chib (1993), for example, demonstrated the application of the Gibbs sampler for two-parameter normal-ogive models. More generally, Patz and Junker (1999a, 1999b) provided details of Metropolis-Hastings procedures within Gibbs for logistic IRT models. Baker (1998), Ghosh, Ghosh, Chen, and Agresti (2000), and Sheng (2010) provide applications of Gibbs sampling. In addition to these applications, Bayesian estimation has been reported for other IRT models, including the 4PL model (Culpepper, 2016), the multi-unidimensional graded response IRT model (Kuo & Sheng, 2015), and the mixture IRT model (Bolt, Cohen, & Wollack, 2001).

MCMC estimation of IRT requires respondents' data and prior distributions for the model parameters. Gibbs sampling is one of the MCMC algorithms that can be used to estimate the parameters of IRT models. This method summarizes the joint posterior distribution of (θ, ξ) by simulation (Albert, 1992). After collecting the examinee responses in the data matrix **Y**, suppose we have a vector (**ω**) including item (ξ) and person (θ) parameters. Implementation of Gibbs sampling starts with initial guesses on this vector. Let $\boldsymbol{\omega}^{(0)} = \left(\boldsymbol{\omega}_1^{(0)}, \ldots, \boldsymbol{\omega}_k^{(0)}\right)$ denote initial values of **ω**. As explained by Wollack, Bolt, Cohen, and Lee (2002, p.340), "a single sample from the joint posterior distribution, $\pi(\omega|\mathbf{Y})$, is approximated, by sampling each parameter from its conditional posterior distribution." A cycle is completed after sampling the first set of parameters. The parameters of the previous cycle serve as conditional values for the next cycle. We need to iterate the cycles *t* times until convergence is achieved. Mean, mode, or median of each parameter's marginal posterior distribution can be used as the final estimates (Baker, 1998). Readers are referred to Albert (1992) and Baker (1998) for more details on Gibbs Sampling.

One characteristic of the implementation of MCMC algorithms is that they can be quite technically complex. Fortunately, implementations of Bayesian estimation algorithms are often available from authors of published articles (Curtis, 2010). Early applications of Bayesian IRT estimation were mainly implemented in WINBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003). Thus, most of the Bayesian estimation programs have been written in the BUGS language (Curtis, 2010). Over the years, other software packages have also been developed for Bayesian estimation of IRT including JAGS (Plummer, 2003), STAN (Stan Development Team, 2016), and OPENBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2010), which is the more recent version of WinBUGS. These packages are designed primarily for Bayesian estimation. Bayesian algorithms have become available in several general purpose software packages including SAS (e.g., Proc MCMC; SAS Institute, 2017), S-PLUS, R packages including DPpackage (Jara, Hanson, Quintana, Mueller, & Rosner 2012), mlirt (Fox, 2007), MCMCpack (Martin, Quinn, & Park, 2011), pscl (Jackman et al., 2017) and MATLAB (The MathWorks Inc., 2010).

Mplus (Muthén, & Muthén, 1998-2019) recently implemented a Bayesian MCMC algorithm option for latent variable models (Asparouhov & Muthén, 2010). Mplus is a general latent variable modeling program that implements the estimation of several statistical model families, including structural equation models, latent class analysis, factor analysis, mixture models with both single- and multi-level data structures. Mplus also estimates a range of IRT models (Embretson, & Reise, 2000) including the one-, two- and three-parameter logistic models (1PL, 2PL, and 3PL; Birnbaum, 1968), the four-parameter logistic model (4PL; Barton & Lord, 1981), the partial credit model (PCM; Masters & Wright, 1997), and the generalized partial credit model (GPCM; Muraki, 1997). Multi-level and mixture extensions of these models are also possible within Mplus. A relatively small number of studies have reported results use of Mplus for Bayesian estimation of IRT models (e.g., Luo & Dimitrov, 2019). Luo and Dimitrov (2019) have shown how to obtain estimates for MCMC/EAP of the ability parameters when using Bayes estimator in Mplus. Their study has focused on ability parameters. However, studies investigating the Bayesian estimation of item parameters of IRT models using Mplus do not yet appear to have been reported. As the popularity of the Mplus software package increases, the user may want to learn how to create syntaxes to be able to estimate item parameters of IRT models using Mplus software package. A didactic paper on this issue would be very helpful for practitioners. To this end, the purpose of this study is to introduce Bayes estimator of Mplus and to demonstrate its application in obtaining item parameter estimates of dichotomous IRT models.

METHOD

Empirical Example

Estimation of dichotomous IRT models on a sample data set using the Bayesian estimation algorithm implemented in Mplus will be demonstrated in this part. In addition, Bayesian estimates were compared with ML based estimates as implemented in Mplus.

Data

Section 6 of the Law School Admission Test data set (LSAT6; Bock & Lieberman, 1970) was analyzed in this study. The LSAT6 data set consists of the item responses of 1,000 examinees to five multiple choice five-choice items on Figure Classification. Multiple choice items are coded as 0 and 1 for wrong and right answers, respectively. LSAT6 data set was preferred in this study as it has been studied in many IRT studies and it is available with most of the IRT software packages. This data set is known to be useful example of testing out IRT procedures and showing the use of Bayesian estimation.

IRT Models

Only Rasch, 1PL, and 2PL dichotomous models were considered in the present study as the Bayes estimator is not currently available in Mplus for other IRT models (e.g., 3PL, 4PL, and polytomous models). Under the 2PL IRT model, the probability of a correct response to an item can be given as follows:

$$P(X_{is} = 1|\theta_s, \beta_i, \alpha_i) = \frac{\exp[\alpha_i(\theta_s - \beta_i)]}{1 + \exp[\alpha_i(\theta_s - \beta_i)]}, \quad (1)$$

where θ_s is the person ability parameter for examinee *s*, and β_i and α_i are the difficulty and discrimination parameters, respectively, for item *i*. The 1PL model can be written from Equation 1 by setting the item discrimination parameters to a constant value (e.g., the average item discrimination). Similarly, the Rasch model equation may be written by setting the item discrimination parameters to 1. Given that conditional independence assumption, the conditional likelihood of a response pattern would be given as

$$L(u_{s1}, u_{s2}, \dots, u_{sI}) = \prod_{i=1}^{I} P_i(\theta_s)^{u_{si}} Q_i(\theta_s)^{1-u_{si}}, \quad (2)$$

where $Q_i(\theta_s) = 1 - P_i(\theta_s)$.

Estimation

Parameter estimates of the model in Equation 1 can be obtained with either MLE or Bayesian estimation. The MLE algorithm focuses on finding the ability-level estimates that maximize the log-likelihood function with iterative procedures like Newton-Raphson (Embretson & Reise, 2000). Bayesian estimation integrates the prior distribution into the likelihood function. In Mplus, Bayesian estimation is on the probit scale, and MLE is on the logistic scale. When using the probit model, the posterior distribution of θs can be defined as

$$p(\theta_s \mid Y) \propto L \times p(\theta_s) \quad (3)$$

where $p(\theta_s)$ is the prior. As can be seen in Equation 3 (right side), prior distributions required to define the posterior distributions. These are used with the specified model in determining the posterior. The central tendency statistics of the posterior distribution can be reported as the final estimates.

Bayesian Estimation with Mplus

Bayesian estimation in Mplus is not currently available for count, continuous-time survival, censored or nominal data, nor is Bayesian estimation currently available for the 3PL, graded response model, partial credit model, or the generalized partial credit model.

The ESTIMATOR option should be defined as "ESTIMATOR=BAYES;" to obtain Bayes estimates in Mplus. Other Bayesian related options in the ANALYSIS command include ALGORITHM, BCONVERGENCE, BITERATIONS, BSEED, CHAINS, FBITERATIONS, POINT, PREDICTOR, PRIOR, STVALUES, and THIN. Descriptions of these options can be found in Mplus User's Guide (Muthén & Muthén, 1998-2019). The non-informative priors are the default for Bayes estimator in Mplus. However, users can specify informative priors such as inverse Gamma, Inverse Wishart, and Dirichlet (Muthén & Asparouhov, 2013).

Analyses

Three dichotomous IRT models (i.e., Rasch, 1PL, and 2PL models) were fit to the LSAT6 data set using MLE and Bayesian options in Mplus. For comparison, the MLE estimator was used as implemented in the Mplus software. Bayesian estimation with Mplus was applied with non-informative default priors $N(0, 5)$ (Muthén & Muthén, 1998-2019, p. 775) for item difficulty and discrimination parameters. A detailed list of the alternative prior distributions can be found in Mplus User's Guide (Muthén & Muthén, 1998-2019, p. 775). The Mplus syntaxes for Bayesian estimation of the three models are presented in Figures 1 to 3, respectively. All model parameters were estimated from the posterior distribution. LSAT6 data have been previously analyzed with MCMC by Kim (2003). In that study, 5,000 iterations were used as burn-in based on Gelman and Rubin diagnostic information. In this study, the same number of iterations was used for burn-in stage. Thus, a total of 10,000 MCMC iterations were run; the first 5,000 iterations were used as burn-in. Posterior means of the sampled values for each parameter were taken as parameter estimates.

In the Mplus syntax for the Bayesian estimation of the Rasch model (see Figure 1), the TITLE command was used to describe the problem. The FILE option was added for the data set (FILE = LSAT6.txt;). As the LSAT6 data set consists of five dichotomous items, the "NAMES = item1-item5;" option was used to label these five items under the VARIABLE command. The CATEGORICAL option was used to specify these items as categorical (CATEGORICAL = item1-item5;). USEVAR option was used to specify to use all of the items in the analysis (USEVAR = item1-item5;). The ANALYSIS command, ESTIMATOR = BAYES; option was specified to obtain Bayesian estimates. The Bayes estimator in Mplus employs Gibbs sampling (line 10) with two processors (line 11) to run two parallel chains (line 12). FBITERATIONS option was used to set the number of iterations to be 10,000 (line 13). That is, the posterior is based on the last half (i.e., 5,000). STVAL=ML; option was used to specify starting value information as ML-based values. In the syntax, we specified the mean of the posterior distribution to be reported in the output (POINT=MEAN;). THIN indicates the number of iterations to be used for estimating the posterior. In this example, thinning was set to be 5 (see line 16), which means every 5^{th} iteration is used for estimating the posterior. The MODEL command (see line 17) indicates a general factor f1 (see line 18) and use the command "by" to indicate the relationship between items and factor f1. As can be seen in Figure 1, f1 by item1@0.587544 item2-item5@0.587544; (line 18) was used to link five items (items1-item5) with factor f1. The @0.587544 part was used to fix item discrimination parameters at 1, which enable us to obtain a model based on Rasch framework. f1@1 option was used fix factor variance to one. In addition, the mean of f1 was fixed at zero by writing [f1@0];. Specifications between lines 17 and 20 were used to obtain Rasch model. The PLOT command was also used to create plots after estimation. As mentioned in the Mplus User's Guide, TYPE = PLOT3; can be used to request graphical displays of several results. STANDARDIZED option was used to obtain standardized solution, and TECH8 was added to request diagnostic information regarding model convergence in the OUTPUT section (line 24). TECH8 also shows the total number of iterations, including the discards. Mplus reports the Potential Scale Reduction (PSR) computed based on Gelman and Rubin's convergence diagnostic (Gelman & Rubin, 1992) when included TECH8 under the OUTPUT command.

In addition, a posterior predictive checking (PPC) statistics and its p-value (PPP) can be obtained for model fit assessment in Mplus. This statistics is based on the usual $\chi 2$ test, which is computed using the replicated and the observed data in MCMC iteration *t*. The $\chi 2$ fit function difference between these two values is calculated using every 5th iteration. For an excellent-fitting model, 95% CI for the difference in $\chi 2$ value should include zero (Wang & Wang, 2020). Poor fit is observed with low PPP values (e.g. <0.05).

```
1   TITLE: Bayesian Estimation of Rasch Model
2   DATA:
3   FILE = LSAT6.txt;
4   VARIABLE:
5   NAMES = item1-item5;
6   CATEGORICAL = item1-item5;
7   USEVAR = item1-item5;
8   ANALYSIS:
9   ESTIMATOR = BAYES;
10  ALGORITHM=GIBBS;
11  PROCESS=2;
12  CHAINS=2;
13  FBITERATIONS=10000;
14  STVAL=ML;
15  POINT=MEAN;
16  THIN=5;
17  MODEL:
18  f1 by item1@0.587544 item2-item5@0.587544; !1/1.702=0.587544
19  f1@1;
20  [f1@0];
21  PLOT:
22  TYPE = PLOT3;
23  OUTPUT:
24  STANDARDIZED TECH8;
```

Figure 1. Mplus Syntax for Bayesian Estimation of Rasch Model.

The Mplus syntax created for Bayesian estimation of 1PL model is presented in Figure 2. Since most parts of this syntax are similar to the syntax of the previous model, only the different parts are explained here. The main difference between the syntaxes in Figure 1 and Figure 2 is that the MODEL section was modified to be able to estimate 1PL model. The f1 by item1-item5* (loading); option (line 18) was used to fix all discrimination parameters to be equal, and the [item1$1-item5$1*] option (line 19) was used to freely estimate the item difficulty parameters.

```
1  TITLE: Bayesian Estimation of 1PL Model
2  DATA:
3  FILE = LSAT6.txt;
4  VARIABLE:
5  NAMES = item1-item5;
6  CATEGORICAL = item1-item5;
7  USEVAR = item1-item5;
8  ANALYSIS:
9  ESTIMATOR = BAYES;
10 ALGORITHM=GIBBS;
11 PROCESS=2;
12 CHAINS=2;
13 FBITERATIONS=10000;
14 STVAL=ML;
15 POINT=MEAN;
16 THIN=5;
17 MODEL:
18 f1 by item1-item5* (loading);
19 [item1$1-item5$1*]
20 [f1@0]; f1@1;
21 PLOT:
22 TYPE = PLOT3;
23 OUTPUT:
24 STANDARDIZED TECH8;
```

Figure 2. Mplus Syntax for Bayesian Estimation of 1PL IRT Model.

The Mplus syntax created for Bayesian estimation of 2PL model is presented in Figure 3. Since most parts of this syntax are similar to the syntax of the previous model, only the different parts are explained here. The most noticeable difference between the syntaxes in Figure 1 and Figure 3 is is that the MODEL section was modified to be able to estimate 2PL model. Lines 18 and 19, f1 by item1-item5; and [item1$1-item5$1*] options were used to freely estimate the item discrimination and difficulty parameters, respectively.

```
1  TITLE: Bayesian Estimation of 2PL Model
2  DATA:
3  FILE = LSAT6.txt;
4  VARIABLE:
5  NAMES = item1-item5;
6  CATEGORICAL = item1-item5;
7  USEVAR = item1-item5;
8  ANALYSIS:
9  ESTIMATOR = BAYES;
10 ALGORITHM=GIBBS;
11 PROCESS=2;
12 CHAINS=2;
13 FBITERATIONS=10000;
14 STVAL=ML;
15 POINT=MEAN;
16 THIN=5;
17 MODEL:
18 f1 by item1-item5;
19 [item1$1-item5$1*]
20 [f1@0]; f1@1;
21 PLOT:
22 TYPE = PLOT3;
23 OUTPUT:
24 STANDARDIZED TECH8;
```

Figure 3. Mplus Syntax for Bayesian Estimation of 2PL IRT Model.

ML estimates of item parameters of three IRT models with LSAT6 data set were also obtained with Mplus software package for comparison. In order to transform the Mplus derived parameter estimates (i.e., location and thresholds) to typical IRT estimates (i.e., discrimination and difficulty), we followed the transformation formula provided by Asparouhov and Muthén (2016, p.6). Given that μ represents factor mean and ψ denotes the factor variance, item discrimination parameter ($\boldsymbol{a_i}$) and the item difficulty parameter ($\boldsymbol{\beta_i}$) can be calculated as below using the location ($\boldsymbol{\lambda_i}$) and threhold ($\boldsymbol{\tau_i}$) parameters from Mplus output:

$$\boldsymbol{\alpha_i = \lambda_i\sqrt{\psi}} \quad (4)$$

$$\boldsymbol{\beta_i = \frac{\tau_i - \lambda_i \mu}{\lambda_i\sqrt{\psi}}} \quad (5)$$

Forero and Maydeu-Olivares (2009) use the normal and logistic versions of the IRT models to place parameter estimates and standard errors in the same metric (within .01 units). The constant D (i.e., 1.702) proposed by Haley (1952) was used in this study as well to convert normal ogive parameters to the logistic parameters.

RESULTS

Tables 1-3 contain approximate posterior means using the MCMC algorithm for the five item parameters for the three IRT models fit the LSAT6 data set. Bayesian estimation and the MLE estimates of item parameters for Rasch, 1PL, and 2PL models were compared. In order to rescale parameters obtained from the Bayesian estimation, all of the parameters were multiplied by 1.7, since the Bayesian estimation in Mplus uses the probit link function.

Traditional fit indices cannot be used when IRT models are estimated with Bayesian estimation. In this case, the convergence of the Markov chain should be checked using Bayesian diagnostic statistics. Mplus reports the potential scale reduction (PSR) for convergence assessment (Asparouhov & Muthén, 2010b). PSR values between 1 and 1.1 indicate good convergence (Wang & Wang, 2020). In this study, PSR values were found to be 1.024, 1.018, and 1.042 for Rasch, 1PL, and 2PL models, respectively. Under the MCMC estimation, the PPC statistics can be used to assess model fit. Mplus provides a $\chi 2$ fit function difference between observed values and replicated estimates, a p-value (PPP), scatter plot and histogram. $\chi 2$ fit function differences were found to be [-14.585, 22.300], [-18.232, 16.381], and [-17.844, 17.061] for Rasch, 1PL, and 2PL models, respectively. Associated p-values were 0.346, 0.545, and 0.505, respectively. All of the 95% CI's include zero, and associated p-values are high indicating that three IRT models fit the data very well. In addition, plots based posterior predictive checking (see Figures A3 and A6) also indicate that the fit of the model is reasonable. Table 1 lists item parameter estimates of the Rasch model from Mplus for the MLE and Bayes estimators. Mplus-ML columns present MLE estimates, and the last column displays Bayesian estimates. The item parameter estimates from Bayesian estimation in Mplus are not on the same scale as those from MLE. Therefore, to convert the item difficulty and item discrimination parameters to the logit scale, we applied Equations 4 and 5 to the Bayesian estimates. The results presented in Table 1 in the right column (headed Mplus-Bayes) are on the logit scale. As shown in Table 1, the transformed Bayesian estimates differed in the first and second decimal places compared to the MLE estimates. The Pearson correlation value between Bayes estimates and ML estimates was .999.

Table 1. Item Parameter Estimates of Rasch Model

Item parameter	Mplus-ML	Mplus-Bayes
β_1	-2.871	-2.791
β_2	-1.062	-1.076
β_3	-0.257	-0.260
β_4	-1.387	-1.396
β_5	-2.218	-2.194
α	1.000	1.000

Note. β=item difficulty; α=item discrimination.

Table 2 presents the item parameter estimations for 1PL IRT model from the MLE and Bayes estimators in Mplus. As shown in Table 2, the Bayesian estimates from Mplus yielded similar estimates to those obtained with the MLE estimator. As shown in Table 2, the transformed Bayesian estimates differed in the first and second decimal places compared to the MLE estimates. The Pearson correlation value between Bayes estimates and ML estimates was .999.

Table 2. Item Parameter Estimates of 1PL IRT Model

Item parameter	Mplus-ML	Mplus-Bayes
β_1	-3.606	-3.571
β_2	-1.319	-1.372
β_3	-0.317	-0.333
β_4	-1.726	-1.783
β_5	-2.773	-2.804
α	0.757	0.745

Note. β=item difficulty; α=item discrimination.

Table 3 presents the item parameter estimates of the 2PL IRT model in Mplus with the two estimators. As can be seen, the Bayes estimator in Mplus also yielded similar estimates to those obtained with MLE (see Table 3). For item discrimination parameters, the Pearson correlation value between Bayes estimates and ML estimates was .87. For item difficulty parameters, the Pearson correlation value between Bayes estimates and ML estimates was .999. As shown in Table 3, the transformed Bayesian estimates differed in the first and second decimal places compared to the MLE estimates.

Several plots are possible in the "View plots" submenu under the "Plots" menu of Mplus screen. A screen opens and shows several options that can be used to request graphical displays of several results. As an example, Bayesian related plots for item 1 estimated with the 1PL IRT model are presented in the Appendix (see Figures A1-A7).

Table 3. Item Parameter Estimates of the 2PL IRT Model

Item parameter	Mplus-ML	Mplus-Bayes
α_1	0.825	0.727
α_2	0.722	0.749
α_3	0.891	0.934
α_4	0.688	0.689
α_5	0.659	0.616
β_1	-3.360	-3.672
β_2	-1.371	-1.373
β_3	-0.280	-0.279
β_4	-1.868	-1.916
β_5	-3.117	-3.323

Note. β=item difficulty; α=item discrimination.

DISCUSSION and CONCLUSION

Dichotomous IRT models are typically estimated with both Bayesian and MLE estimation algorithms. Bayesian estimation of IRT models is sometimes preferable to MLE as MLE needs numerical integration, which can be slow or prohibitive depending on the numbers of dimensions of integration as a function of the numbers of latent variables. Tutorials do not yet appear to be presented in the psychometric literature for estimating dichotomous IRT models in Mplus using Bayesian estimation. In this paper, we provide a simplified step-by-step method for the estimation of dichotomous IRT models with Bayesian estimation.

Specifically, three dichotomous IRT models were analyzed using the five-item LSAT6 data set. Parameter estimates of these five items were compared for MLE and Bayesian estimations. Results suggested that there were some differences between item parameter estimates from MLE and Bayesian

estimation. This is consistent with the results of previous research that showed comparable estimation results for MLE and MCMC (e.g., Kieftenbeld & Natesan, 2012; Luo, 2018; Paek, Cui, Öztürk Gübeş, & Yang, 2018; Wollack, Bolt, Cohen, & Lee, 2002). For instance, Luo (2018) found very close estimates between MCMC and robust ML (MLR) estimation of 2PL testlet model in Mplus. Kieftenbeld and Natesan (2012) have found little difference between the estimates obtained from MML and MCMC estimation of the graded response model. Similarly, Wollack et al. (2002) have also demonstrated that item parameter estimates from MMLE and MCMC methods were very similar under the nominal response model.

Bayesian estimation with non-informative priors should give asymptotically similar estimates as MLE. As shown in this study, non-informative priors are the default for the Bayes estimator in Mplus, although several informative priors can be defined. The prior distribution is one of the most important aspects of Bayesian estimation. This is because prior distributions can substantially affect the posterior distribution especially for small sample sizes (Natesan, Nandakumar, Minka, & Rubright, 2016). Only non-informative priors were used in this study. However, it is possible to add informative or slightly informative priors to the estimation in Mplus software. Further studies may reveal the differential effect of using different priors in Mplus. The addition of the Bayesian estimation algorithm in Mplus makes Mplus a viable and useful software package for the estimation of dichotomous IRT models.

REFERENCES

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17*(3), 251–269.

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association, 88*(422), 669–679.

Asparouhov, T., & Muthén, B. (2010a). *Bayesian analysis of latent variable models using Mplus*. (Technical Report). Version 4. Retrieved from http://www.statmodel.com/download/BayesAdvantages18.pdf

Asparouhov, T., & Muthén, B. (2010b). *Bayesian analysis using Mplus, version 4*. Technical report. Los Angeles: Muthén and Muthén. www.statmodel.com.

Asparouhov, T., & Muthén, B. (2016). *IRT in Mplus* (Technical report). Los Angeles, CA: Muthén & Muthén.

Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement, 22*(2), 153–169.

Barton, M.A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model* (ETS RR- 81-20). Princeton, NJ: Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick, *Statistical theories of mental test scores* (pp. 395-479). Reading, Massachusetts: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443–459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35*(2), 179–197.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics, 26*(4), 381–409.

Culpepper, S. A. (2016). Revisiting the 4-parameter item response model: Bayesian estimation and application. *Psychometrika, 81*(4), 1142–1163.

Curtis, S. M. (2010). BUGS code for item response theory. *Journal of Statistical Software, 36*(1), 1–34.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Hillsdale, NJ: Erlbaum.

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods, 14*(3), 275–299.

Fox, J. P. (2007). Multi-level IRT modeling in practice with the Package mlirt. *Journal of Statistical Software, 20*(5), 1–16.

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer.

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.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–472.

Ghosh, M., Ghosh, A., Chen, M. H., & Agresti, A. (2000). Non-informative priors for one-parameter item response models. *Journal of Statistical Planning and Inference, 88*(1), 99–115.

Haley, D. C. (1952). *Estimation of the dosage mortality relationship when the dose is subject to error.* (Tech. Rep. No.15, Office of Naval Research Contract No. 25140, NR-342–022). Stanford, CA: Stanford University, Applied Mathematics and Statistics Laboratory.

Jackman, S., Tahk, A., Zeileis, A., Maimone, C., Fearon, J., Meers, Z., ... & Imports, M. A. S. S. (2017). Package 'pscl'. See http://github. com/atahk/pscl.

Jara, A., Hanson, T., Quintana, F. A., Mueller, P., & Rosner, G. (2012). DPpackage: Bayesian nonparametric modeling in R. R package version, 1-1.

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 36*(5), 399–419.

Kim, S. H. (2003). An investigation of Bayes estimation procedures for the two-parameter logistic model. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics* (pp. 389-396). Tokyo: Springer.

Kuo, T. C., & Sheng, Y. (2015). Bayesian estimation of a multi-unidimensional graded response IRT model. *Behaviormetrika, 42*(2), 79–94.

Lee, S. Y., & Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research, 39*(4), 653–686.

Luo, Y. (2018). A short note on estimating the testlet model with different estimators in Mplus. *Educational and Psychological Measurement, 78*(3), 517–529.

Luo, Y., & Dimitrov, D. M. (2019). A short note on obtaining point estimates of the IRT ability parameter with MCMC estimation in Mplus: How many plausible values are needed?. *Educational and Psychological Measurement, 79*(2), 272–287.

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). Mcmcpack: Markov chain Monte Carlo in R. *Journal of Statistical Software, 42*(9), 1–21.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–121). New York, NY: Springer.

Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multi-level SEM?. *Survey Research Methods, 3*(1), 45–58.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–164). New York, NY: Springer.

Muraki, E., & Bock, R. D. (1996). PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks (Version 3) [Computer software]. Chicago, IL: Scientific Software.

Muthén, B., & Asparouhov, T. (2013). Item response modeling in Mplus: A multi-dimensional, multi-level, and multi-timepoint example. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of item response theory: Models, statistical tools, and applications* (Vol 1, pp. 527-539). Boca Raton, FL: Chapman & Hall/CRC Press.

Muthén, L. K., & Muthén, B. O. (1998-2019). Mplus (version 8.3)[computer software]. Los Angeles, CA: Muthén & Muthén.

Natesan, P., Nandakumar, R., Minka, T., & Rubright, J. D. (2016). Bayesian prior choice in IRT estimation using MCMC and variational Bayes. *Frontiers in Psychology, 7*, 1422.

Paek, I., Cui, M., Öztürk Gübeş, N., & Yang, Y. (2018). Estimation of an IRT model by Mplus for dichotomously scored responses under different estimation methods. *Educational and Psychological Measurement, 78*(4), 569–588.

Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), 146–178.

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*(4), 342–366.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Proceedings of the 3rd international workshop on distributed statistical computing (Vol. 124, p. 125). Technische Universita¨t Wien, Vienna, Austria.

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling, 11*(3), 424–451.

SAS Institute. (2017). *Base SAS 9.4 procedures guide: Statistical procedures*. SAS Institute.

Sheng, Y. (2010). A sensitivity analysis of Gibbs sampling for 3PNO IRT models: Effects of prior specifications on parameter estimates. *Behaviormetrika, 37*(2), 87–110.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). WinBUGS version 1.4 [Computer program]. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2010). *OpenBUGS Version 3.1.1 user manual*. Retrieved from http://www.openbugs.net/Manuals/Manual.html

Stan Development Team. (2016). *Stan modeling language users guide and reference manual* (Version 2.12.0). Retrieved from http://mc-stan.org/documentation/

The MathWorks Inc. (2010). MATLAB. [Computer software]. Natick, MA: The MathWorks Inc.

Thissen, D. (1991). MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory (Version 6.0) [Software manual]. Chicago, IL: Scientific Software.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (2013). *Handbook of modern item response theory*. Springer Science & Business Media.

Wang, J., & Wang, X. (2020). *Structural equation modeling: Applications using Mplus*. Hoboken, NJ: John Wiley & Sons.

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y. S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 26*(3), 339–352.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG [computer software]. Lincolnwood, IL: Scientific Software.

Mplus'ta Bayes Kestirimi ile IRT Modellerinin Madde Parametre Kestirimlerinin Elde Edilmesine İlişkin Kısa Bir Not

Giriş

Madde tepki kuramı (MTK) modelleri 1970'lerden beri çeşitli test uygulamalarında kullanılmaktadır (van der Linden ve Hambleton, 2013). MTK modellemesinde en önemli adımlardan birini parametre tahmini oluşturmaktadır. MTK modeline ait madde ve yetenek parametrelerinin tahmini tipik olarak maksimum olabilirlik tahmini (MLE; Bock & Aitkin, 1981) ve Markov zinciri Monte Carlo tahmini (MCMC; Patz ve Junker, 1999a) gibi yöntemler kullanılarak yapılmaktadır. MLE tabanlı tahmin yöntemleri, BILOG (Zimowski, Muraki, Mislevy ve Bock, 2003), MULTILOG (Thissen, 1991) ve PARSCALE (Muraki ve Bock, 1996) gibi yazılımların geliştirilmesinden bu yana MTK modellemesinde yaygın olarak kullanılmaktadır. MTK modellerinin tahmini için MCMC yöntemlerinin uygulamaları 1990'ların başında gösterilmeye başlanmıştır (ör. Albert, 1992; Albert ve Chib, 1993). Kararlı sonuçlar üretmek adına MLE yöntemi, genellikle Bayesci yöntemlerin gerektirdiği bir koşul olmayan, büyük örneklemlere ihtiyaç duyar (örneğin, Asparouhov ve Muthén, 2010; Meuleman ve Billiet, 2009). Bu bağlamda, Bayesci kestirim yöntemleri MLE yöntemleriyle ilişkili bazı sorunların üstesinden gelebilecek avantajlara sahiptir. Lee ve Song (2004) Bayesci kestirim yöntemlerinin asimptotik olarak dağılımsız serbest tahminler sağlayabileceğini ve normal olmayan yetenek dağılımına sahip daha küçük örneklemlerle daha doğru sonuçlar verebileceğini belirtmektedir. Fox (2010), MLE'den farklı olarak Bayesci kestirim yöntemlerinin, makul önsel (prior) bilgiler mevcut olduğunda marjinal maksimum olabilirlik tahmininden daha küçük standart hatalar sunmasının yanı sıra model parametrelerini tahmin etmek için ek bilgilerin kullanılmasını sağladığını ileri sürmektedir. Böylece, Metropolis-Hastings ve Gibbs örneklemesi gibi MCMC uygulamaları, 1990'lardan sonra, özellikle nispeten karmaşık modeller için veya verilerin az olduğu ve asimptotik teorinin tutulmasının pek mümkün olmadığı durumlarda MTK modellemesi için giderek daha popüler hale gelmiştir (Rupp, Dey ve Zumbo, 2004). Örneğin Albert (1992) ve Albert ve Chib (1993), iki parametreli normal-ogive modelleri için Gibbs örnekleyicisinin uygulanabileceğini göstermişlerdir. Daha genel olarak, Patz ve Junker (1999a, 1999b) lojistik MTK modelleri için Gibbs içinde Metropolis-Hastings prosedürlerinin ayrıntılarını göstermiştir. Baker (1998), Ghosh, Ghosh, Chen ve Agresti (2000) ve Sheng (2010) çalışmaları Gibbs örneklemesi uygulamaları sunmaktadır. Bu uygulamalara ek olarak, 4 parametreli MTK modeli (Culpepper, 2016), çok boyutlu derecelendirilmiş yanıt MTK modeli (Kuo ve Sheng, 2015) ve karma MTK modeli (Bolt, Cohen, ve Wollack, 2001) dahil olmak üzere diğer MTK modelleri için Bayesci kestirimi kullanılmıştır. MTK'nin MCMC tahmini, yanıtlar için bir olasılık modelinden, model parametreleri için önsel dağılımlardan ve muhtemelen hiperparametreler için önsel dağılımlardan oluşur. MCMC algoritmalarının uygulanmasının bir özelliği, teknik olarak oldukça karmaşık olmalarıdır. Neyse ki, Bayesci kestirim algoritmalarının uygulamaları genellikle yayınlanmış

makalelerin yazarlarından elde edilebilmektedir (Curtis, 2010). Bayesci MTK tahmininin ilk uygulamaları esas olarak WINBUGS yazılımında uygulanmıştır (Spiegelhalter, Thomas, Best ve Lunn, 2003). Bu nedenle, Bayesci tahmin programlarının çoğu BUGS dilinde yazılmıştır (Curtis, 2010). Yıllar içinde, JAGS (Plummer, 2003), STAN (Stan Geliştirme Ekibi, 2016) ve OPENBUGS (WinBUGS'ın yeni sürümü; Spiegelhalter, Thomas, Best ve Lunn, 2010) dahil olmak üzere MTK'nin Bayesci tahmini için başka yazılım paketleri de geliştirilmiştir. Bu paketler öncelikle Bayesci kestirim için tasarlanmıştır. Bayesci algoritmaları, MATLAB (The MathWorks Inc., 2010), SAS (ör. Proc MCMC; SAS Enstitüsü, 2017), S-PLUS, MCMCpack (Martin, Quinn ve Park, 2011), pscl (Jackman ve diğerleri, 2017), DPpackage (Jara, Hanson, Quintana, Mueller ve Rosner 2012), mlIRT (Fox, 2007) gibi R paketleri de dahil olmak üzere çeşitli genel amaçlı yazılım paketlerinde kullanılabilir hale gelmiştir.

Mplus (Muthén ve Muthén, 1998-2019) kısa bir süre önce örtük değişken modelleri için bir Bayesci MCMC algoritma seçeneği uygulamaya başlamıştır (Asparouhov ve Muthén, 2010). Mplus, yapısal eşitlik modelleri, örtük sınıf analizi, faktör analizi, hem tek hem de çok seviyeli veri yapılarına sahip karma modelleri de dâhil olmak üzere birçok istatistiksel model ailesinin tahminini uygulayan genel bir örtük değişken modelleme programıdır. Mplus ayrıca 1 parametreli lojistik model (1PL), 2 parametreli lojistik model (2PL), 3 parametreli lojistik model (3PL; Birnbaum, 1968), 4 parametreli lojistik model (4PL; Barton ve Lord, 1981), kısmi kredi modeli (PCM; Masters ve Wright, 1997) ve genelleştirilmiş kısmi kredi modeli (GPCM; Muraki, 1997) dahil olmak üzere bir dizi MTK modelini (Embretson ve Reise, 2000) tahmin edebilmektedir. Bu modellerin çok düzeyli ve karma uzantıları Mplus içerisinde de mümkündür. Nispeten az sayıda çalışma, MTK model parametrelerinin Bayesci kestirimi için Mplus kullanımı sonuçlarını raporlamıştır (örn. Luo ve Dimitrov, 2019). Mplus kullanarak MTK modellerinin madde parametrelerinin Bayesci tahminini araştıran çalışmalar henüz rapor edilmemiştir. Bu çalışmanın amacı Mplus kullanarak Bayesci kestirimine bir giriş sağlamak ve bunun iki kategorili (dikotomus) MTK modellerinin tahmininde uygulanmasını göstermektir.

Yöntem

Bu bölümde, Mplus'ta uygulanan Bayesci kestirim algoritmasını kullanarak ampirik bir veri seti üzerinde iki kategorili MTK modellerinin madde parametre tahminleri gösterilmektedir. Ayrıca, bu madde parametre tahminleri Mplus'ta uygulanan maksimum olabilirlik tahminininden elde edilen değerlerle karşılaştırılmaktadır. Bu çalışmada analiz edilen Hukuk Fakültesi Kabul Testi 6 (LSAT6; Bock ve Lieberman, 1970) veri seti 1000 adayın beş seçenekli beş maddeye verdiği cevaplardan oluşmaktadır. Çoktan seçmeli maddeler yanlış yanıt için 0 ve doğru yanıt için 1 olarak kodlanmıştır. Bu çalışmada, Bayes kesitiricisi şu anda Mplus'ta 3PL, 4PL ve polytomous modeller gibi diğer MTK modelleri için mevcut olmadığından sadece Rasch, 1PL ve 2PL MTK modellerine odaklanılmıştır. Bu 3 modele ait yetenek ve madde parametresi tahminleri, MLE veya Bayes kestirimi ile elde edilebilir. MLE algoritması, Newton-Raphson (Embretson & Reise, 2000) gibi yinelemeli kestirimlerle olabilirlik fonksiyonunu en üst düzeye çıkaran yetenek ve madde düzeyi tahminlerini bulmaya odaklanır. Bayes kestiricisi ise, önsel dağılımı olabilirlik fonksiyonuna entegre eder. Mplus'ta Bayesci kestirimi probit ölçeğinde ve MLE kestirimi lojistik skalada gerçekleştirilmektedir. Bayesci kestirim yönteminde posterior dağılımın modu, medyanı veya ortalaması, merkezi eğilimin nihai tahmini olarak alınmaktadır. Mplus'ta Bayesci kestirimi şimdilik sayım, sürekli zamanlı yaşam kalım, sansürlü veya nominal veriler için mevcut değildir ayrıca 3PL, kademeli yanıt modeli, kısmi kredi modeli veya genelleştirilmiş kısmi kredi modeli için de Bayesci kestirimi mevcut değildir. Bu sebeple, bu çalışmada Rasch, 1PL ve 2PL modelleri Mplus'taki MLE ve Bayes seçenekleri kullanılarak LSAT6 veri setine uygulanmıştır. Karşılaştırma için Mplus yazılımında uygulandığı şekliyle MLE tahmincisi kullanılmıştır. Mplus ile Bayesci kestirimi bilgilendirici olmayan varsayılan önsellerle uygulanmıştır. Üç modelin Bayesci kestirimi için Mplus sözdizimleri sırasıyla Şekil 1 ila 3'te sunulmuştur. Tüm model parametreleri sonsal (posterior) dağılımdan hesaplanmıştır. LSAT6 verileri daha önce Kim (2003) tarafından MCMC ile analiz edilmiş ve Gelman ve Rubin tanılayıcı bilgilerine dayanarak burn-in (yanma) sayısı olarak 5000 iterasyon (yineleme) belirlenmiştir. Bu çalışmada da, yanma aşaması için aynı sayıda iterasyon kullanılmıştır. Böylece toplam 10000 MCMC yinelemesi gerçekleştirilmiş ve ilk 5000 yineleme burn-

in olarak kullanılmıştır. Her bir parametre için örneklenen değerlerin sonsal ortalamaları nihai tahminler olarak kullanılmıştır. LSAT6 veri setine sahip üç MTK modelinin madde parametrelerinin ML tahminleri de karşılaştırma için Mplus yazılım paketi ile elde edilmiştir. Mplus türevli parametre tahminlerini (yani konum [location] ve eşikler [thresholds]) tipik MTK tahminlerine (yani ayırt edicilik [discrimination] ve güçlük [difficulty] değerlerine) dönüştürmek için Asparouhov ve Muthén (2016, s.6) tarafından sunulan dönüşüm formülleri kullanılmıştır. Forero ve Maydeu-Olivares (2009), D = 1.702 ölçekleme sabitini kullanarak parametre tahminlerini ve standart hataları aynı metriğe (.01 birim içinde) yerleştirmek için MTK modellerinin normal ve lojistik varyantlarını kullanır (Haley, 1952). Bu çalışmada da D sabiti, kullanılan modelin normal ogive parametrelerini lojistik parametrelerin ölçeğine koymak için kullanılmıştır. Tablo 1-3, LSAT veri setine uyan Rasch, 1PL ve 2PL MTK modelleri için beş madde parametresi için MCMC algoritmasını kullanan yaklaşık sonsal (posterior) ortalamaları göstermektedir. Bulgular bölümünde, Rasch, 1PL ve 2PL modelleri için Bayesci kestirimi ile madde parametrelerinin MLE tahminleri karşılaştırılmaktadır. Bayesci kestiriminden elde edilen parametreleri yeniden ölçeklendirmek için Mplus'taki Bayesci kestirimi probit bağlantı fonksiyonunu kullandığından tüm parametreler 1.7 ile çarpılmıştır. Tablo 1-3'te MLE ve Bayesci tahmincileri için Mplus'tan Rasch, 1PL ve 2PL modellerinin madde parametre tahminleri listelenmektedir. Mplus-ML sütunları MLE tahminlerini sunar ve son sütun Bayesci tahminlerini gösterir. Mplus'ta Bayesci kestiriminden alınan madde parametresi tahminleri, MLE ile aynı ölçekte değildir. Bu nedenle, madde güçlüğü ve madde ayırt edicilik parametrelerini logit ölçeğine dönüştürmek için Bayesci tahminlere dönüşüm formülleri uygulanmıştır. Tablo 1-3'te görülebileceği gibi, dönüştürülmüş Bayesci kestirimleri, birinci ve ikinci ondalık basamaklarda MLE tahminlerine göre farklılaşmıştır. Bayesci tahminleri ile ML tahminleri arasındaki Pearson korelasyon değeri .999 civarında bulunmuştur. MLE ve MCMC arasındaki en büyük farklılıklar 2PL MTK modeline ait parametrelerde bulunmuştur.

Sonuç ve Tartışma

İki kategorili MTK modelleri tipik olarak hem Bayesci hem de MLE tahmin algoritmaları ile tahmin edilir. MTK modellerinin Bayesci kestirimi bazen MLE'ye tercih edilir, çünkü MLE'nin örtük değişkenlerin sayısının bir fonksiyonu olan entegrasyon boyutlarının sayısına bağlı olarak yavaş veya engelleyici olabilen sayısal entegrasyona ihtiyacı vardır. İki kategorili MTK modellerinin Bayesci kestirimi kullanarak Mplus'ta nasıl tahmin edileceğini psikometri literatüründe gösteren didaktik bir çalışma yer almamaktadır. Bu çalışmada, Bayesci kestirimi ile iki kategorili MTK modellerinin tahmini için gerekli işlemler adımlar halinde sunulmaktadır. Spesifik olarak, beş maddelik LSAT6 veri seti kullanılarak iki kategorili üç MTK modeli analiz edilmiştir. Bu beş maddenin parametre tahminleri MLE ve Bayesci tahminleri için karşılaştırılmıştır. Sonuçlar, MLE ile Bayesci kestirimlerinden elde edilen madde parametre tahminleri arasında bazı farklılıklar olduğunu göstermiştir. Bu, MLE ve MCMC için karşılaştırılabilir tahmin sonuçları gösteren önceki araştırmaların sonuçlarıyla tutarlıdır (örneğin, Kieftenbeld & Natesan, 2012; Luo, 2018; Paek, Cui, Öztürk Gübeş ve Yang, 2018; Wollack, Bolt, Cohen ve Lee, 2002). Bilgilendirici olmayan (non-informative) önsellerle Bayesci kestirimi, MLE ile asimptotik olarak benzer tahminler vermelidir. Bu çalışmada gösterildiği gibi, Mplus'ta Bayesci kestirimi için varsayılan, bilgilendirici (informative) önseller kolayca belirtilmesine rağmen, bilgilendirici olmayan önseller kullanmaktır. Önsel dağılım, Bayesci tahminin en önemli kısımlarından biridir. Bunun nedeni, önsel dağılımların özellikle küçük örneklem büyüklükleri için sonsal dağılımı önemli ölçüde etkileyebilmesidir (Natesan, Nandakumar, Minka ve Rubright, 2016). Bu çalışmada sadece bilgilendirici olmayan önseller kullanılmıştır. Bununla birlikte, Mplus yazılımındaki tahminlere bilgilendirici veya biraz bilgilendirici önseller eklemek mümkündür. Mplus'a Bayesci kestirim algoritmasının eklenmesi, Mplus'ı iki kategorili MTK modellerinin kestirimi için uygun ve kullanışlı bir yazılım paketi haline getirmiştir.

Appendix. Mplus Plots for 1PL IRT Model with Bayes Estimator

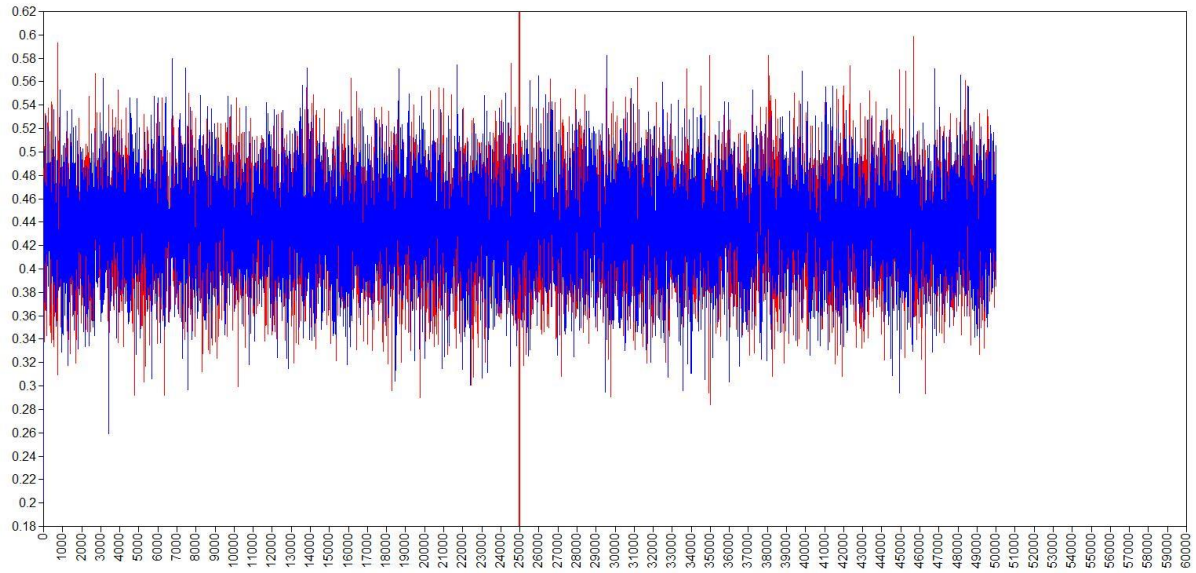


Figure A1. Bayesian posterior parameter trace plots.

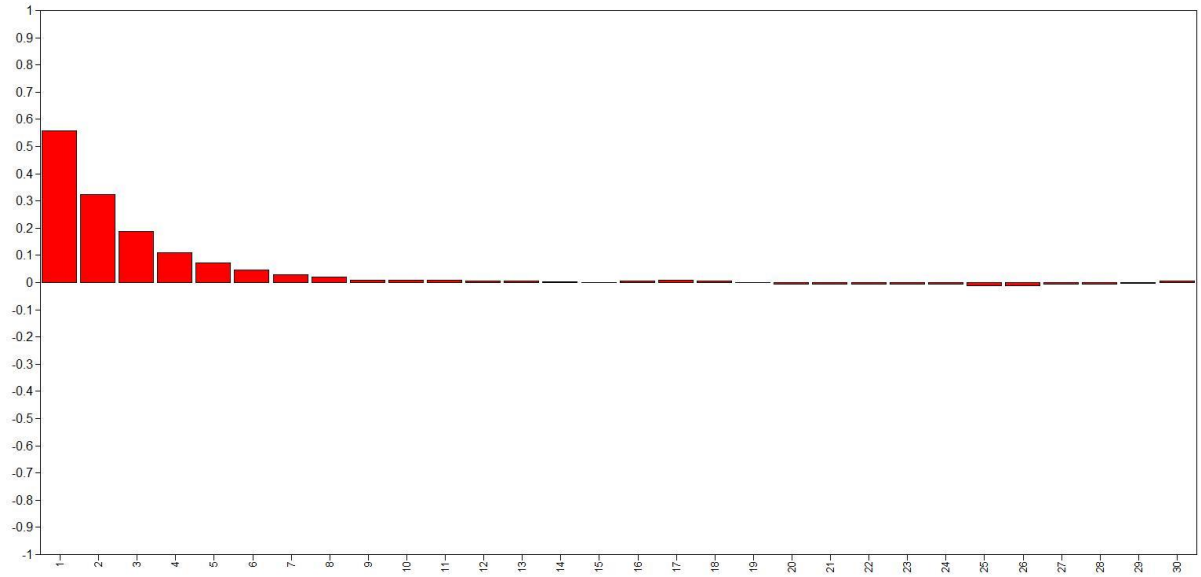


Figure A2. Bayesian autocorrelation plot.

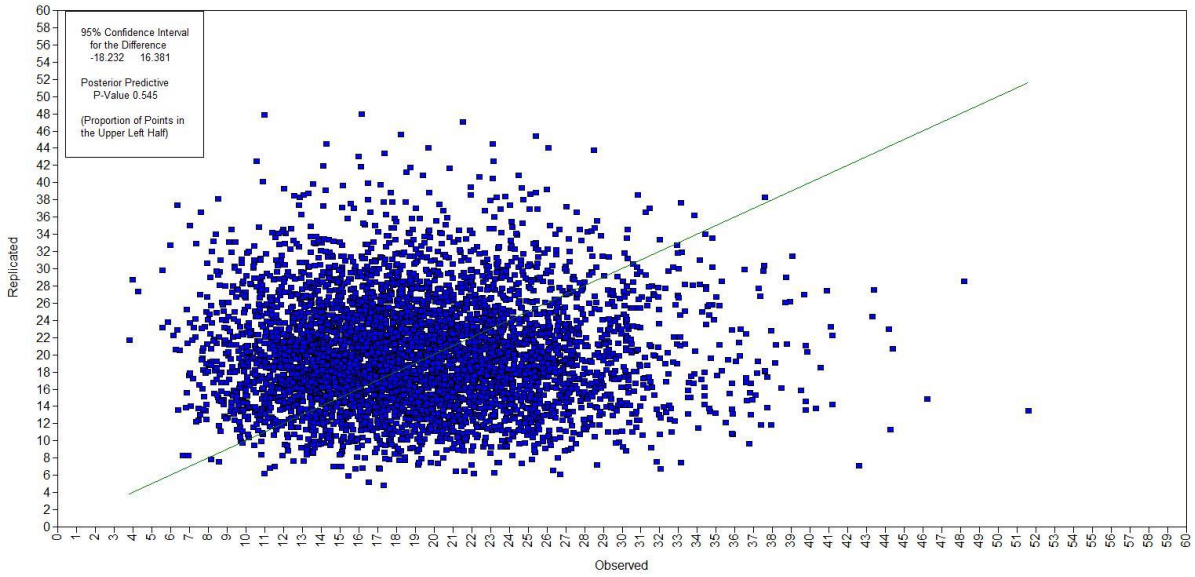


Figure A3. Bayesian posterior predictive checking scatterplots.

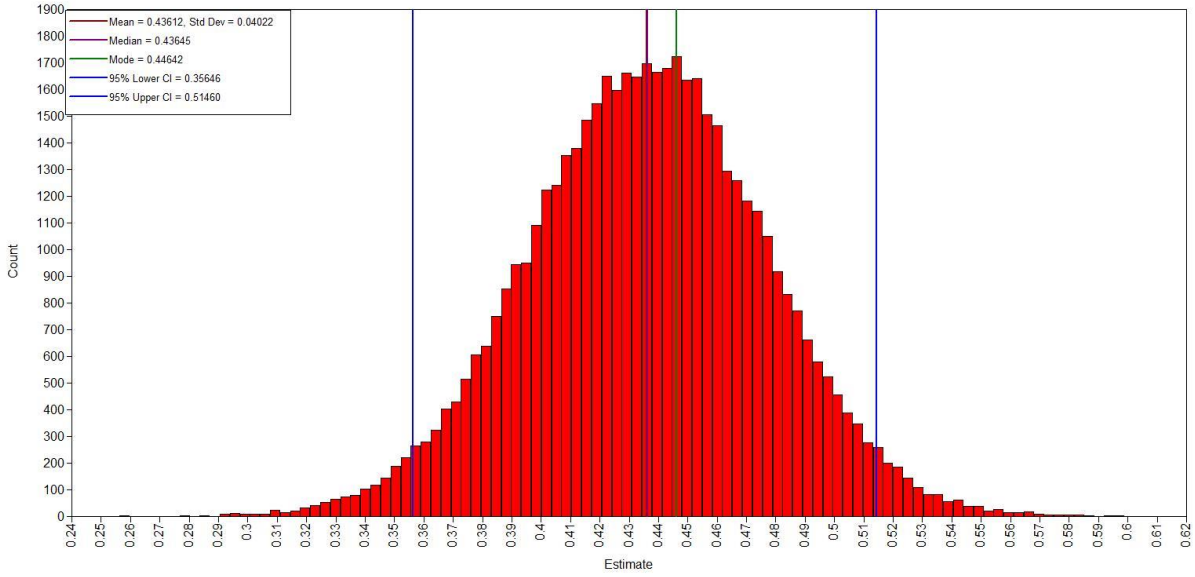


Figure A4. Bayesian posterior parameter distributions.

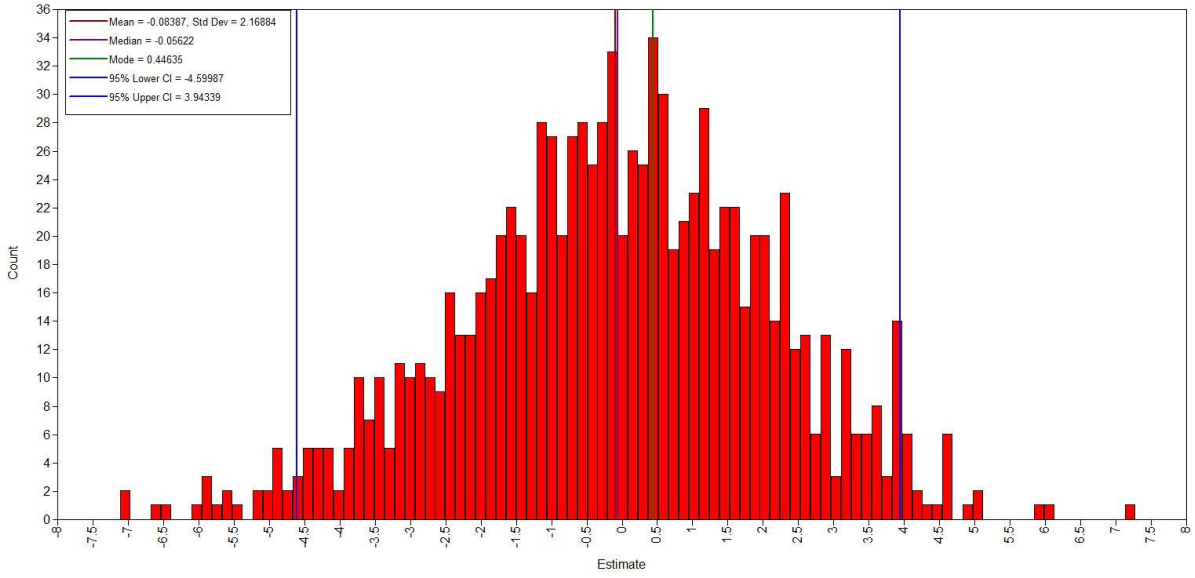


Figure A5. Bayesian prior parameter distributions.

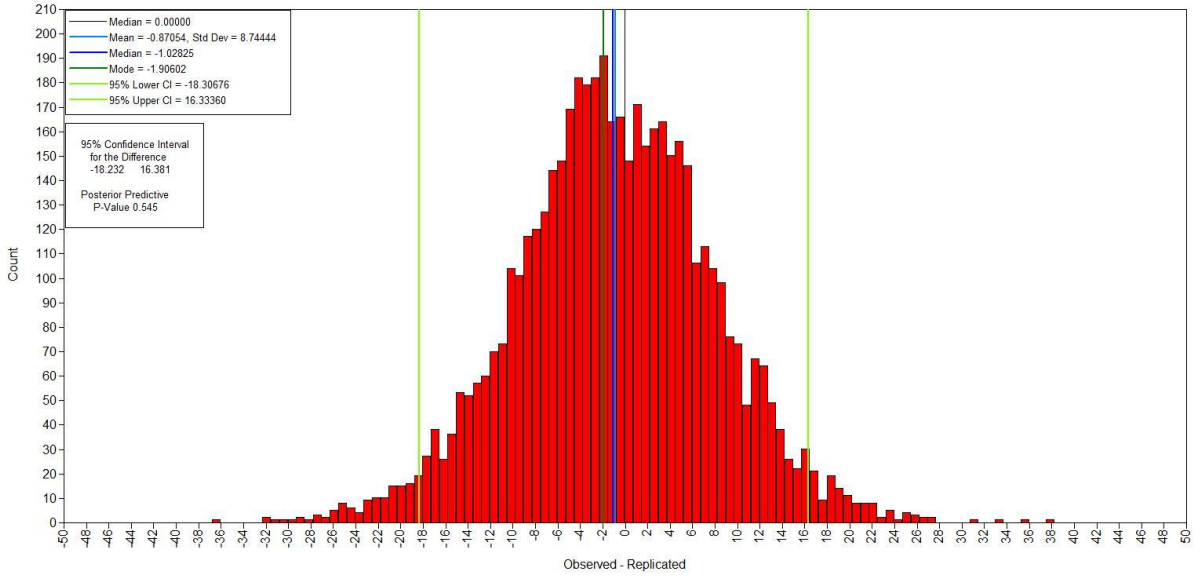


Figure A6. Bayesian posterior predictive checking disctibutions plots.

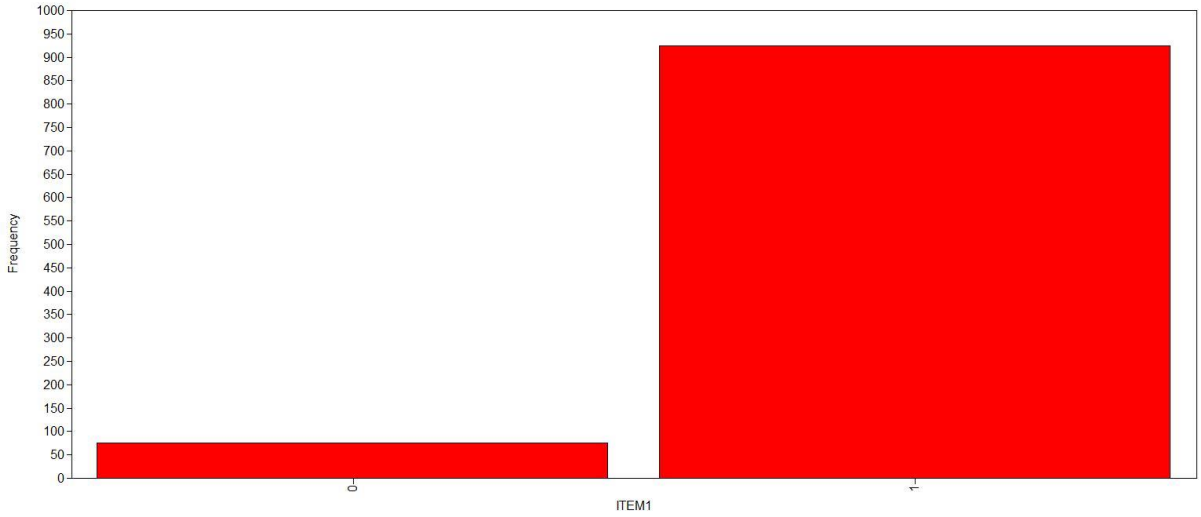


Figure A7. Histogram of sample values.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(3); 283-301

Analysis of Differential Item Functioning of PISA 2015 Mathematics Subtest Subject to Gender and Statistical Regions *

Mustafa ÇELİK ** Yeşim ÖZER ÖZKAN ***

Abstract
It is accepted that the presence of Differential Item Functioning (DIF) in large scale examinations may be an indication of bias. The aim of the present study was to analyze whether dichotomous (1-0) items in the PISA 2015 mathematics subtest exhibit DIF with regard to gender and statistical regions. The study was carried out using the data of 2409 students who took part in PISA 2015 and answered all mathematics questions. Rasch model was used via Winsteps software to determine whether the items exhibit DIF or not. Confirmatory Factor Analysis (CFA) was applied to the sixty-three mathematics questions in the clusters. The modification indices and goodness of fit values were examined for CFA and a total of eight items that disrupt the model structure were excluded from the test. DIF analyses were carried out for the 55 items that were observed to be one-dimensional. The analysis results based on gender indicated that five items exhibit DIF. Two of these items exhibit DIF in favor of girls, while three in favor of men. Statistically significant DIF findings were observed in all items when the analyses results based on statistical regional units were analyzed. While at least 10 DIF cases were observed as a result of the binary territory comparison on an item basis, maximum 38 DIF cases were observed. Minimum DIF was observed in item Q4 as a result of regional comparisons, whereas maximum DIF was observed in items Q47 and Q50.

Key Words: Differential item functioning, Rasch model, bias, statistical regions, PISA.

INTRODUCTION

Structuring of labor via qualified education is directly related to education quality and policy. Continuous advancement of scientific developments by way of innovations increases the importance of cooperative education quality and education policies. Measurement and evaluation tools are used for determining the personal qualifications of individuals who undergo a certain education. Individual outputs determined subject to the implementation purpose provide insight into the competence of the individuals (Baykul, 2000). Large scale examinations are conducted for the global evaluation of educated individuals. Large scale examinations in Turkey are conducted by the Ministry of National Education (MoNE), General Directorate of Measurement, Evaluation and Examination Services and Student Selection and Placement Center.

Large scale examinations are carried out worldwide such as Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS) are carried out which enable the comparison of many education-related outputs.

PISA, organized by the Organization for Economic Cooperation and Development (OECD) has been applied in Turkey once every three years since 2000. The fundamental knowledge and skills of the students in the fields of science, mathematics and reading are assessed as part of the PISA project for a 15-year-old group of students (MoNE, 2013). PISA goes beyond assessing whether students can

* This study was a part of thesis was produced from the first author's master thesis.
** Teacher, Ministry of National Education, Gaziantep-Turkey, mstfclk47@gmail.com, ORCID ID: 0000-0002-9501-8726
*** Assoc. Prof., Gaziantep University, Faculty of Education, Gaziantep-Turkey, e-posta:yozer80@gmail.com, ORCID ID:0000-0002-7712-658X

To cite this article:
Çelik, M. & Özer-Özkan, Y.(2020). *Analysis of Differential Item Functioning of PISA 2015 Mathematics Subtest Subject to Gender and Statistical Regions, 11*(3), 283-301. doi: 10.21031/epod.715020.

Received: 05.04.2020
Accepted: 10.07.2020

reproduce what they have learned in school, thus focusing on determining their ability to apply their knowledge in real life, solve problems in novel situations as well as their abilities to make use of skills such as extrapolation and reasoning (MoNE, 2010). The data acquired from these surveys contribute to the interpretation of cognitive data. Turkey has taken part in school and student surveys within the scope of PISA 2015 (MoNE, 2015).

Individuals who take part in international large scale examinations vary with regard to characteristics such as ethnic origins, language, culture, etc. The balance of presence in life of female-male individuals differs among different societies, especially within the context of gender. These differences make it difficult to adapt the tests into different languages and cultures in intercultural studies (Van de Vijver & Tanzer, 2004). Large scale examinations should be prepared without allowing for any inequalities by taking the aforementioned circumstances under control. The prepared test items should not provide any advantage or disadvantage to any group (Öğretmen, 1995).

The validity and reliability of the test scores of the individuals for the characteristics to be compared with regard to desired ability or success may have an impact on the accuracy of the decisions made based on these data (Gierl, 2000). The fact that large scale examinations used as a resource for important decisions are free of errors brings forth the validity of the test or test items and thus, the presence of bias observed as systematic error. Item and test bias is among the threats against meeting the validity requirement (Clauser & Mazor, 1998). In this regard, unbiasedness is considered as a criterion in order for a test or test item to meet the validity requirement (Camili & Shepard, 1994). Characteristics of measurement such as the ability level, item discrimination, item difficulty, distribution, reliability vary subject to the group (Özer Özkan, 2012). In this regard, it is expected that the psychometric characteristics of the measurement tool are the same for all responders (Kıbrıslıoğlu Uysal & Atalay Kabasakal, 2017).

Bias is accepted as the presence of systematic error in the test items (Osterlind, 1983). It can be indicated that the value of the variables is systematically low or high in case of bias in the test item (Çıkrıkçı Demirtaşlı & Uluştaş, 2015). It is expected that the individuals will have the same probability of responding to the items correctly if the test or test items have the same construct validity for all individuals in the group (Camili & Shepard, 1994). Determining the test or test item that exhibits DIF is important for validity. In this case, the validity of the study carried out will be at risk in case studies are not carried out for determining the biased items in large scale tests.

It is observed that studies on DIF are carried out frequently in different countries in order to test the validity of international large scale examinations. DIF and bias studies based on gender, culture and language are observed frequently in literature. It has been observed that DIF and bias studies have been carried out for international or national scale examinations subject to *gender* (Acar, 2011; Alkaline, 2014; Amour, AL-Gadarene Alomar, & AL Ruairi, 2015; Arıffin, Idris, & Ihsak, 2010; Ateşok Deveci, 2008; Bakan Kalaycıoğlu, 2008; Bakan Kalaycıoğlu & Kelecioğlu, 2011; Bekci, 2007; Berberoğlu, 1995; Birjandi & Amini, 2007; Doolite & Cleary, 1987; Gamer & Engelhard, 1999; Hanna, 1986; Haris & Cartlon, 1993; Karakaya, 2012; Karakaya & Cult, 2012; Ken, Sunbelt, & Omar, 2013;Kıbrıslıoğlu & Atalay Kabasakal, 2017; Kurnaz, 2006;Latifi, Bulut, Gierl, Christie, & Jeeva, 2016; Le, 1999,2009; Lyons-Thomas, Sandilands, & Ercikan, 2014; Öğretmen & Doğan, 2004; Özer Özkan & Fincan, 2017; Satıcı & Özer Özkan, 2016; Sunna, 2012;Şenferah, 2015; Taylor & Lee, 2012;Turkman, 2014;Ultras, 2012;Yilin & Tavşancıl, 2015;Yurdugül & Aşkar, 2004; Zenisky, Hambleton, & Robin, 2004; Zwick & Ercikan, 1989); *school type* (Bakan Kalaycıoğlu, 2008; Bekci, 2007; Karakaya & Kutlu, 2012; Şenferah, 2015); *regions and cultures* (Ercikan & Kim, 2009; Gök, Atalay Kabasakal, & Kelecioğlu, 2014; Özmen, 2014; Ulutaş, 2012; Yurdugül & Aşkar, 2004; Zwick & Ercikan, 1989).

It is observed that bias studies in Turkey have been carried out since the early 1990s on data for national tests of Secondary School Institutions Student Selection and Placement Test, Student Selection Test, Level Determination Exam and Transition from Primary to Secondary Education. There has been an increase in DIF and bias studies on PISA, TIMSS and PIRLS following the increase in the popularity of international large scale examinations after the 2000s. Bias studies have been included in the PISA 2015 technical report published by the OECD and information has been presented on how the biased items are controlled (OECD, 2015). It is observed that there is no mention of any DIF or bias towards gender

and statistical regions in both the PISA 2015 technical report published by OECD and the PISA 2015 National Report published by MoNE General Directorate of Measurement, Evaluation and Examination Services and Student Selection. Moreover, the fact that the impacts of gender and culture yield results that cannot be ignored in DIF and bias studies in literature increase the importance of the study. In this regard, it is expected that the present study aiming to examine the DIF of PISA 2015 mathematics subtest subject to gender and statistical regions will contribute to the related studies in the literature.

The Aim of the Study

The aim of the study is to determine whether the dichotomous (1-0) items in the PISA 2015 application mathematics subtest exhibit DIF subject to gender and territory or not, according to the Item Response Theory (IRT) via the Rasch Model. Answers to the following questions were sought within the framework of this aim:

1. Do the binary scored items in the PISA 2015 mathematics literacy subtest exhibit DIF based on the analyses via the Rasch method?
2. Do the binary scored items in the PISA 2015 mathematics literacy subtest exhibit DIF with regard to the statistical regions in Turkey based on the analyses via the Rasch method?

METHOD

The study examines whether the dichotomous (1-0) items in the PISA 2015 mathematics literacy subtest exhibit DIF or not with regard to gender and regions. The study is designed based on considering the current situation from different perspectives, defining and comparing the relations between them and expressing them in a holistic and circumspect way (Büyüköztürk et al., 2013). In this regard, this is a descriptive study in the survey model.

Population and Sample

The 15-year-old student population who can take part in the PISA 2015 Turkey application was determined as 925.366. A total of 187 schools from 61 provinces representing the 12 regions in the Turkey Nomenclature of Territorial Units for Statistics. (NUTS) 1st level took part in the PISA 2015 application. Turkey's NUTS classifications are officially termed statistical regions. Therefore, in this study, the term statistical region is used. School sample groups were determined during the PISA 2015 study via a stratified random sampling method, while the students to take part in the application were selected via random sampling method (MoNE, 2015). The present study was carried out using the data of 2409 students who responded to the PISA 2015 mathematics subtest. The regionals DIF findings of the items in the PISA 2015 mathematics subtest were examined based on the NUTS-1 classification presented in Table 1. Table 1 presents the distribution subject to gender and statistical regions for the students who took the mathematics subtest in 2015.

Table 1. Sample Group Distribution for the Students Subject to Gender and Statistical Regions

	Female		Male		Total	
Territory	f	%	f	%	f	%
İstanbul	221	18.45%	210	17.34%	431	17.89%
Western Marmara	53	4.42%	46	3.80%	99	4.11%
Aegean	154	12.85%	130	10.73%	284	11.79%
Eastern Marmara	102	8.51%	107	8.84%	209	8.68%
Western Anatolia	112	9.35%	113	9.33%	225	9.34%
Mediterranean	160	13.36%	182	15.03%	342	14.20%
Central Anatolia	61	5.09%	70	5.78%	131	5.44%
Western Black Sea	54	4.51%	77	6.36%	131	5.44%
Eastern Black Sea	35	2.92%	45	3.72%	80	3.32%
Northeastern Anatolia	44	3.67%	36	2.97%	80	3.32%
Central Eastern Anatolia	60	5.01%	59	4.87%	119	4.94%
Southeastern Anatolia	142	11.85%	136	11.23%	278	11.54%
Total	1198	100.00%	1211	100.00%	2409	100.00%

It can be observed when Table 1 is examined that the students who responded to the PISA 2015 mathematics subtest have similar gender distributions. Whereas, it can be indicated that the distribution according to the statistical regions of the students who responded to the PISA 2015 is similar within the scope of NUTS.

Data Collection Tools

Procedure

PISA 2015 cognitive test results published at the OECD official website were used in the study. PISA 2015 was applied in Turkey by way of a computer-based assessment method instead of as a pencil-paper test. The items in the mathematics subtest were included in 36 of the 66 booklets prepared for the implementation of this method. PISA 2015 cognitive test data were downloaded after which the data related to the Turkey mathematics subtest were sorted out. The data for 2409 Turkish students who responded to all of the items in the mathematics literacy subtest were used (OECD, 2015).

Data Analysis

The items in PISA 2015 mathematics subtest were classified by the OECD into 6 + 1 (equivalent form) different clusters comprised of 11 or 12 questions. Each booklet used for implementation in PISA 2015 was prepared by including the question group that makes up one or two mathematics clusters. Annex 1 presents the additional data regarding the mathematics subtest items included in each booklet implemented in PISA 2015.

The forms used in PISA 2015 implementation were prepared consecutively, placing to the booklets each of the six different clusters. The students answered the items in the mathematics cluster in the booklet. Each student did not take the complete mathematics subtest. Instead, they answered the mathematics items in one or two of the six mathematics clusters determined by the OECD. The following sample distribution method was used for selecting the items in the forms prepared for PISA 2015 implementation. Table 2 presents the sample distribution method for the number of items included in the mathematics subtest clusters.

Table 2. Mathematics Subtest Items Sample Cluster Distribution

	Cluster01	Cluster02	Cluster03	Cluster04	Cluster05	Cluster06a
Forms	11 Items	12 Items	12 Items	11 Items	12 Items	11 Items
Form33	x	x				
Form34		x	x			
Form35			x	x		
Form36				x	x	
Form37					x	x
Form38	x					x
Form39		x			x	

While the items in Cluster01 and Cluster02 in Form33 were observed when Table 2 is examined, the items in Cluster05 and Cluster06 were answered in Form37. Two equivalent forms were prepared for the final cluster in the PISA 2015 mathematics subtest. One of the equivalent forms was applied in Turkey. The equivalent form was named as "Cluster06a" throughout the study. The data regarding the Turkey mathematics subtest declared by the OECD were extracted by taking the data related to gender, territory, form number and question responses. Six multi-scored items were excluded from the data cluster in addition to one item excluded by the OECD since the study was going to be carried out using the dichotomous (1-0) items.

Six copies of the data file were prepared since the responses to each of the mathematics subtest items included in each cluster will be analyzed separately. Each data file was renamed as such (e.g.; Cluster01, Cluster02, etc.), after which they were cleaned up so as to include only the responses to the items in that cluster. Hence, each file was prepared to include gender, education territory and the responses to the items in that cluster. One copy of each cluster file was made in order to ensure that the data are in compliance with the Winsteps 3.80.1 software. Microsoft Office Excel software was used to arrange one of the files to include gender and item responses and the other file to include regions and item responses. As a result, 12 data files were prepared by sorting the PISA 2015 mathematics subtest items for analysis in six clusters and two variables. Winsteps was run and each Excel file was transformed into a text document for analysis. The proper syntax was written for the file transformed into a text document in accordance with the Winsteps software. The related package software was used in the study for the analysis of the assumptions of IRT and for data sorting. The normality graph and the skewness-kurtosis coefficients were examined for the normal distribution assumption of the data of each cluster. The skewness and kurtosis coefficients of each cluster were determined to be between +1 and -1. CFA was applied on the PISA 2015 mathematics subtest for the unidimensionality analysis on a cluster basis. CFA was used to examine RMSEA, GFI, NFI, RFI, IFI, AGFI, CFI and SRMR with regard to validity and goodness of fit values. Two items in Cluster02, three items in Cluster04, one item in Cluster05 and two items in Cluster06a were excluded from the study since they did not meet the unidimensionality assumption. It was observed that the items of each cluster included in the study meet the normality, unidimensionality and local independence assumptions. Finally, it was accepted that the items in the PISA 2015 mathematics subtest are structured in accordance with the Rasch model, according to IRT.

RESULTS

Gender-Related DIF Findings for the Mathematics Subtest Items

This section examines the gender-related DIF values for the dichotomous (1-0) items included in six clusters in PISA 2015 mathematics subtest. DIF measurement value regarding the responses of female and male students to the nine items in Cluster01 of the mathematics subtest subject to the item codes,

the contrast value between the DIF measurement values and t values were examined with the findings presented in Table 3.

Table 3. DIF Values of Items in Cluster01 Subject to Gender

Item	Focus Group	DIF Measurement	Reference Group	DIF Measurement	DIF Contrast	t value
Q1		-1.94		-1.82	-.12	-.63
Q2		-1.12		-1.37	.26	1.41
Q3		-1.03		-1.34	.31	1.69
Q4		-.33		-.24	-.09	-.45
Q5	Female	.37	Male	.56	-.19	-.88
Q6		.11		.19	-.09	-.43
Q7		2.25		1.64	.60	1.91
Q8		1.13		1.76	-.62	-2.35
Q9		.59		.71	-.12	-.56

It was observed when Table 3 was examined that item Q7 exhibits DIF (.60) in favor of males, whereas item Q8 exhibits DIF (-.62) in favor of females. Figure 1presents the change in DIF for the items in Cluster01 subject to gender.

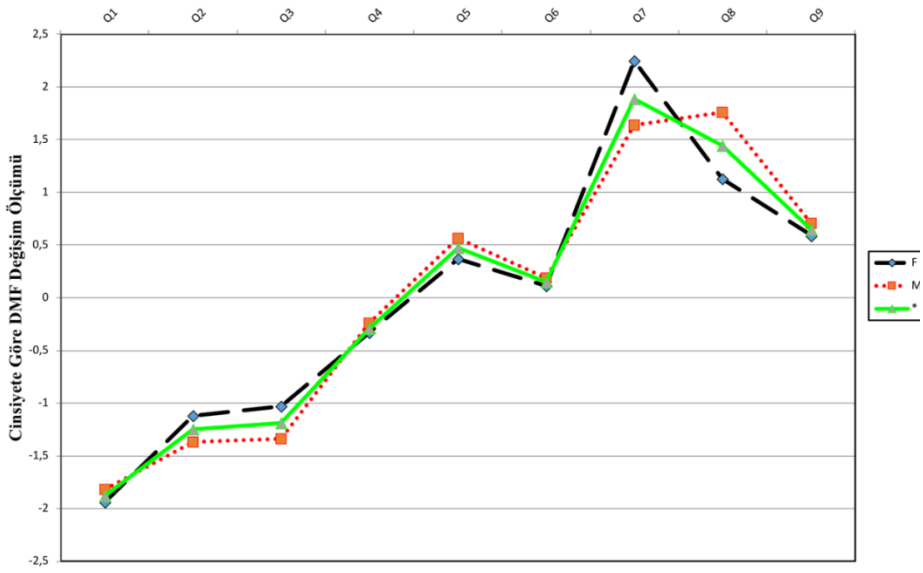


Figure 1. DIF Change Graph Subject to Gender for the Nine Items in Cluster01

It can be understood from the DIF change graph in Figure 1 that the items with contrast values outside the range of .5 and -.5 logit exhibit DIF at a statistically significant level. It can be observed that items Q7 and Q8 have the highest divergence from the mean value for the female and male students. In conclusion, it can be indicated when the DIF values of the nine items in Cluster01 subject to gender are examined that items Q7 and Q8 exhibit DIF at a statistically significant level. Table 4 presents the DIF value subject to gender for the items in Cluster02 of the mathematics subtest.

Table 4. DIF Values Subject to Gender for the Items in Cluster02

Item	Focus Group	DIF Measurement	Reference Group	DIF Measurement	DIF Contrast	t value
Q10	Female	-.03	Male	.44	-.47	-2.49
Q11		.83		.68	.15	.74
Q12		-.64		-.87	.22	1.25
Q13		-2.80		-2.94	.15	.62
Q14		2.70		2.05	.66	2.11
Q16		.31		.64	-.33	-1.71
Q18		-.31		.06	-.36	-2.01
Q19		.21		-.22	.43	2.35

It is observed in Table 4 that the item Q14 in Cluster02 (.66) exhibits DIF in favor of males. Figure 2 presents the DIF change graph of items in Cluster02 subject to gender.

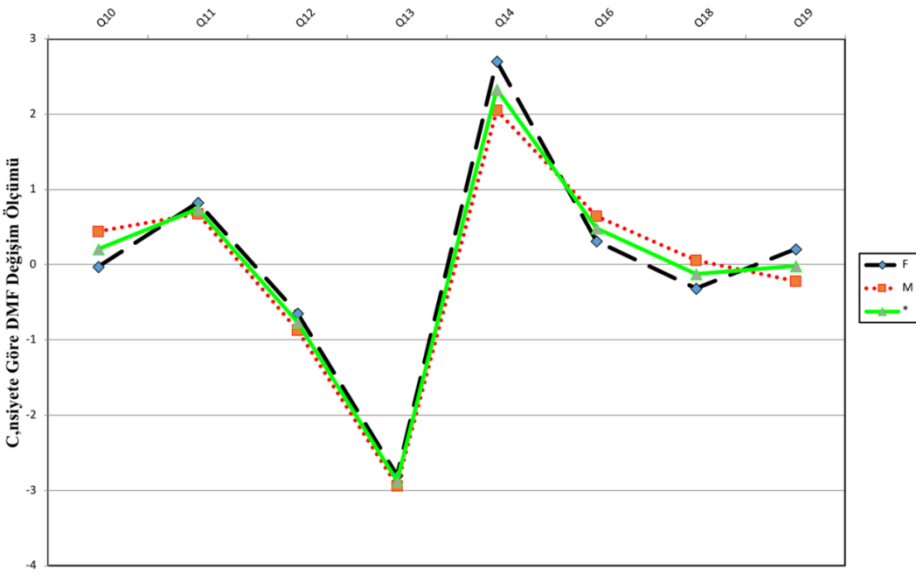


Figure 2. DIF Change Graph of the Eight Items in Cluster02 Subject to Gender

It can be stated when the DIF values of the eight items in Cluster02 are examined that the Q14 coded item exhibits DIF at a statistically significant level. Table 5 presents the DIF value of the items in the mathematics subtest Cluster03 subject to gender.

Table 5. DIF Values Subject to Gender for the Items in Cluster03

Item	Focus Group	DIF Measurement	Reference Group	DIF Measurement	DIF Contrast	t value
Q20		-.80		-.85	.05	.26
Q21		.37		-.09	.46	2.43
Q22		.29		.13	.17	.87
Q23		.35		.35	.00	.00
Q24		-1.62		-1.36	-.26	-1.39
Q25	Female	4.07	Male	4.56	-.49	-.80
Q26		-.65		-.78	.14	.74
Q27		-.77		-.59	-.18	-.97
Q28		1.15		1.25	-.10	-.44
Q29		1.18		1.56	-.38	-1.64
Q30		-3.86		-3.86	.00	.00

The DIF contrast values of the 11 items in Cluster03 have been calculated between 0.5 and -0.5 logit in Table 5. Thus, it can be stated that the items in Cluster03 do not work for or against any group. Figure 3 presents the change in DIF graph for the items in Cluster03 subject to gender.

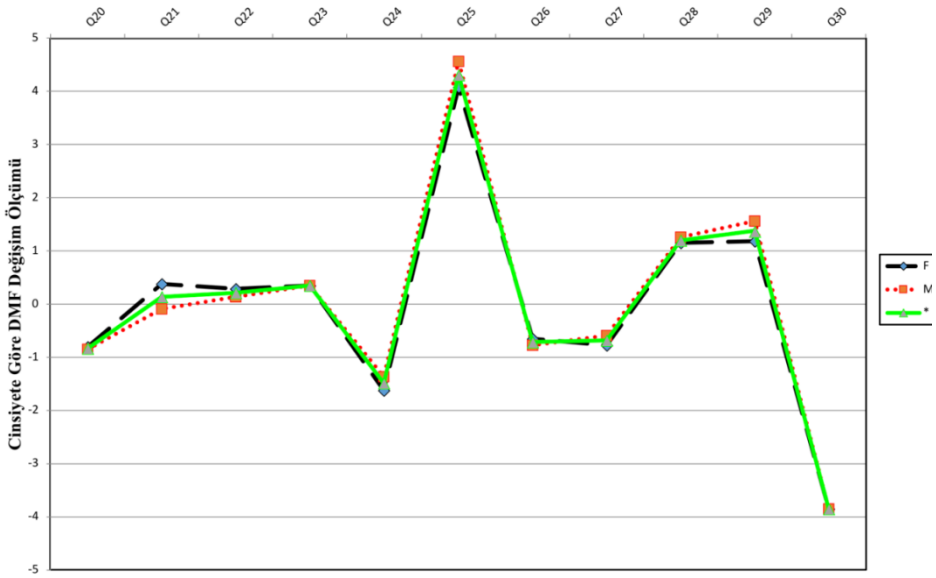


Figure 3. DIF Change Graph of the 11 Items in Cluster03 Subject to Gender

It can be stated when the change in DIF subject to the gender of the 11 items in Cluster03 presented in Figure 3 is examined that there are no items that exhibit DIF at a statistically significant level. Table 6 presents the DIF values subject to the gender of the items in Cluster04 of the mathematics subtest.

Table 6. DIF Values of the Items in Cluster04 Subject to Gender

Item	Focus Group	DIF Measurement	Reference Group	DIF Measurement	DIF Contrast	t Value
Q31		-2.00		-2.00	.00	.00
Q33		-1.29		-1.18	-.11	-.63
Q34		.34		.54	-.20	-1.04
Q35	Female	-.89	Male	-1.08	.19	1.09
Q37		4.34		4.67	-.34	-.51
Q38		.01		-.07	.08	.43
Q39		-.74		-.68	-.06	-.37
Q40		.06		-.07	.13	.70

It can be stated when the DIF contrast values subject to the gender of the eight items in Cluster04 presented in Table 6 are examined that the DIF contrast value calculated between 0.5 and -0.5 logit and that the items do not exhibit DIF subject to gender.

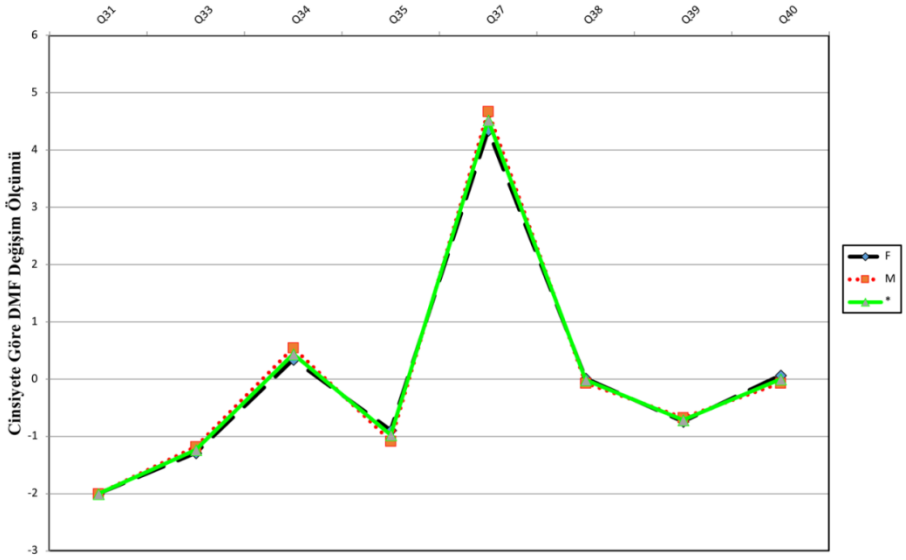


Figure 4. DIF Change Graph Subject to Gender for the Eight Items in Cluster04

It can be understood when Figure 4 is examined that there are no items with contrast values outside the boundaries of .5 and -.5 logit, or in other words, that the items in Cluster04 do not exhibit DIF at a statistically significant level. Table 7 presents the DIF values subject to the gender of the items in Cluster05 of the mathematics subtest.

Table 7. DIF Values Subject to Gender of the Items in Cluster05

Item	Focus Group	DIF Measurement	Reference Group	DIF Measurement	DIF Contrast	t Value
Q42	Female	-1.75	Male	-1.92	.17	.96
Q44		1.54		1.32	.22	.75
Q45		-1.12		-1.17	.05	.30
Q46		.18		.50	-.32	-1.53
Q47		2.26		2.09	.17	.44
Q48		-1.26		-1.26	.00	.00
Q49		-.83		-.96	.13	.77
Q50		3.37		3.13	.23	.39
Q51		-1.72		-1.55	-.17	-1.02
Q52		-.43		-.38	-.06	-.30

It was observed when Table 7 was examined that the items in Cluster05 do not exhibit DIF subject to gender. Figure 5 presents the DIF change graph subject to the gender of the items in Cluster05.

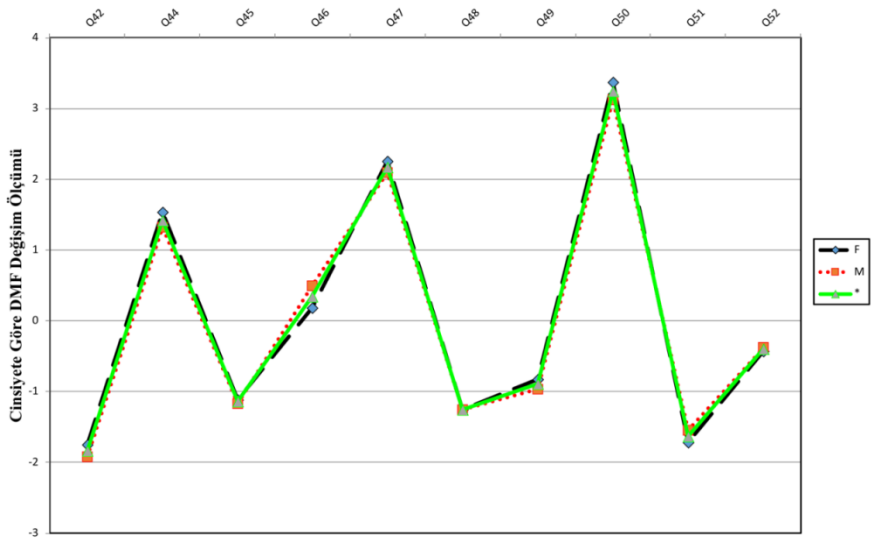


Figure 5. DIF Change Graph Subject to Gender of the 10 Items in Cluster05

It is understood when Figure 5 is examined that the items in Cluster05 do not display DIF at a statistically significant level. The DIF values subject to the gender of the items in Cluster06a of the mathematics subtest are presented in Table 8.

Table 8. DIF Values Subject to Gender of the Items in Cluster06a

Item	Focus Group	DIF Measurement	Reference Group	DIF Measurement	DIF Contrast	t Value
Q53		-1.74		-2.47	.73	3.82
Q54		.69		.27	.42	1.78
Q56		-.12		-.27	.16	.76
Q57		-1.60		-1.12	-.48	-2.55
Q58	Female	.48	Male	1.23	-.75	-2.94
Q59		2.85		2.82	.03	.06
Q60		-1.37		-1.45	.08	.42
Q62		.46		.59	-.13	-.55
Q63		.27		.61	-.35	-1.48

It is observed when Table 8 is examined that the item Q58 in Cluster06a operates in favor of females based on its DIF contrast (-.75) value, whereas item Q53 (.73) operates in favor of males. Figure 6 presents the DIF change graph subject to the gender of the items in Cluster06a.

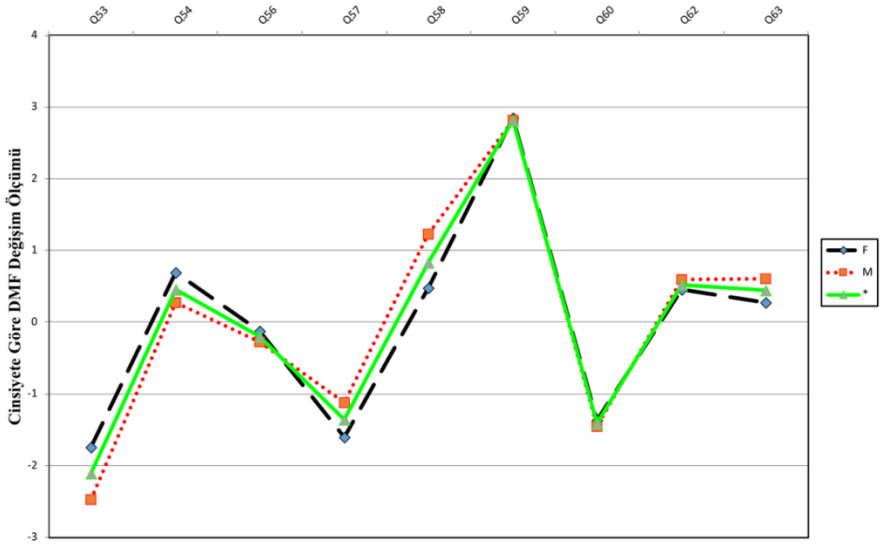


Figure 6. DIF Change Graph Subject to Gender of the Nine Items in Cluster06a

It can be observed when Figure 6 is examined that the female and male students stray away from the mean at the maximum level in items Q53 and Q58. In conclusion, it can be stated that items Q53 and Q58 exhibit DIF at a statistically significant level.

Statistical Regions Related DIF Findings of the Items in the Mathematics Subtest

This section focuses on the DIF values of the dichotomous (1-0) items in the PISA 2015 mathematics subtest. The DIF change graphs subject to statistical regions are presented for the items in the six clusters of the mathematics subtest. Figure 7 presents the change in DIF graph subject to the statistical regions for the nine items in Cluster01.

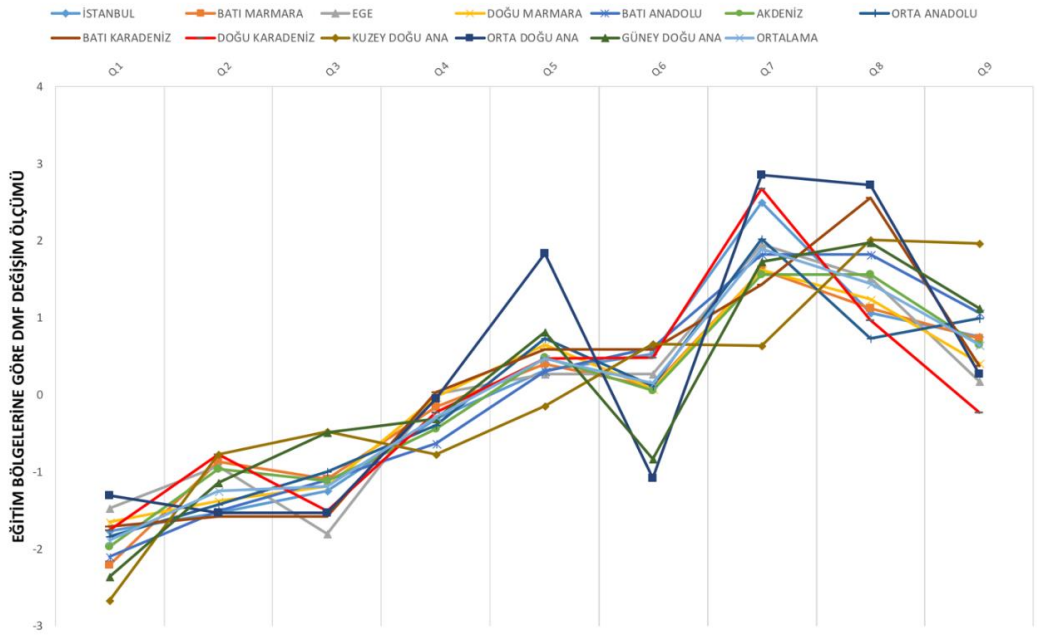


Figure 7. Change in DIF Subject to Statistical Regions for the Nine Items in Cluster01

In Figure 7, with the blue dotted line for İstanbul, Western Marmara with the orange checkered line for, Aegean with the grey triangle line, Eastern Marmara with the yellow crossed line, Western Anatolia with the blue starred line, the Mediterranean with the green dotted line, Central Anatolia with the navy blue perpendicular line, Western Black Sea with the brown line, Eastern Black Sea with the red line, Northeastern Anatolia with the brown checkered line, Central Eastern Anatolia with the navy blue squared line, Southeastern Anatolia with the green triangled line and the educational territory mean value with the blue crossed line. It is observed that items Q3, Q5, Q6, Q7, Q8 and Q9 exhibit the biggest change from among the items in Figure 1. It is seen from the DIF change graph of the items in Cluster01 subject to statistical regions that the maximum divergence from the mean value is in the Eastern Black Sea, Northeastern Anatolia, Central Eastern Anatolia and Southeastern Anatolia region. Figure 8 presents the change in DIF of the eight items in Cluster02 subject to statistical regions.

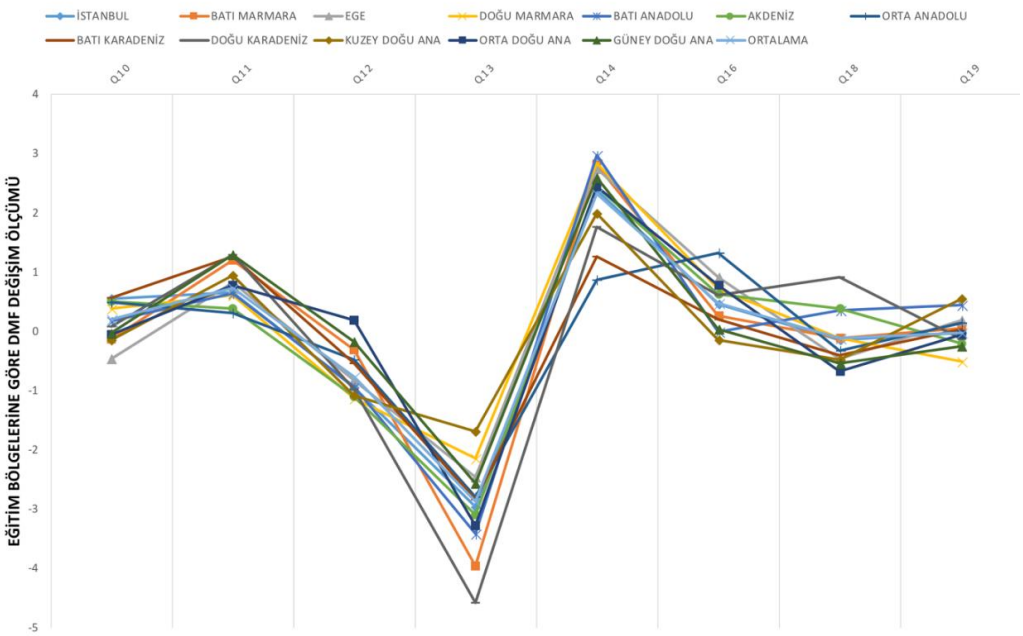


Figure 8. Change in DIF Subject to Statistical Regions for the Eight Items in Cluster02

Figure 8 illustrates that Q13, Q14, Q16 and Q18 items exhibit the highest rate of change. As can be seen from the change in DIF of the items in Cluster02 subject to statistical regions, Northeastern Anatolia, Central Eastern Anatolia and Southeastern Anatolia are the regions that have diverged the most from the mean value. Figure 3 shows the DIF change graph of the 11 items in Cluster03 subject to statistical regions.

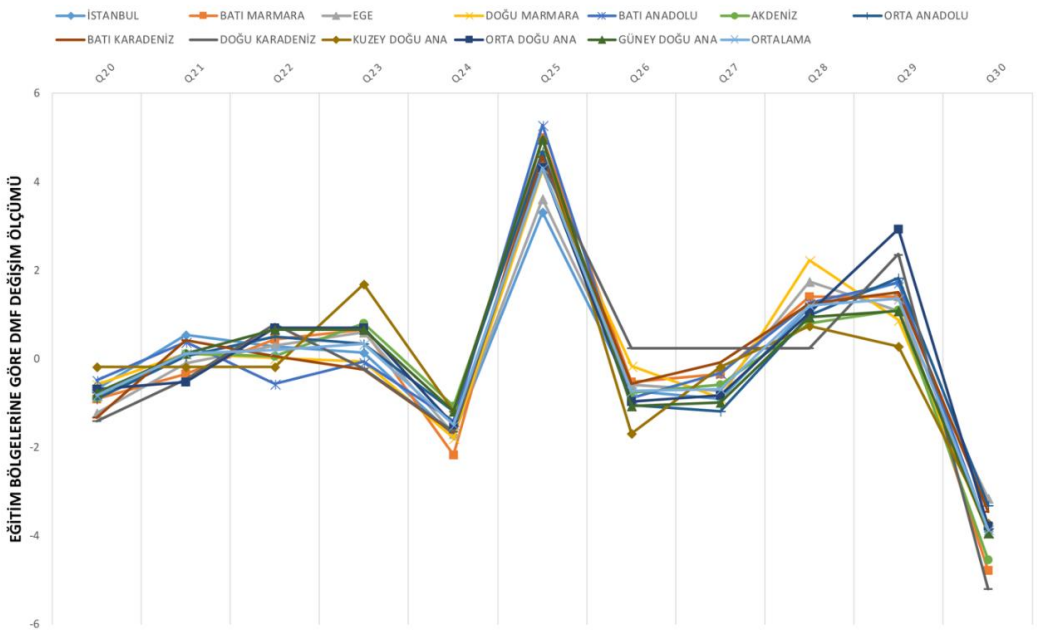


Figure 9. Change in DIF Subject to Statistical Regions for the Eleven Items in Cluster03

Figure 9 illustrates that Q23, Q25, Q26 and Q28 items exhibit the highest rate of change. As can be seen from the change in DIF of the items in Cluster03 subject to statistical regions, Eastern Marmara, Northeastern Anatolia, Central Eastern Anatolia and Southeastern Anatolia are the regions that have diverged the most from the mean value. Figure 10 shows the DIF change graph of the eight items in Cluster04 subject to statistical regions.

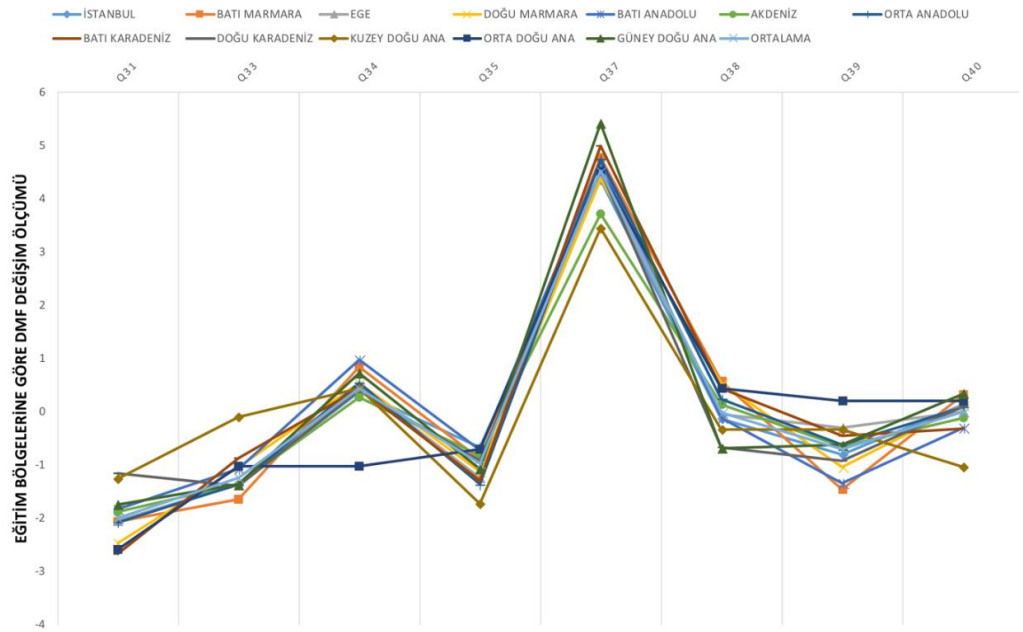


Figure 10. Change in DIF Subject to Statistical Regions for the Eight Items in Cluster04

It can be observed when Figure 10 is examined that Q33, Q34 and Q39 are the items from among the eight items of Cluster04, which display the highest rate of change subject to statistical regions. It is presented in the DIF change graph subject to regions for the items in Cluster04 that Northeastern

Anatolia and Central Eastern Anatolia are the regions that have diverged the most from the mean value. The results indicate that Northeastern Anatolia territory for items Q33 and Q40 and the Central Eastern Anatolia regions for items Q34 and Q39 have diverged from the mean value at a significant level. Figure 11 presents the DIF change subject to regions for the 10 items in Cluster05.

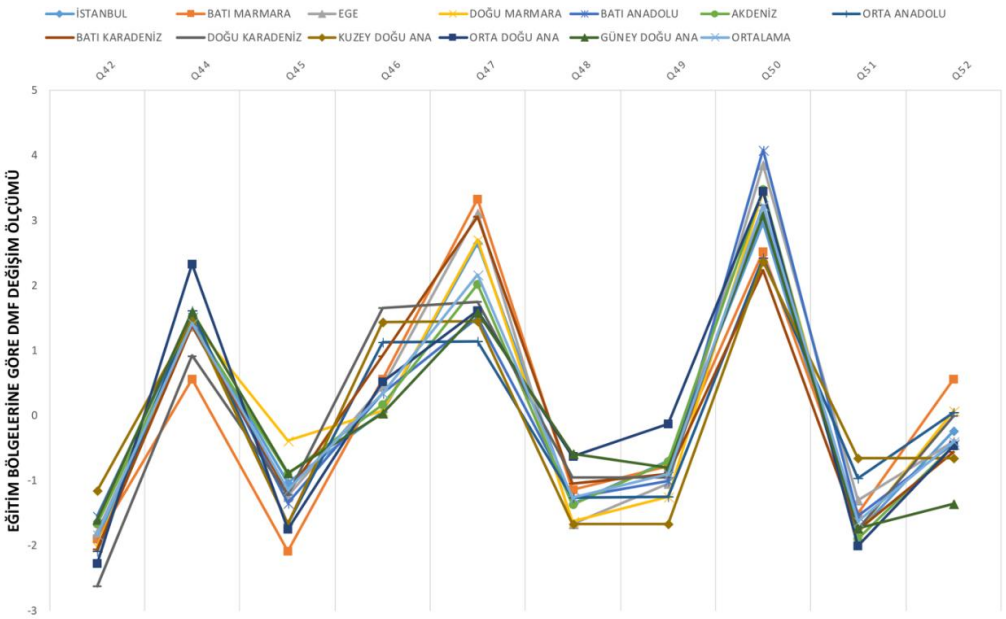


Figure 11. Change in DIF Subject to Statistical Regions for the Ten Items in Cluster05

Figure 11 reveals that the highest rate of change is observed in items Q44, Q46, Q47, Q49 and Q52. It can be seen from the graph showing the DIF change subject to statistical regions for the items in Cluster05 that the greatest divergence from the mean value has been observed in Western Marmara, Eastern Black sea, Northeastern Anatolia, Central Eastern Anatolia and Southeastern Anatolia regions. Figure 12 shows the DIF change graph of the nine items in Cluster06a subject to statistical regions.

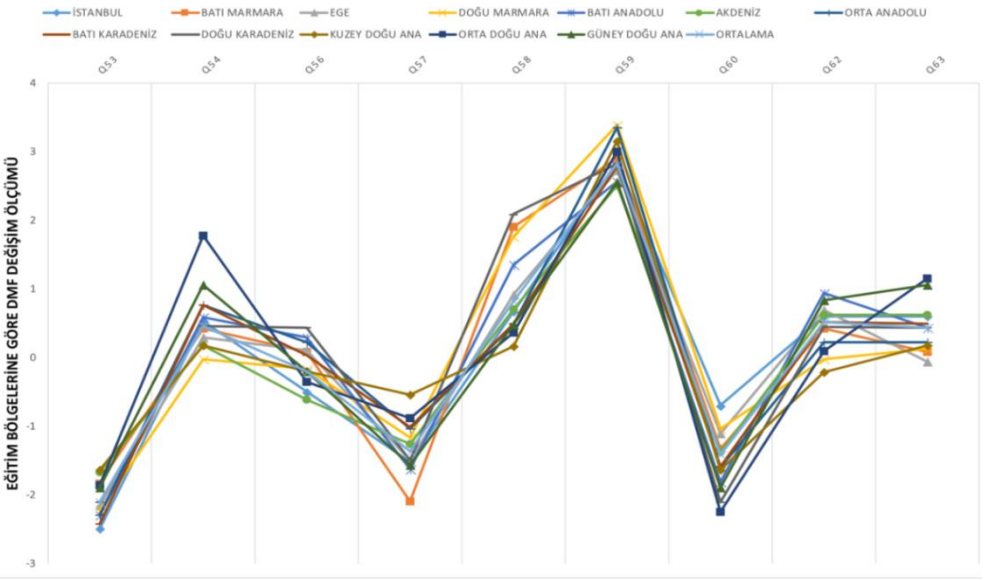


Figure 12. Change in DIF Subject to Statistical Regions for the Nine Items in Cluster06

It can be observed from Figure 12 that the highest rate of change in the graph is observed in items Q54, Q57, Q58, Q62 and Q63.

DISCUSSION and CONCLUSION

The aim of the present study was to determine whether the dichotomous (1-0) items in the PISA 2015 mathematics literacy subtest exhibit DIF subject to gender and statistical regions. IRT based Rasch model method was used for examining whether the items included in the study exhibit DIF or not.

Large scale examinations have a significant impact by way of their results on the education shareholders. It is important that the results of the examinations on which decisions related to individuals and the educational systems of countries are based contain minimum error. This will contribute to the accuracy of the decisions taken in accordance with the test results. Hence, it is expected that the tests applied in the field of education are free from bias. In other words, it is expected that the responses to the items in the examinations are not affected by factors such as gender, socioeconomic level, language, culture, territory, graduated school type, etc. excluding the abilities of the students. It is an important problem with regard to validity when the items operate for or against a certain group. Determining bias as a source of systematic error for examinations is also important for accountability.

Statistically significant DIF finding was observed subject to gender and statistical regions in the dichotomous (1-0) items of the mathematics subtest in PISA 2015 Turkey implementation. Statistically significant DIF findings were observed in five items in the mathematics subtest with regard to gender and in all items with regard to statistical regions.

Gender-based analyses of the items in the mathematics subtest of PISA 2015 revealed that item Q7 in Cluster01 operates in favor of males, whereas item Q8 operates in favor of females. Moreover, it was also understood that the item Q14 in Cluster02 operates in favor of males, whereas items Q58 and Q53 in Cluster06a operate in favor of females. It was understood as a result of the DIF analysis subject to gender for the 55 items of the mathematics subtest included in the study that five items exhibit. Of these items, two exhibited DIF in favor of females and three in favor of males. Demir and Köse (2014) carried out a study for examining whether the items included in PISA 2009 mathematics literacy subtest exhibit DIF subject to gender and culture. The study results put forth that two questions exhibit DIF subject to gender-based on the MH technique, three questions based on the LR technique and four questions based on the SIBTEST technique. In addition, DIF findings in favor of female students have been obtained as a result of the study by Akour et al. (2015) examining whether the PISA 2012 mathematics subtest results exhibit DIF or not. Atalay Kabasakal and Kıbrıslıoğlu Uysal (2017) conducted a study examining whether the PISA 2015 science subtest exhibits DIF subject to gender or not as a result of which it was observed that the number of items that exhibit DIF varies between two and six. Çıkrıkçı Demirtaşlı and Ulutaş (2015) carried out a study examining whether the items in the PISA 2006 science literacy subtest exhibit DIF subject to culture and gender or not. It was observed based on the DIF analysis subject to the item and item type that all multiple-choice items operate in favor of females, whereas two-thirds of the open-ended questions and a short response item operate in favor of males. It is observed especially in recent studies reporting gender DIF in large scale examinations that the number of items that exhibit DIF varies between three and six. The results obtained from these studies and the aforementioned literature findings are in accordance.

Statistically significant DIF findings were observed in all items when the results obtained from the analyses of the items in PISA 2015 mathematics subtest subject to statistical regions are examined. While at least 10 DIF cases were observed in the item based binary educational territory comparison, the maximum DIF cases observed were 38.

It is also very important for the implementation of the national and international tests in PISA 2015 mathematics subtest to take into consideration the impact of different demographic characteristics on the measurement results. It has been reported when the result indicating that the reasons for DIF in examinations carried out at the national level include variables such as gender and school type is taken

into consideration that it is inevitable for large scale examinations at the international scale such as PISA to include items that exhibit DIF (Bakan Kalaycıoğlu & Kelecioğlu, 2011). According to Sachse and Haag (2017), DIF can be observed due to the margin for error calculated for large scale examinations. In this regard, they have mentioned the need to reevaluate the methods used for calculating the standard error for national tendencies and taking into consideration the errors due to different points. Arikan, Van de Vijder and Yağmur (2018) expressed as a result of their study that less DIF is observed when tendency scores are used in DIF analyses.

It can be stated when the results of studies examining DIF subject to the gender of large scale examinations along with the results of the present study are taken into consideration that similar results have been attained. It is expressed when DIF subject to statistical regions is examined that different demographic characteristics should be taken into consideration and that different DIF prediction methods should be used. In addition, it is also observed that even though DIF description techniques yield similar results, they do not yield the same results due to the presence of algorithms and breakpoints at different classifications (Ardıç & Gelbal, 2017).

REFERENCES

Acar, T. (2011). Sample size in differential item functioning: An application of hierarchical linear modeling. *Kuram ve Uygulamada Eğitim Bilimleri, 11*(1), 284-288.

Akalın, Ş. (2014). *Kamu Personeli Seçme Sınavı genel yetenek testinin madde yanlılığı açısından incelenmesi.* (Doktora Tezi, Ankara Üniversitesi, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Akour, M., AL-Baddareen, G., H., A., & AL Duwairi, A. (2015). Exploring the jordanian gender gap in a large-scale assessment in mathematics. *Jordan Journal of Educational Sciences, 11*(1), 101-111.

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

Ariffin, S. R., Idrıs, R., & Ishak, N. M. (2010). Differential item functioning in Malaysian generic skills instrument (MyGSI). *Journal Pendidikan Malaysia, 35*(1), 1-10.

Arikan, S., Van de Vijver, F. R., & Yagmur, K. (2018). Propensity score matching helps to understand sources of DIF and mathematics performance differences of Indonesian, Turkish, Australian, and Dutch students in PISA. *International Journal of Research in Education an Science, 4*(1), 69-82. doi: 10.21890/ijres.382936

Atalay Kabasakal, K., ve Kıbrıslıoğlu Uysal, N. (2017). Öğrenci özelliklerinin cinsiyete dayalı değişen madde fonksiyonuna etkisi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *4*(8), 373-390. doi: 10.21031/epod.333451

Ateşok Deveci, N. (2008). *Üniversitelerarası kurul yabancı dil sınavının madde yanlılığı bakımından incelenmesi.* (Doktora Tezi, Ankara Üniversitesi, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Bakan Kalaycıoğlu, D. (2008). *Öğrenci Seçme Sınavı'nın madde yanlılığı açısından incelenmesi*. (Doktora Tezi, Hacettepe Üniversitesi, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Bakan Kalaycıoğlu, D., ve Kelecioğlu, H. (2011). Öğrenci Seçme Sınavının madde yanlılığı açısından incelenmesi. *Eğitim ve Bilim, 36*(161), 3-13.

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik Test Teorisi ve uygulanması.* Ankara: ÖSYM Yayınları.

Bekci, B. (2007). *Ortaöğretim Kurumları Öğrenci Seçme Sınavının değişen madde fonksiyonlarının cinsiyete ve okul türüne göre incelenmesi.* (Yüksek lisans Tezi, Hacettepe Üniversitesi, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Berberoğlu, G. (1995). Differential item functioning analysis of computation, word problem and geometry questions across gender and SES groups. *Studies in Educational Evaluation, 21*(4), 439-456. doi: 10.1016/0191-491X(95)00025-P

Birjandi, P., & Amini, M. (2007). Differential item functioning (Test Bias) analysis paradigm across manifest and latent examinee groups (On the construct validity of IELTS). *Human Sciences*, 55 (Special Issue On Linguistics),153-172.

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö., Karadeniz, Ş., ve Demirel, F. (2013). *Bilimsel araştırma yöntemleri.* (15. Baskı). Ankara: Pegem Akademi.

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

Çıkrıkçı Demirtaşlı, N., ve Uluştaş, S. (2015). A study on detecting of differential ıtem functioning of PISA 2006 science literacy ıtems in Turkish and American samples. *Eurasian Journal of Educational Research, 58,* 41-60 doi: 10.14689/ejer.2015.58.3

Demir, S., ve Köse, İ. A. (2014). Mantel-Haenszel, SIBTEST ve Logistik Regresyon yöntemleri ile değişen madde fonksiyonu analizi. *International Journal of Human Sciences, 11*(1), 700-714. doi: 10.14687/ijhs.v11i1.2798

Doolittle, A. E., & Cleary, T. A. (1987). Gender-based differantial item performance in mathematics achievement items. *Journal of Educational Measurement, 24*(2), 157- 166.

Ercikan, K., & Kim, K. (2009). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing, 5*(1), 23-35. doi: 10.1207/s15327574ijt0501_3

Özer Özkan, Y., & Fincan, F. B. (2017). An Investigation of Item Bias in Free Boarding and Scholarship Examination in Turkey. *International Test Commission (ITC) 2016 Conference*. Canada

Gamer, M., & Engelhard Jr, G. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education*, 12(1), 29-51.

Gierl, M. J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education, 25*(4), 280-296.

Gök, B., Atalay Kabasakal K., ve Kelecioğlu, H. (2014). PISA 2009 öğrenci anketi tutum maddelerinin kültüre göre değişen madde fonksiyonu açısından incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 5*(1), 72-87. DOI: 10.21031/epod.64124

Hanna, G. (1986). Sex differences in the mathematics achievement of 8th graders in Ontario. *Journal for Research in Mathematics Education, 17*, 231-237.

Harris, A.M., & Carlton, S.T. (1993). Patterns of Gender Differences on Mathematics Items on the Scholastic Aptitude Test. *Applied Measurement In Education*, 6(2),137-151.

Kıbrıslıoğlu Uysal, N., ve Atalay Kabasakal, K. (2017). Öğrenci özelliklerinin cinsiyete dayalı değişen madde fonksiyonuna etkisi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 8(4), 373-390.

Kan, A., Sünbül, Ö., ve Ömür, S., (2013). 6.-8. Sınıf seviye belirleme sınavları alt testlerinin çeşitli yöntemlere göre değişen madde fonksiyonlarının incelenmesi. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 9*(2), 207-222.

Karakaya, İ. (2012). Seviye belirleme sınavındaki fen ve teknoloji ile matematik alt testlerinin madde yanlılığı açısından incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri, 12*(1), 215-229.

Karakaya, İ., ve Kutlu, Ö. (2012). Seviye belirleme sınavındaki Türkçe alt testlerinin madde yanlılığının incelenmesi, *Eğitim ve Bilim, 37*(165), 348-362.

Kurnaz, F. B. (2006). *Peabody resim kelime testinin madde yanlılığı açısından incelenmesi*. (Yüksek Lisans Tezi, Hacettepe Üniversitesi, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Latifi, S., Bulut, O., Gierl, M., Christie, T., & Jeeva, S. (2016). *Differential Performance on National Exams: Evaluating item and bundle functioning methods using English, Mathematics, and Science Assessments*. SAGE Open, 1-14.

Le, L. (2009). Investigating gender differential item functioning across countries and test languages for PISA science items. *International Journal of Testing,* 9(2), 122-133. DOI: 10.1080/15305050902880769

Lyons-Thomas, J., Sandılands, D., & Ercikan, K. (2014). Gender differential item functioning in mathematics in four international jurisdictions. *Education and Science Large- Scale Assessment Special Issue, 39*(172), 20-32.

MoNE, (2010). PISA 2009 Uluslararası Öğrenci Değerlendirme Programı Ulusal Ön Raporu, Ankara: MEB Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı.

MoNE, (2013). PISA 2012 Ulusal ön raporu, Millî Eğitim Bakanlığı, Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü, Ankara.

MoNE, (2015). PISA 2015 Uluslararası Öğrenci Değerlendirme Programı Ulusal Raporu, Ankara: MEB Ölçme Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü.

OECD, (2015). OECD PISA Technical Report. OECD: http://www.oecd.org/pisa/data/2015-technical-report/ adresinden alındı. (Erişim Tarihi: 27.06.2018)

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

Öğretmen, T. (1995). *Differential item functioning (DIF) analysis of the verbal ability section of first stage of the university entrance examination in Turkey*. (Doctoral Dissertation, Middle East Tecnic University, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Öğretmen, T., ve Doğan, N. (2004). Öğretmen, T., & Doğan, N. (2004). OKÖSYS matematik alt testine ait maddelerin yanlılık analizi. *İnönü Üniversitesi Eğitim Fakültesi Dergisi, 5* (8), 61-76.

Özer Özkan, Y. (2012). *Öğrenci başarılarının belirlenmesi sınavından (ÖBBS) klasik test kuramı, tek boyutlu ve çok boyutlu madde tepki kuramı modelleri ile kestirilen başarı puanlarının karşılaştırılması*. (Doktora Tezi, Ankara Üniversitesi, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Özmen, D.T. (2014). PISA 2009 Okuma testi maddelerinin yanlılığı üzerine bir çalışma. *Eğitim Bilimleri ve Uygulama Dergisi*, *13*(26),147-165.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance", *Psychological Bulletin*, 114(3), 552-566.

Sachse, K. A., & Haag, N. (2017). Standard errors for national trends in international large-scale assessments in the case of cross-national differential item functioning. *Applied Measurement in Education*, *2*(30), 102-116. DOI: 10.1080/08957347.2017.1283315

Satıcı, K., ve Özer Özkan, Y. (2017). Temel eğitimden ortaöğretime geçiş sınavının (2014-Kasım) cinsiyet açısından madde yanlılığının incelenmesi. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, *13*(1), 254-274. DOI: 10.17860/mersinefd.305954

Suna, H.E. (2012). TIMSS 2007 Fen bilimleri testindeki maddelerin dil ve cinsiyet yanlılığı açısından incelenmesi. (Yüksek Lisans Tezi, Ankara Üniversitesi, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Şenferah, S. (2015). *2010 Seviye belirleme sınavı matematik alt testi için değişen madde fonksiyonlarının ve madde yanlılığının incelenmesi*.(Doktora Tezi, Gazi Üniversitesi, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed ıtem formats. *Applied Measurement in Education*, *25*(3), 246-280. DOI: 10.1080/08957347.2012.687650

Türkan, A. (2014). 2012- *Seviye Belirleme Sınavının Rasch modeline göre cinsiyet değişkeni açısından yanlılığının incelenmesi*. (Yüksek Lisans Tezi, Gaziantep Üniversitesi, Gaziantep). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Ulutaş, S. (2012). *PISA 2006 fen okuryazarlığı testindeki maddelerin yanlılık bakımından araştırılması*. (Doktora tezi, Ankara Üniversitesi, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Yalçın, S., ve Tavşancıl, E. (2015). TIMSS 2011 Fen uygulamasında cinsiyete göre farklılaşan madde fonksiyonunu açıklayan değişkenler. *Eğitim Bilimleri ve Uygulama*, *14*(27), 1-21.

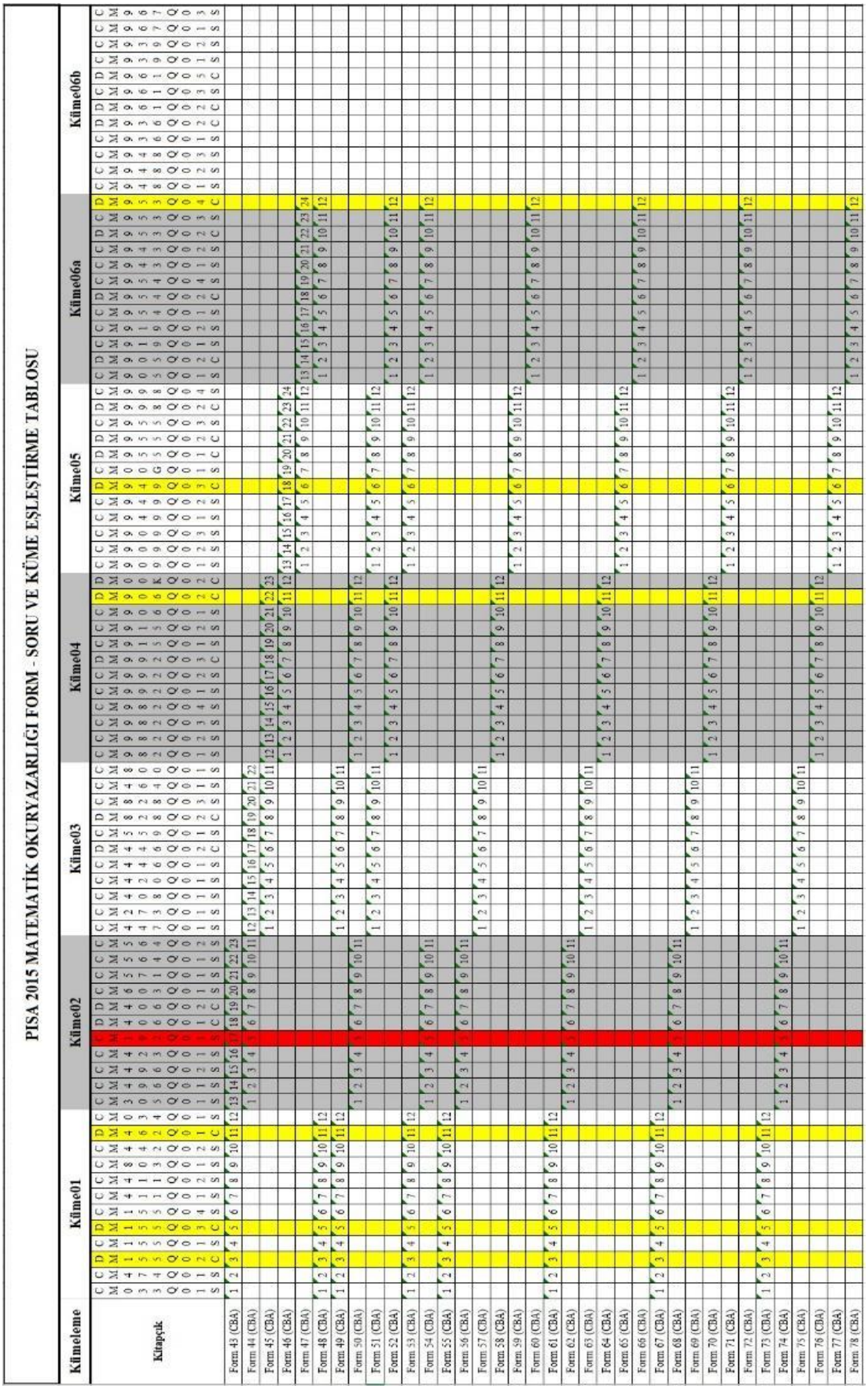
Yurdugül, H., ve Aşkar, P. (2004). Ortaöğretim kurumları öğrenci seçme ve yerleştirme sınavının, öğrencilerin yerleşim yerlerine göre, diferansiyel madde fonksiyonu açısından incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *27*(2004), 268-275.

Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural Assessment. European *Journal of Applied Physiology*, *54*(2004), 119–135.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2004). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment*, *9*(12), 61-78. doi: 10.1207/s15326977ea0901&2_3

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP. *History Assessment Journal of Educational Measurement*, *26*(1), 55-66.

Appendix A. PISA 2015 Mathematics Literacy Form-Question and Cluster Matching Table



ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(3); 302-310

Development of a Short Form: Methods, Examinations and Recommendations

Hakan KOĞAR *

Abstract
The aim of this review is to explain the methods that can be used when developing short form of a measurement tool and to examine some short form development studies in the field of health sciences literature by taking into consideration the criticisms of short form development studies. It is seen that short form development studies are especially concentrated in the fields of health sciences. The main reason for this situation has been shown that clinicians need fast and reliable measurement tools to reduce the pressure on them. The review results of the 12 articles selected for this research show that there are very few studies that follow the guidelines for short form development. Researchers are advised to develop the short form of the scale by taking into account the criteria mentioned in this study. It is recommended to select measurement instruments which are developed in accordance with ethical rules and have sufficient psychometric properties. Clinical researchers should be aware that the perception that measuring instruments containing less items are less valid does not show the truth. The same psychometric standards are sought for each measurement tools.

Key Words: Short form, scale development, classical test theory, item response theory

INTRODUCTION

Attempts to develop a short form of an existing measurement tool started at the beginning of the 20th century when it was questioned if it was essential to use all the items on Doll's (1917) Binet-Simon intelligence test to measure intelligence. Studies on the development of short forms, the number of which increased in the 1950s, initially focused on the measurement tools used for clinical assessments as an outcome of criticisms made against numerous items on intelligence and ability tests (Levy, 1968). Levy (1968), who examined the short form development studies in that period, criticized these kinds of studies in his study by claiming that studies aiming to produce short forms diverted from their real purposes and became a commonplace academic activity.

Why Short Forms?

The primary aim of studies on the development of short forms in the mid-20th century was for effective use of the time available (Levy, 1968). The aim was to establish a balance between economic use of time and energy and accurate test estimations (Doppelt, 1956). Today, however, there are different reasons underlying efforts to develop short forms. Some of these are as follows: finding the use of short forms convenient in studies involving multiple cultures with multiple variables, saving time by measuring fewer behaviors, the possibility of developing a child form, reaching the goals of selection and placement more quickly and developing a short form having the same validity as that of a long form. Studies on the development of a short form are observed to be more common in the field of health sciences. This is primarily attributed to the health specialists' need for a quick and reliable measurement tool to relieve the pressure they are under (Smith, McCarthy & Anderson, 2000).

* Assoc. Prof. Dr., Akdeniz University, Faculty of Education, Antalya-Turkey, e-mail: hkogar@gmail.com, ORCID ID: 0000-0001-5749-9824

To cite this article:
Koğar, H. (2020). Development of a Short Form: Methods, Examinations, and Recommendations. *Journal of Measurement and Evaluation in Education and Psychology*, *11*(3), 302-310. doi: 10.21031/epod.739548.

Received: 18.05.2020
Accepted: 08.09.2020

Psychometric Theories Used In Developing Short Forms

Various methods were used in developing short forms in the mid-20th century, some of which are selecting the item set yielding the highest correlation with the long form of the measurement tool, forming an item sample based merely on item statistics, and selecting a factor or factors with the highest validity (Levy, 1968). It was revealed that among these methods, it was the selection of an item sample based on classical item statistics that was used most frequently; in addition to these statistics, some other statistics, such as Guttman's scalogram analyses were also found to be utilized. These methods, the use of which are limited today, as well as other methods that started to be used with the advancements in technology as of 1970, are explained below in detail in association with the psychometric theories they are based on.

Classical Test Theory (CTT)

Classical item statistics

The most important of the classical statistics that go way back to the times when intense interest in scale development studies started in the field of social sciences are item difficulty index, item discrimination index and item total correlation coefficient. The item difficulty index refers to the difficulty level of an item with respect to the ability level of the individuals in a group. According to Henning (1987), an item being too easy or too difficult can indicate that the score distribution is skewed, which may show that the item prepared is not compatible with the ability level of the group. The item discrimination index, the purpose of which is to distinguish a high scoring group from the low scoring group in reference to the total score, is an important index value that determines the place of an item in a scale. As for the item total correlation coefficient, it displays the relationship between the trait the item or the content is testing and the trait that the total score of the test is measuring. Each item score should be associated with the total score. Items that show a high level of relationship with the total score are those items that highly account for the variance in the total score, as in the factor load of a factor analysis. In other words, these items have a high level of validity. These statistical techniques are frequently utilized in the 21st century as their calculations are relatively easy. However, particularly item total correlation is known to result in misleading findings as it is based on the Pearson's product-moment correlation coefficient (Raykov & Marcoulides, 2011)

Biggers' (1976) Spearman-Brown prediction method

Biggers (1976), who criticized the use of classical item statistics, stated that the long form is the unity of n number of parallel short forms, and that the short form developed is merely one of these parallel forms; thus, it is not possible to determine which short form is a more appropriate selection. Moreover, generating a short form by eliminating or choosing items is an irreversible experimental method; that is, he stated that it was not possible to initially develop a short form and then add items to try to obtain the long form of the test. For this purpose, Spearman-Brown proposed the prediction method as an alternative to developing a short form. He, first of all, developed the short form of a 40-item dogmatism scale with the aid of classical item analyses. Subsequently, he divided the test into two parts based on odd-and even-numbered items, and calculated the correlation coefficient between the total score of the short form, obtained using the classical item analyses, and the total score of the long form of the scale. It was found that the coefficient between the scores obtained from one half of the scale based on odd-numbered items and the scores from the long form of the scale was .92, while the correlation between the scores of the other part of the scale based on even-numbered items and the scores obtained from the long form of the scale was .93.

Factor analysis

Factor analysis is a multivariate statistical technique by which items are associated with one or more latent items by means of a model constructed based on the relationships among the observed variables. It is the most frequently used statistical technique in studies on scale development and adaptation as well as in short form scale development studies. However, sample studies in which factor analysis is accurately conducted is rarely encountered. According to Goretzko, Pahm and Buhner (2019), in studies where factor analysis is utilized, problems are experienced particularly in identifying the sample size, in choosing the correct rotation method and the correct technique for selecting the factor-revealing technique. Based on the studies they examined, Fabrigar, Wegener, MacCallum and Strahan (1999) made some recommendations for studies in which factor analysis would be used. According to researchers, the number of items that needs to be included in a factor is at least four, and the sample size needs to be at least 400. In cases where multivariate normality is obtained, mostly likelihood estimation, and in other conditions such techniques as data rotation methods or principal axis factoring should be used. Smith, McCarthy and Anderson (2000) stated that factor analysis was frequently used in short form development studies and criticized the formation of the short form by applying a factor analysis to the data set obtained from the long form of a scale. This kind of an approach is based on the assumption that the long and short forms of a scale have the same structure. However, there is no certainty that the long and the short forms of the scale have the same factor structure. As a solution to this problem, they proposed running a separate factor analysis on the items of the short form. If these findings are similar to those obtained from the long form, then this means that the two forms of the scale can be alternatives to each other. On the other hand, significant differences between the factor structures of the short and long forms can indicate that these two forms measure different traits.

Item Response Theory (IRT)

Item response theory (IRT) was developed to overcome the various limitations of CTT and particularly the inadequate approaches in determining psychometric properties of scales. It includes two approaches, namely parametric (Birnbaum, 1968; Rasch, 1960) and non-parametric approaches (Mokken & Lewis, 1982). Researchers should choose one of these fundamental approaches based on the purpose of the research study and on the extent to which the assumptions are met. When there is a symmetrical relationship between a latent trait and responses to the item, and when uni-dimensionality and a large sample size can be ensured, parametric IRT models can be utilized. On the other hand, when there is an asymmetrical distribution and a small sample size, non-parametric IRT models can be used. It is known that parametric and non-parametric IRT models show resistance to conditions where the unidimensionality of IRT models are violated (Embretson & Reise, 2000; Sodano & Tracey, 2011).

Parametric Item Response Theory

Like factor analysis, techniques based on CTT can obtain information based only on relationships among independent items. Moreover, all the statistical findings obtained are dependent on the sample. The greatest advantage of the item response theory (IRT) is that it eliminates the dependence on the sample by claiming invariance of the item parameters. The standard errors in IRT are calculated separately for each level of latent trait. In this way, the group's fixation to one error value is overcome. This topic is important in terms of the decisions made especially in clinical measurements. IRT obtains information from the items that can distinguish groups with high and low ability. Furthermore, as IRT yields item characteristic curves (ICC) at each trait level and for each dimension, the amount of information necessary to obtain the short form of the scale can be estimated. While it is possible to determine the level of ability with a higher level of certainty with items yielding higher amounts of information, determining ability level with items yielding lower amounts of information is possible with lower level of certainty. The items yielding the highest amounts of information can be selected in accordance with the range of the trait being measured. By selecting the better performing items providing adequate information across different levels of the trait, it is possible to develop a short form with high psychometric properties. In addition, rather than obtaining a single coefficient yielded as in reliability

measuring techniques based on CTT, such as Cronbach alpha, test information functions (TIF) in IRT allows the assessment of the certainty for each level of the structure being measured. Thanks to TIF, ability levels that include high amounts of information and thus include low amounts of error can be determined and, in this way, a high level of local reliability can be obtained (Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991). TIF can be developed by means of ICCs. Hence, in short form development studies, the aim should be to reach the same amount of information that the long form possesses by selecting items yielding high amounts of information.

Non-Parametric Item Response Theory and the Mokken Scale Analysis

Non-parametric Item Response Theory (IRT) is an approach, the use of which has become widespread as of the beginning of the 21st century owing to the very low number of assumptions it has. Its interpretation is also easy for researchers. It is commonly used particularly for exploratory purposes. Like in parametric IRT, ICCs are also obtained in non-parametric IRT. ICCs can be obtained in all kinds of distributions — monotonically decreasing, monotonously non-decreasing, symmetrical or asymmetrical distributions (Meijer & Baneke, 2004). Non-parametric IRT models are categorized into two: Mokken scale analyses and non-parametric regression prediction models. The Mokken scale analysis is the extended probabilistic version of the Guttman scale. It has two approaches, namely the Monotone Homogeneity Model (MHM) and the Dual Monotone Model (DMM). MHM defines the relationship between individuals and items that belong to unidimensional item groups and that have an item response function displaying a latent trait and a monotonic relationship. It is the simplified version of DMM with fewer assumptions. The primary aim of these models is to order items and individuals (Koğar, 2015). Parametric and non-parametric IRT follows the algorithm for simultaneous selection (Lei, Dunbar, & Kolen, 2004).

Ant Colony Optimization

Even though it is not a psychometric theory, the Ant Colony Optimization (ACO), one of the most current and effective techniques developed with the aim of developing short forms, is based on the algorithm of ants' search for food (Dorigo & Stützle, 2004). It is believed that this algorithm, which calculates the shortest route between the ant colony and the food source, can be used in short form development studies. It is modeled by utilizing the Structural Equality Model (SEM). The ACO algorithm aims to reveal the model with the highest compatibility by converging towards the appropriate model. It tries to produce the best short form based on the repetition of this process.

Purpose and Importance of the Research

In the present study, some of the methods frequently utilized to develop the short form of a measurement tool are explained. Even though the number of studies based on developing short forms is quite high and has a long history, discussions in this area continue to exist. Criticisms against studies on developing short forms can be examined from two basic aspects. First, these studies prioritize the validity of the measurement tool over any other property. According to psychometric theories, the validity of a measurement tool is obligatory. Such factors that relate to convenience, such as reducing item numbers or using time more effectively, are of secondary importance. Hence, while developing a short form of a scale, the primary aim should be to obtain a short form that is at least as valid as the long form of the scale. However, it is noticed in literature that there are research studies that divert from this aim. The second criticism is that during the development of the short form of a measurement tool, methodology errors are frequently made, the short form is developed carelessly and imprecisely, and the short form is not compared to the long form. This could be attributed to the limited information regarding the methodology of developing short forms in the literature (Smith, McCarthy & Anderson, 2000). The aim of the current study is to explain the methods that can be used to develop a short form of a measurement tool and to examine the methodology that some studies employed to develop short forms in the literature

of education and health sciences by taking into consideration the criticisms made against studies on short form development.

What Needs to be Taken into Account in Short Form Development Studies and The Examination of Some Studies

In this part of the study, what needs to be taken into account while developing short forms are explained and itemized based on the studies of Levy (1968), Smith, McCarthy and Anderson (2000), and Hagtvet and Sipos (2016). For the present study, 12 short form development studies published in journals indexed in the ERIC and PUBMED databases between the years 2011 and 2019 were selected. In all of these studies, the aim was to develop a new, short form. The present study examined whether or not the short form in each was developed in accordance with the principles stated below. Identification regarding these studies is presented in Table 1.

Table 1. Identification of the Studies Examined

Source	The Short Form of the Scale
Baiocco, Pallini & Santamaria (2014)	Adolescent Friendship Attachment Scale
Woudstra, Meppelink, Maat, Oosterhaven, Fransen & Dima (2019)	Short Assessment of Health Literacy
Lim & Chapman (2013)	Attitudes Toward Mathematics Inventory
Jenkinson, Kelly, Dummett & Morley (2019)	The Oxford Participation and Activities Questionnaire
Rogers, M. E., Creed, P. A., Searle, J., & Hartung, P. J. (2011)	Physician Values in Practice Scale
Ferrario, Panzeri, Anselmi & Vidotto (2019)	Illness Denial Questionnaire
Morin, Valois, Crocker & Robitaille (2019)	Intellectual Disability Questionnaire
Nimon & Zigarmi (2015)	Work Intention Inventory
Milavic, Padulo, Grgantov, Milic, Mannarini, Manzoni, Ardigo & Rossi (2019)	The Psychology Skills Inventory For Sports
Park & Hill (2017)	Occupational Work Ethic Inventory
Siefert, Sexton, Meehan, Nelson, Haggerty, Dauphin & Huprich (2019)	DSM–5 Levels of Personality Functioning Questionnaire
Bohlmeijer, Klooster, Fledderus, Veehof & Baer (2011)	Five Facet Mindfulness Questionnaire

1. Initially, the long form of a measurement tool should be sufficiently reliable and valid:

When a short form is to be developed, the first step to be taken is to evaluate the reliability and validity values of this scale long form. If a measurement tool is not reliable nor valid, then any short form of this tool will most likely have inaccurate validity and reliability values. Two of the 12 studies examined explained the psychometric traits of the long form in detail. Other studies sufficed by merely reporting reliability coefficients or stating that the long form is valid and reliable measurement tool.

2. If a short form does not have the same psychometric traits as those of the long form of a measurement tool, then it is not a single short form, but one of the alternative short forms:

The item set in a short form should be formed by randomly selecting an item set from the long form of the scale that best explains the structure. The next phase is to make the decision as to whether the short form is an "equivalent" or "exchangeable" form. The "equivalent" short form has the same psychometric traits as those of the long form and, therefore, can be used as an alternative to the long form. The "exchangeable" short form, however, does not possess psychometric traits to the same degree as those of the long form. Hence, in another study replicated with a similar method, it is likely to obtain similar forms. "Exchangeable" short forms generally have a lower validity than the long form. In this case, the researcher should reveal and discuss the different forms, the different factor structures, and the different items or item sets that can be alternatives to this form. Otherwise, this form cannot be an alternative to

the long form. Having fewer items in the short form does not mean a lower level of validity is sufficient. This issue is so important that it cannot be disregarded. In two of the studies examined, it was deduced that the form assumed an "equivalent" nature. The short forms developed in these studies were at least as valid and as reliable as the long form. However, in the remaining ten studies, since there was no sufficient information about the reliability and validity of the long form, no interpretation could be made about these studies.

3. A transition should be made from the population behavior (items in the long form) to the the sample behavior (items to be included in the short form) by ensuring that it reflects the nature of the trait which the measurement tool is measuring:

One other factor that needs attention is related to the selection of items for the short form from the item pool in the long form. The selected items that will make up the sample of the behavior should be able to reflect the population behavior in the long form. This topic is as important as psychometric properties and is related to content validity. A well-explained and well-defined content is a topic of priority that is of vital importance for construct validity. In order to maintain the content domain, not only statistical evidence but also expertise in the field is important in the selection of the items to be included in the sample. Only one of the studies examined was observed to have discussed the content of the long form in detail and took into consideration the content as well as the statistical analyses when choosing items for the short form. In all the other studies, only statistical evidence was taken into consideration.

4. The view that "if the long form of the measurement tool is valid, then its short form is also valid" is wrong:

Even if a short form includes the items in the long form as well, this does not ensure that the short form will be reliable and valid. The short form includes fewer items and less content. From this respect, it is psychometrically at a disadvantage. For this reason, the psychometric properties must definitely be statistically proven. In all the studies examined, statistical evidence was sought for the reliability and validity of the short form.

5. In measurement tools with multiple dimensions, the content and psychometric properties should be analyzed for each dimension:

In structures with multiple dimensions, the psychometric properties of the scale should be examined by associating each item of the scale with the relevant dimension. In this case, evidence should be presented to prove that each dimension is reliable and valid. For example, if item selection is to be made based on item-total correlations, the total score should be the factor score, not the overall total of the measurement tool. It should be ensured that there are at least four items in one dimension. If it is essential to omit one dimension completely from the scale, then the relevant theoretical and statistical foundation should be presented in detail. It should be noted that the lower the number of items are, the the lower the content validity will be. One of the 12 studies examined was disregarded because it had a unidimensional structure. 10 of the remaining studies was found to have taken into consideration the multidimensional structure and run the statistical analyses. Even though the present study had a multidimensional structure, it obtained the proofs for the latent trait by means of the total score of the scale.

6. Evidence regarding reliability should be obtained within the scope of various types of reliability:

Construct validity should be the primary concern in determining the validity of a short form. However, in reporting reliability, different kinds of evidence for reliability such as internal consistency reliability, inter-rater agreement in measurements of behavior, and stability reliability need to be obtained. Reliability is a concept related to error and it is not possible to mention only one type of error in a measurement process. Hence, reliability coefficients that take error into consideration from different perspectives should be used. Only one of the 12 studies that were examined obtained internal consistency and stability reliability coefficients. To this end, the Cronbach alpha and the test-retest reliability coefficients were used. It was observed that in one of the research studies the reliability coefficient was not reported. All the remaining studies were found to have reported the internal consistency reliability coefficient. Eight of these studies reported the Cronbach alpha reliability coefficient, one reported the

Raykov's maximum reliability coefficient and one reported the person reliability and person discrimination coefficients.

7. The psychometric properties of the short form should be examined independent of the long form:

The short form of a measurement tool is a copy of the long form which displays a high degree of association. However, this high degree of association does not prove that the short form is reliable and valid. The concepts of validity and reliability are not transferrable and transitive. For this reason, the psychometric properties of the short form must definitely be examined independent of the long form and evidence should be reported. The proofs obtained from one independent group should be compared with the reliability and validity proofs of the long form. While half of the studies examined were found to have obtained the reliability and validity coefficients independent of the long form, the other half of the studies remained limited to merely reducing the number of items in the long form.

8. In clinical and behavioral measurement tools, the classification accuracies of the short form should also be examined:

The aim of some clinical measurement tools is to make classifications. The aim should be to refrain from negative classification (diagnosing an individual with a syndrome as having no syndrome) and positive classification (diagnosing an individual without a syndrome as having a syndrome). Thus, proofs independent of the long form should be obtained. An accurate classification and diagnosis by the long form does not guarantee that the short form can serve the same purposes as well. Four of the studies examined can be used for clinical purposes. None of these studies reported any proof for accuracy of classification.

9. That the time saved by developing a short form is meaningful and important should be justified:

One of the concrete aims of developing a short form is to save time. However, as previously mentioned, validity and reliability are more important than time. Hence, the researcher should explain how much time was saved and show that the time saved did not impact the the psychometric properties. On average 40 minutes is needed to fill in a long form with 80 items. Assuming that the short form of such a form would include 40 items, it can be said that 20 minutes will be saved. However, it should be noted that a reduction of 40 items will have negative impacts on the reliability and validity. The degree of these effects should be discussed in the study. One of the studies examined the time to be saved by developing a short form and discussed this by taking into consideration the psychometric properties of the measurement tool. The other studies, however, merely stated that time would be used more effectively.

CONCLUSION

While developing a short form of a measurement tool, one of the greatest misconceptions of researchers is the idea that the reliability and validity of the short form and the original measurement tool are the same. This causes some researchers to disregard psychometric properties such as reliability and validity, and prevents some researchers from paying the necessary importance to this issue. In the development of a short form, the observed number of items decreases. Therefore, the content and coverage are narrowed, which makes it difficult for these two test forms to be alternatives to each other.

The 12 research studies selected for the present study were screened in two important indexed databases in the fields of health and social sciences. The results which the examinations yielded show that the number of studies conforming to the rules of developing short forms is limited. This shows that short form development studies, which have been under discussion since mid-20th century, are still subject to discussion. When the study by Levy (1968) is compared to that of Smith, McCarthy and Anderson (2000), it is true that the examined studies performed a more accurate study. However, in the examined studies the following problems were identified: not reporting a detailed account of the reliability and validity information of the long form of the measurement tool, not paying attention to the fact that the short form must be as reliable and valid as the long form, not being aware of the fact that the concept of "exchangeable" short form emerges in cases where the reliability and validity of the short form is not at the same level as those of the long form, limiting reliability to merely reporting the internal

consistency coefficients, obtaining the psychometric properties of the short form independent of the long form, not providing a detailed explanation of the content of the long form of the measurement tool, and not obtaining proof regarding the fact that the content of the short form can be generalized to the long form as well. In these studies, there are also deficiencies in terms of not explaining how much time is saved, which is one of the primary aims of developing short forms, and how this impacts psychometric properties. Furthermore, it was observed that in clinical measurement tools, classification accuracy was not tested.

When developing short forms, researchers are recommended to use the long form of the measurement as a starting point and take the criteria mentioned in the present study into consideration. On the other hand, studies aiming to adapt short forms of the measurement tool are not recommended owing to some important points such as the psychometric properties of short forms may not be precise. For this reason, instead of conducting a short form adaptation study, initially adapting the long form of a measurement tool to the related culture, and then developing the short form of the adapted measurement tool is recommended to be a more sound approach.

Particularly from the clinical researchers perspective, it is not sufficient to choose a measurement tool whose short form is already developed merely because it was published in a refereed journal and because it will save more time. Measurement tools that were developed in accordance with ethical principles and have sufficient psychometric properties are recommended to be selected. Clinical researchers should note that the perception that measurement tools with fewer items are less valid is not true. The same psychometric standards should be sought in each measurement tool. Moreover, in selection of a short form, it is recommended that one should critically analyze whether or not the steps outlined in the present study were followed during its developmental process; a short form that has the essential properties can be utilized for clinical or other purposes.

REFERENCES

Baiocco, R., Pallini, S., & Santamaria, F. (2014). The development and validation of an Italian short form of the adolescent friendship attachment scale. *Measurement and Evaluation in Counseling and Development, 47*(4), 247-255. https://doi.org/10.1177/0748175614538060

Biggers, J.L. (1976). An a priori approach for developing short-forms of tests and inventories. *The Journal of Experimental Education, 44*(3), 8-10. https://doi.org/10.1080/00220973.1976.11011528

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M Lord, & R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397–479). Reading, MA: MIT Press.

Bohlmeijer, E., Ten Klooster, P. M., Fledderus, M., Veehof, M. , & Baer, R. (2011). Psychometric properties of the five facet mindfulness questionnaire in depressed adults and development of a short form. *Assessment, 18*(3), 308-320. https://doi.org/10.1177/1073191111408231

Doll, E. A. (1917). A brief Binet-Simon scale. *Psychological Clinic, 11*, 197-211.

Doppelt, J. E. (1956). Estimating the full scale score on the Wechsler Adult Intelligence Scale from scores on four subjects. *Journal of Consulting Psychology, 20*(1), 63. https://doi.org/10.1037/h0044293

Dorigo, M., & Stützle, T. (2004). *Ant colony optimization*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/1290.001.0001

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: L. Erlbaum Associates.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Ferrario, S. R., Panzeri, A., Anselmi, P., & Vidotto, G. (2019). Development and psychometric properties of a short form of the Illness Denial Questionnaire. *Psychology Research and Behavior Management, 12*, 727. https://doi.org/10.2147/PRBM.S207622

Jenkinson, C., Kelly, L., Dummett, S., & Morley, D. (2019). The Oxford Participation and Activities Questionnaire (Ox-PAQ): development of a short form and index measure. *Patient Related Outcome Measures, 10*, 227-232. https://doi.org/10.2147/PROM.S210416

Goretzko, D., Pham, T. T. H., & Bühner, M. (2019). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology, (In Press)*, 1-12. https://doi.org/10.1007/s12144-019-00300-2

Hagtevt, K. A., & Sipos, K. (2016). Creating short forms for construct measures: The role of exchangeable forms. *Pedagogika, 66*(6), 689-713.

Hambleton, R. K., Swaninathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.

Henning, G. (1987). *A guide to language testing- development, evaluation, Research*. London: Newbury House Publisher.

Koğar, H. (2015). Madde tepki kuramına ait parametrelerin ve model uyumlarının karşılaştırılması: Bir Monte Carlo Çalışması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 6*(1), 142-157. https://doi.org/10.21031/epod.02072

Lei, P. W., Dunbar, S. B., & Kolen, M. J. (2004). A comparison of parametric and nonparametric approaches to item analysis for multiple choice tests. *Educational and Psychological Measurement, 64*, 565–587. https://doi.org/10.1177/0013164403261760

Leite, W. L., Huang, I. C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivariate Behavioral Research, 43*(3), 411-431. https://doi.org/10.1080/00273170802285743

Levy, P. (1968). Short-form tests: A methodological review. *Psychological Bulletin, 69*(6), 410. https://doi.org/10.1037/h0025736

Lim, S. Y., & Chapman, E. (2013). Development of a short form of the attitudes toward mathematics inventory. *Educational Studies in Mathematics, 82*(1), 145-164. https://doi.org/10.1007/s10649-012-9414-x

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case seng for nonparametric item response theory modeling. *Psychological Methods, 9*, 354–368. https://doi.org/10.1037/1082-989X.9.3.354

Milavic, B., Padulo, J., Grgantov, Z., Milić, M., Mannarini, S., Manzoni, G. M. ..., & Rossi, A. (2019). Development and factorial validity of the Psychological Skills Inventory for Sports, Youth, version–Short Form: Assessment of the psychometric properties. *PloS one, 14*(8), 1-17. https://doi.org/10.1371/journal.pone.0220930

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous responses. *Applied Psychological Measurement, 6*, 417–430. https://doi.org/10.1177/014662168200600404

Morin, D., Valois, P., Crocker, A. G., & Robitaille, C. (2019). Development and psychometric properties of the Attitudes Toward Intellectual Disability Questionnaire–Short Form. *Journal of Intellectual Disability Research, 63*(6), 539-547. https://doi.org/10.1111/jir.12591

Nimon, K., & Zigarmi, D. (2015). Development of the work intention inventory short-form. *New Horizons in Adult Education and Human Resource Development, 27*(1), 15-28. https://doi.org/10.1002/nha3.20090

Park, H., & Hill, R. B. (2018). Development and validation of a short form of the occupational work ethic inventory. *Journal of Career and Technical Education, 32*(1), 9-28. https://doi.org/10.21061/jcte.v32i1.1588

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmark's Paedagogiske Institut.

Rogers, M. E., Creed, P. A., Searle, J., & Hartung, P. J. (2011). The physician values in practice scale-short form: Development and initial validation. *Journal of Career Development, 38*(2), 111-127. https://doi.org/10.1177/0894845310363593

Sodano, S. M., & Tracey, T. J. (2011). A brief Inventory of Interpersonal Problems–Circumplex using nonparametric item response theory: Introducing the IIP–C–IRT. *Journal of Personality Assessment, 93*(1), 62-75. https://doi.org/10.1080/00223891.2010.528482

Raykov, T., & Marcoulides, G. A. (2011). Classical item analysis using latent variable modeling: a note on a direct evaluation procedure, *Structural Equation Modeling, 18*(2), 315-324. https://doi.org/10.1080/10705511.2011.557347

Siefert, C. J., Sexton, J., Meehan, K., Nelson, S., Haggerty, G., Dauphin, B., & Huprich, S. (2019). Development of a short form for the DSM–5 levels of personality functioning questionnaire. *Journal of Personality Assessment, (In Press)*, 1-11. https://doi.org/10.1080/00223891.2019.1594842

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12*(1), 102-111. https://doi.org/10.1037/1040-3590.12.1.102

Woudstra, A. J., Meppelink, C. S., Maat, H. P., Oosterha, ven, J., Fransen, M. P., & Dima, A. L. (2019). Validation of the short assessment of health literacy (SAHL-D) and short-form development: Rasch analysis. *BMC Medical Research Methodology, 19*(1), 122-131. https://doi.org/10.1186/s12874-019-0762-4

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(3); 311-323

Investigation of the Effect of Missing Data Handling Methods on Measurement Invariance of Multi-Dimensional Structures *

Mehmet Ali IŞIKOĞLU ** Burcu ATAR ***

Abstract
The purpose of this study was to compare the missing data handling methods on measurement invariance of multi-dimensional structures. For this purpose, data of 10857 students who participated in PISA 2015 administration from Turkey and Singapore and fully responded to the items related to affective characteristics of science literacy was used. Data with different percentages of missing data (5%, 10%, and 20% missing data) were generated from the complete data set with missing completely at random (MCAR) mechanism. In all data sets, missing data was completed with listwise deletion (LD), serial mean imputation (SMI), regression imputation (RI), expectation maximization (EM), and multiple imputation (MI) methods. Measurement invariance of the construct being measured between countries on completed data sets was investigated with multiple-group confirmatory factor analysis (MG-CFA). Findings from each dataset were compared with reference values. In the results of the study, RI and MI methods in the data set with 5% missing, EM method in the data set with 10% missing, and MI method in the data set with 20% missing gave the more similar results to the reference values than the other methods.

Key Words: Missing data handling methods, measurement invariance, multiple-group confirmatory factor analysis, PISA 2015, science literacy.

INTRODUCTION

Measurement instruments are of great importance in education systems. In order to train qualified workforce in accordance with the needs of the society, placement of individuals in educational institutions and programs, making changes and improvements in educational systems can be made based on the findings obtained from measurement instruments. As a result of national and international assessment studies, countries can even change their educational policies. In particular, the results of large-scale assessment studies that enable international comparisons are followed with interest by all stakeholders of the education.

PISA (Program for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study) aim to make cross-country comparisons. PISA and TIMSS are large-scale studies that aim to make comparisons between countries and can affect educational policies at national and international levels. The comparability of the results, especially in international assessments, is of great importance in the evaluation of countries. To be able to interpret the findings from different groups who took the same measurement instrument, the measurement instrument should have the same meaning for all groups. In this context, the concept of measurement invariance emerges. Drasgow (1984) defines measurement invariance as the similar relationships between observed test scores and latent traits across all subgroups.

The data obtained by the measurement instruments are not always complete. For reasons caused by the examinee, the measurement instrument, or the administrator, some data may be missing on data sets.

* This study is based on Mehmet Ali Işıkoğlu's master thesis titled "Comparison of Influence of the Missing Data Handling Methods on Measurement Invarıance".

** Teacher, Ministry of National Education of Republic of Turkey, Ankara-Turkey, mali.isikoglu@gmail.com, ORCID ID: 0000-0001-5104-5661
*** 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:
Işıkoğlu, M. A., & Atar, B. (2020). Investigation of the effect of missing data handling methods on measurement invariance of multi-dimensional structures. *Journal of Measurement and Evaluation in Education and Psychology, 11*(3), 311-323. doi: 10.21031/epod.749370.

Received: 08.06.2020
Accepted: 21.09.2020

Missing values arise as a problem since they directly affect the results of the statistical analyses of data sets. As in all other statistical analyses, in the measurement invariance studies, the missing data needs to be checked and managed before the analyses. The presence of missing data can affect the results of many analyses, including confirmatory factor analysis. Since excluding examinees with missing values from data sets will reduce the sample size, the power to generalize the results to the population decreases. In addition, the presence of missing values can cause type I and type II errors. Even the difference in the methods used to handle the missing data problem may lead to different findings from the analysis (Harrington, 2009).

Many techniques have been developed to handle missing data. Allison (2001) classified the missing data handling methods as traditional methods, methods based on Maximum likelihood, and multiple imputation approaches. Listwise deletion (LD) is the method that enables the complete data set to be obtained by removing all cases with unobserved data in any of the variables in the data set. If the missing data has the missing completely at random (MCAR) mechanism, the standard error estimates will be close to the standard error estimates of the real data, since the data set obtained by removing the missing data will be a random sample of the original data set (Allison, 2003). However, if each missing value is in different observations, the sample size will be greatly affected by this situation. This can cause problems even if the missing data has the MCAR mechanism (Enders, 2010). Serial mean imputation (SMI) assigns the mean of the observed data in the variable where the missing data is located, instead of missing data (Little & Rubin, 2002). Since the average of the variable is imputed to the missing data, it does not change the mean value of the variable. However, it reduces the distance of the missing data from the mean to zero, and it underestimates the variance (Enders, 2010; Tabachnick & Fidell, 2013). In the regression imputation (RI) method, the missing variables are imputed values with a regression equation obtained from the observed variables. However, the imputed values have some disadvantages, such as better fit than expected due to estimation from other variable and reducing the variance because it will most likely impute a value close to the mean. And, when the other variables are not a good predictor of the variable with missing value, there is no difference between regression imputation and mean imputation (Tabachnick & Fidell, 2013). Expectation maximization (EM), which is a method based on maximum likelihood, is a method consisting of two steps: expectation (E) and maximization (M), and consists of sequential steps based on a series of regressions. The disadvantage of this method is that the standard errors obtained from this method are not consistent with the actual standard errors (Allison, 2003). In the multiple imputation (MI) method, the random variance is added to the values estimated by regression, unlike EM method. However, different results can be obtained each time due to the addition of random variance (Allison, 2003).

There are two commonly used approaches in measurement invariance tests: confirmatory factor analysis and item response theory (Reise, Widaman & Pugh, 1993). Measurement invariance is generally examined by the multiple-group confirmatory factor analysis (MGCFA) method, which includes hierarchical steps (Whitaker & McKinney, 2007). In order to control the measurement invariance between groups with MGCFA method, configural invariance which requires equality of factor structures between groups, metric invariance which requires equality of factor loadings between groups, scalar invariance which requires equality of intercepts between groups, and strict invariance which requires equality of residual variances between groups must be tested hierarchically (Schoot, Lugtig & Hox, 2012).

Purpose of the Study

The purpose of this study was to investigate the effect of missing data handling methods on measurement invariance of multi-dimensional structures. In this context, the answer to the following problem is sought: "What is the effect of listwise deletion (LD), serial mean imputation (SMI), regression imputation (RI), expectation maximization (EM), and multiple imputation (MI) methods used to handle missing data on the measurement invariance in data sets with different percentages of missingness?".

General Background

In the literature, Reise, Widaman, and Pugh (1993) investigated the effects of confirmatory factor analysis and item response theory models on the invariance of psychological measures. The actual psychological data collected from Minnesota and China were examined by both methods, and their advantages and disadvantages were investigated. Cheung and Rensvold (2002) investigated how GFI goodness of fit statistic changed in MGCFA, which is generally used in measurement invariance studies. As a result of the invariance study performed in the simulation data consisting of two groups, it was suggested to use ΔCFI, ΔGamma, and ΔMcDonald's indices from 20 different fit indices based on GFI. Chen, Wang, and Chen (2012) conducted a simulation study on data sets with different rates of missingness in order to compare the missing data handling methods in exploratory and confirmatory factor analysis. In the study where six different methods were examined, all the methods produced appropriate results for exploratory and confirmatory factor analyses. It was concluded that the most suitable method for exploratory factor analysis was EM. In the case of less than 20% missing, no statistically significant difference was found between the methods. However, when the missing data is more than 30%, it is suggested to use the SMI and linear trend methods It is seen that studies on measurement invariance are generally based on real data among different groups such as gender and culture (Schnabel, Kelava, Vijver & Seifert, 2015; Wang, Willett & Eccles, 2011).Some of the studies were also used to compare the goodness of fit indices used when examining the measurement invariance (Chen, 2007; Cheung & Rensvold, 2002).

Studies on the effect of missing data on test and item parameters and model data fit (Akbaş & Tavşancıl, 2015; Çüm & Gelbal, 2015; Demir, 2013; Köse, 2014) were conducted. However, there are not many studies about the effect of missing data handling methods on measurement invariance under different conditions. In one of these studies, Selvi, Alıcı & Uzun (2020) examined the effect of EM RI, and SMI methods on measurement invariance on the data obtained from the School Attitude Scale developed by Alıcı (2013) under the condition of 5% missing. Findings of the study show that different methods can change measurement invariance decisions. It has been suggested by the researchers to do more research on different missing data structures and different proportions of missing data.

When the studies related to the missing data handling methods were examined, it is generally aimed to determine which method is more successful in handling missing values (Allison, 2003; Chen, Wang & Chen, 2012; Downey & King, 1998; Olinsky, Chen & Harlow, 2003). The data sets used are generally simulation data, and it is seen that the successful methods change in the data sets with different sample sizes and different percentages of missingness. Missing data studies have recently increased. The problem of missing data is no longer ignored, and efforts are being made to solve the problem.

In this context, it is thought that examining the performance of the missing data handling methods at different missing rates in measurement invariance studies on multi-dimensional structures is important in terms of shedding light on the problem of missing data in measurement invariance studies. Five methods frequently used in researches are discussed within the scope of this study.

METHOD

Participants

The sample was 10857 15-years old students (5109 from Turkey and 5748 from Singapore) who participated in PISA 2015 administration from Turkey and Singapore. Students who have fully responded to items on "enjoyment of science, instrumental motivation, and epistemological beliefs about science" were used in the study. Measurement invariance studies between Turkey and Singapore were conducted on a complete data set of 10857 students in total.

Since PISA results are generally used for cross-country comparisons, it was decided to evaluate the measurement invariance between countries in the data set. It was decided to use Turkey and Singapore data whose mean science score distance from the OECD average is approximately equal in absolute value in terms of mean science score. Singapore has 556 mean science score, Turkey has 425 mean

science score, and OECD average is 493. It is also taken into account that Singapore is the most successful country in terms of average science score. Similarly, the percentage of variation in science performance explained by students' socio-economic status was also considered.

Data Collection Instruments/Data Collection Methods/Data Collection Techniques

The data used in this study was obtained from the PISA 2015 administration organized by OECD and aimed to evaluate the educational systems of countries. PISA is an administration to measure the level of knowledge and skills necessary for students to participate in modern society. In addition to focusing on key areas such as science, mathematics, and reading, the 2015 administration included collaborative problem solving and financial literacy as an innovative field (OECD, 2016).

In this study, the model including the items of enjoyment of science, instrumental motivation, and epistemological beliefs was used. Enjoyment of science is represented by five items, instrumental motivation by four items, and epistemological beliefs by six items. Each item has four response categories, such as strongly disagree, disagree, agree, and strongly agree. Some sample items are shown in the Table 1.

Table 1. Sample Items of the Model

ST094	How much do you disagree or agree with the statements about yourself below? (Please select one response in each row.)				
		Strongly disagree	Disagree	Agree	Strongly agree
ST094Q01NA	I generally have fun when I am learning <broad science> topics.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST094Q02NA	I like reading about <broad science>.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST113	How much do you agree with the statements below? (Please select one response in each row.)				
		Strongly disagree	Disagree	Agree	Strongly agree
ST113Q01TA	Making an effort in my <school science> subject(s) is worth it because this will help me in the work I want to do later on.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST113Q02TA	What I learn in my <school science> subject(s) is important for me because I need this for what I want to do later on.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST131	How much do you disagree or agree with the statements below? (Please select one response in each row.)				
		Strongly disagree	Disagree	Agree	Strongly agree
ST131Q01NA	A good way to know if something is true is to do an experiment.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST131Q03NA	Ideas in <broad science> sometimes change.	$\square_1$	$\square_2$	$\square_3$	$\square_4$

Data Analysis

From the complete data set, 5%, 10%, and 20% of values were deleted randomly on the basis of all cells in the dataset with the help of the R program, and missing data with different percentages of missingness were generated. To determine the mechanism of the missing data in the data sets, Little's MCAR test was performed in each data set. MCAR test was examined separately for each country's datasets. Accordingly, for Turkey p= 0.864 (chi-square=3474.455) in the data set with 5% missing, p= 0.909 (chi-square=8279.206) in the data set with 10% missing, and p= 0.921 (chi-square=21341.920) in the data set with 20% missing were found. For Singapore p= 0.976 (chi-square=3458.673) in the data set with 5% missing, p= 0.990 (chi-square=8840.290) in the data set with 10% missing, and p= 0.645 (chi-

square=23308.247) in the data set with 20% missing were found. Accordingly, it can be said that the missing data in all data sets have MCAR mechanism. Afterwards, LD, SMI, RI, EM, and MI with five imputation methods were applied to each data set to handle the missing data problem, and inter-country measurement invariance was examined by MGCFA approach on completed data sets.

For cross-country measurement invariance, enjoyment of science, instrumental motivation, and epistemological beliefs model is shown in Figure 1.

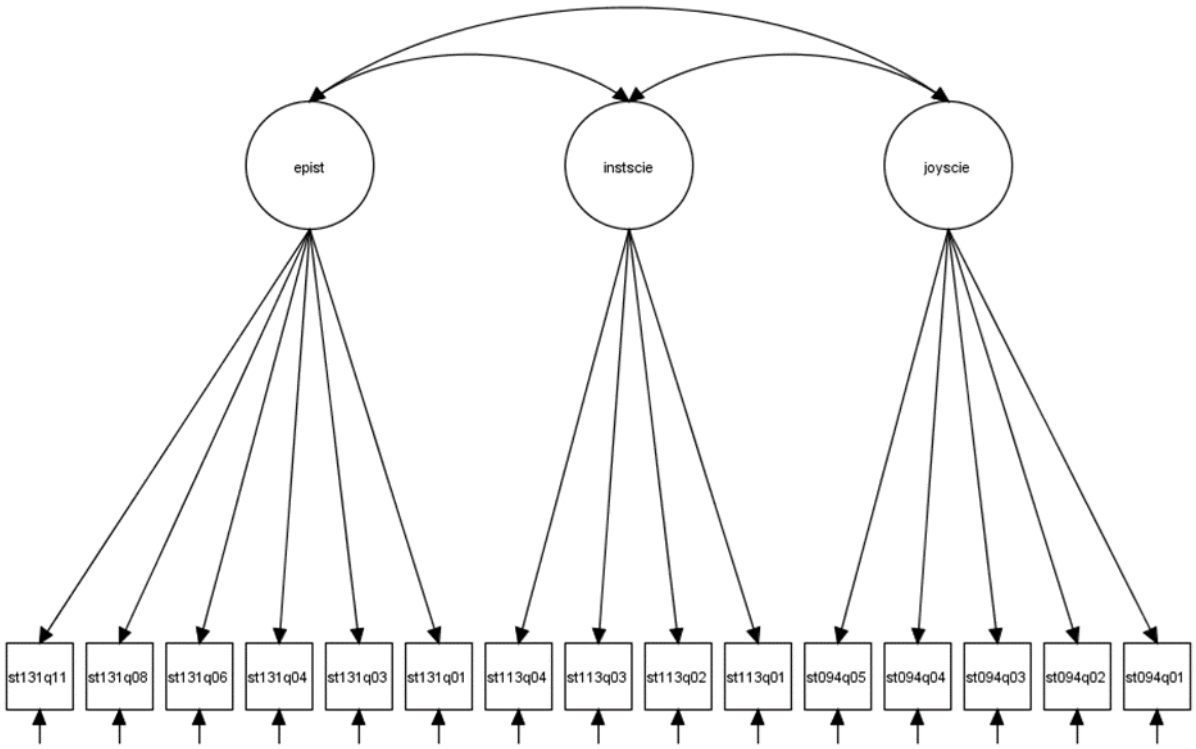


Figure 1. Enjoyment of Science, Instrumental Motivation, and Epistemological Beliefs Model

Before starting the analysis, it is necessary to check the missing values, normality, outliers, and multicollinearity in the data set. The kurtosis and skewness values of each data were examined for normality assumption. According to the findings, the skewness values of the variables ranged from -0.942 to -0.471, and the kurtosis values ranged from -0.296 to 0.913. Tabachnick and Fidell (2013) stated that the closeness of kurtosis and skewness values to zero shows that the distribution is close to normal distribution. According to obtained kurtosis and skewness values, it can be said that each variable was distributed normally. To determine the outliers, z distributions were examined. $|z|>3.29$ indicates that the variable contains outliers (Tabachnick & Fidell, 2013). According to the findings, z scores of the variables ranged between -2.78 and 1.42. In this case, it can be concluded that there are no outliers in the data set. VIF and tolerance values were examined to determine if there was a multicollinearity problem. VIF values ranged between 2.178 and 4.882, and the tolerance values ranged between 0.205 and 0.459. Based on this finding, it was concluded that there is no multicollinearity problem in the data set.

In order to compare the results obtained from a measurement instrument applied to groups with different characteristics, it is important to ensure the measurement invariance between groups. There are different approaches to test measurement invariance, such as MGCFA and item response theory. In this study, the measurement invariance was examined with the MGCFA approach with ML estimator. MGCFA aims to compare the means, variance, and covariance of the latent variable between the groups while

testing the measurement invariance (Asparouhov & Muthen, 2014). In this context, configural invariance, metric invariance, scalar invariance, and strict invariance were tested hierarchically. ΔCFI was examined to determine whether measurement invariance was provided at each stage. A difference of less than .01 supports the less parameterized model (Chung et al., 2016).

RESULTS

In this section, the findings of the research are given. Firstly, the reference values to compare the data sets with different percentages of missingness were obtained by performing a hierarchical measurement invariance in the complete data set.

Before moving on to measurement invariance studies in the whole data set, confirmatory factor analysis was performed in Turkey and Singapore datasets separately, and model-data fits were examined. Fit indices obtained from Turkey and Singapore datasets are presented in Table 2.

Table 2. Fit Indices in the Singapore and Turkey Data Sets

	χ^2	df	χ^2/df	SRMR	RMSEA	CFI	TLI
Singapore	3546.635	87	40.766	.033	.083	.949	.938
Turkey	2036.208	87	23.405	.022	.066	.968	.961

When Table 2 is examined, it is seen that the data for both countries fit the model. After that, a cross-country measurement invariance study was conducted for the complete data set, and reference values were obtained. Reference values obtained from the complete data set are provided in Table 3.

Table 3. Fit Indices in The Complete Data Set

	χ^2	df	χ^2/df	SRMR	RMSEA	CFI	TLI	ΔCFI
Configural	5582.843	174	32.085	.028	.076	.958	.949	
Metric	5723.250	186	30.770	.034	.074	.957	.951	-.001
Scalar	6222.092	198	31.425	.038	.075	.953	.950	-.005
Strict	11469.299	216	53.099	.226	.098	.912	.914	-.046

When the fit indices in the Table 1 were examined, it was seen that configural invariance, metric invariance, and scalar invariance were achieved in the complete data set, but not the strict invariance ($|\Delta CFI| \leq .01$). The values related to fit indices from the reference data set was used to compare with the completed data sets. Then, the results of the measurement invariance studies were included in the data sets with 5% missing, 10% missing, and 20% missing and completed with LD, SMI, RI, EM, and MI methods.

Influence of Missing Data Handling Methods on Measurement Invariance in the Data Set with 5% Missing

The data set with 5% missing was completed with LD, SMI, RI, EM, and MI methods, and measurement invariance was hierarchically tested on completed data sets. The fit indices obtained at each stage of measurement invariance according to different methods are provided in Table 4.

Table 4. Fit Indices in the Data Set with 5% Missing and Completed with the Methods

Method	*Invariance*	χ^2	*df*	χ^2/df	*SRMR*	*RMSEA*	*CFI*	*TLI*	ΔCFI
LD	Configural	2982.518	174	17.141	.030	.080	.953	.944	
	Metric	3055.602	186	16.428	.035	.078	.952	.946	-.001
	Scalar	3301.304	198	16.673	.039	.079	.948	.945	-.005
	Strict	5661.534	216	26.211	.222	.100	.909	.912	-.044
SMI	Configural	4363.479	174	25.077	.028	.067	.962	.955	
	Metric	4488.792	186	24.133	.033	.065	.961	.956	-.001
	Scalar	4915.130	198	24.824	.037	.066	.958	.955	-.004
	Strict	10062.546	216	46.586	.228	.092	.912	.914	-.050
RI	Configural	5543.385	174	31.856	.028	.075	.958	.949	
	Metric	5661.530	186	30.438	.033	.074	.957	.952	-.001
	Scalar	6139.553	198	31.008	.036	.074	.954	.951	-.004
	Strict	10845.374	216	50.210	.221	.095	.917	.919	-.041
EM	Configural	6153.287	174	35.364	.028	.080	.955	.946	
	Metric	6275.856	186	33.741	.033	.078	.954	.949	-.001
	Scalar	6766.297	198	34.173	.037	.078	.951	.948	-.004
	Strict	11998.191	216	55.547	.228	.100	.912	.914	-.043
MI	Configural	5413.041	174	31.109	.028	.074	.959	.950	
	Metric	5531.746	186	29.741	.033	.073	.958	.952	-.001
	Scalar	6002.515	198	30.316	.036	.073	.954	.951	-.005
	Strict	10786.742	216	49.939	.224	.095	.916	.919	-.040

When the fit indices in the tables were examined, it was seen that the first three stages of measurement invariance between countries were achieved in all data sets, but not the strict invariance ($|\Delta CFI| \leq .01$). When the fit indices obtained for each method were compared with the reference values given in Table 1, it was observed that the indices obtained from SMI, RI, EM, and MI methods gave more similar results to the reference values. But dissimilarly, LD and SMI methods showed χ^2/df less than the reference value. All indices, especially ΔCFI, were compared with the reference data set. Methods giving more similar results to the reference values were determined. RI and MI methods yielded the closest results.

Influence of Missing Data Handling Methods on Measurement Invariance in the Data Set with 10% Missing

The data set with 10% missing was completed with LD, SMI, RI, EM, and MI methods. Measurement invariance was hierarchically tested on completed data sets. The fit indices obtained at each stage of measurement invariance according to different methods are provided in Table 5.

Table 5. Fit Indices in the Data Set with 10% Missing and Completed the Methods

Method	Invariance	χ^2	df	χ^2/df	SRMR	RMSEA	CFI	TLI	ΔCFI
LD	Configural	1423.852	174	8.183	.035	.080	.950	.940	
	Metric	1444.182	186	7.764	.038	.078	.950	.944	.000
	Scalar	1526.376	198	7.709	.041	.078	.947	.944	-.003
	Strict	2786.312	216	12.900	.245	.103	.898	.901	-.052
SMI	Configural	3186.870	174	18.315	.025	.056	.969	.963	
	Metric	3275.686	186	17.611	.030	.055	.968	.964	-.001
	Scalar	3625.250	198	18.309	.033	.056	.965	.963	-.004
	Strict	8700.191	216	40.279	.232	.085	.913	.915	-.056
RI	Configural	4943.032	174	28.408	.027	.071	.962	.955	
	Metric	5060.230	186	27.206	.032	.069	.961	.957	-.001
	Scalar	5433.588	198	27.442	.035	.070	.959	.956	-.003
	Strict	9705.969	216	44.935	.217	.090	.925	.927	-.037
EM	Configural	6318.480	174	36.313	.028	.081	.956	.947	
	Metric	6446.346	186	34.658	.033	.079	.955	.949	-.001
	Scalar	6873.543	198	34.715	.036	.079	.952	.949	-.004
	Strict	12000.284	216	55.557	.230	.100	.916	.918	-.040
MI	Configural	4898.776	174	28.154	.027	.071	.962	.954	
	Metric	5015.456	186	26.965	.032	.069	.961	.956	-.001
	Scalar	5407.393	198	27.310	.035	.070	.958	.956	-.004
	Strict	9761.137	216	45.190	.222	.090	.923	.926	-.039

When the fit indices in the tables were examined, it was seen that all the missing data handling methods are provided all the invariance stages except strict invariance as in reference data set (|ΔCFI|≤.01). When the fit indices obtained for each method were compared with the reference values given in Table 1, it was seen that the EM method gives results very close to the reference values. Dissimilarly, LD and SMI methods showed χ^2/df less than the reference value. And the SMI method showed CFI and TLI values to be more than they were.

Influence of Missing Data Handling Methods on Measurement Invariance in the Data Set with 20% Missing

The data set with 20% missing was completed with LD, SMI, RI, EM,and MI methods, and measurement invariance between countries was hierarchically tested on completed data sets. The fit indices obtained from the measurement invariance studies are provided in Table 6.

Table 6. Fit Indices in the Data Set with 20% Missing and Completed with the Methods

Method	Invariance	χ^2	df	χ^2/df	SRMR	RMSEA	CFI	TLI	ΔCFI
LD	Configural	417.859	174	2.401	.043	.085	.948	.938	
	Metric	425.802	186	2.289	.049	.082	.949	.943	.001
	Scalar	448.435	198	2.265	.051	.081	.947	.944	-.001
	Strict	694.988	216	3.218	.199	.107	.899	.902	-.049
SMI	Configural	2168.515	174	12.463	.023	.046	.973	.968	
	Metric	2227.010	186	11.973	.027	.045	.973	.969	.000
	Scalar	2484.778	198	12.549	.030	.046	.969	.967	-.004
	Strict	7600.786	216	35.189	.239	.079	.901	.904	-.072
RI	Configural	4736.906	174	27.224	.026	.070	.963	.956	
	Metric	4831.875	186	25.978	.030	.068	.963	.958	.000
	Scalar	5153.266	198	26.027	.033	.068	.960	.958	-.003
	Strict	8454.797	216	39.143	.207	.084	.934	.936	-.029
EM	Configural	7818.130	174	44.932	.028	.090	.950	.940	
	Metric	7928.744	186	42.628	.032	.088	.949	.943	-.001
	Scalar	8317.908	198	42.010	.035	.087	.947	.944	-.003
	Strict	13517.876	216	62.583	.234	.107	.913	.916	-.037
MI	Configural	4916.012	174	28.253	.026	.071	.961	.953	
	Metric	4995.423	186	26.857	.030	.069	.960	.955	-.001

Scalar	5328.987	198	26.914	.033	.069	.958	.955	-.003
Strict	8937.861	216	41.379	.216	.086	.928	.930	-.033

When the fit indices in the tables were examined, it was seen that all the missing data handling methods provided all the invariance stages except strict invariance (|ΔCFI|≤.01). When the fit indices obtained for each method were compared with the reference values given in Table 1, it was seen that the MI method gives results close to the reference values. The MI method shows χ^2/df close to the reference value, but dissimilarly, LD and SMI methods shows χ^2/df lower than it is, and the EM method shows χ^2/df higher than it is.

DISCUSSION and CONCLUSION

In this study, the effect of completing data sets with missing values with LD, SMI, RI, EM, and MI methods on measurement invariance was investigated. As a result of measurement invariance studies between countries performed in data sets completed with different missing data handling methods in all missing percentages, it was observed that all the invariance stages except strict invariance were provided in accordance with the complete data set. Although the data sets were completed with different methods, there was no result that would show the measurement invariance between countries different from the reference data set.

The research was limited in terms of missing data handling methods, missing data mechanisms, and measurement invariance approaches. LD, SMI, RI, EM, and MI methods were used as missing data handling methods. The data sets have MCAR mechanism. Data sets with multi-dimensional structures were used in the study. And, measurement invariance was handled by MG-CFA approach. The findings and discussion in this study are based on a single data set obtained from the PISA 2015 administration. No replication was done in the study. Please consider this situation as a limitation.

In the literature, methods based on the likelihood approach and the multiple imputation approach are proposed as the strategy of handling the missing data in CFA models (Allison, 2003; Brown, 2006). The findings of the research show that EM and MI methods which are based on the likelihood approach yielded more successful results in accordance with the literature.

Selvi, Alıcı & Uzun (2020) tested the measurement invariance with structural equation modeling in the complete data matrix and in cases of handling the missing data tested using EM, Regression-Based Imputation, and Mean Substitution methods. They concluded that different methods can change the decisions of measurement invariance. But, in the findings of this study it was seen that not all methods change measurement invariance decisions.

Allison (2003) stated that MI has good statistical properties, and it can be used in almost any situation. Schafer and Graham (2002) recommended EM algorithm for maximum likelihood and MI method. Similar to the studies, the results obtained from EM and MI methods were found to be more appropriate to reference data in this study.

As a result of comparing the fit indices obtained from each data set with the fit indices obtained from the complete data set, the data sets completed with RI and MI in the data set with 5% missing yielded closer results to the reference values. In the data set with 10% missing, closer results were obtained from the EM method than the other methods. And in the data set with 20% missing, the missing data handling method which gave the closest results to the reference values was MI. While making comparisons, based on ΔCFI change, the methods whose fit indices give the closest results to the reference values were determined descriptively. As a result of the research, recommendations for implementation are as follows: In the measurement invariance studies to be performed in multi-dimensional data sets, data sets with 5% missing can be completed by RI and MI methods. The EM method works better than other methods if there are around 10% missing. And, if the data set has about 20% missing, the MI method can be used to complete the data set.

REFERENCES

Akbaş, U., & Tavşancıl, E. (2015). Farklı örneklem büyüklüklerinde ve kayıp veri örüntülerinde ölçeklerin psikometrik özelliklerinin kayıp veri baş etme teknikleri ile incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 6*(1), 38-57.

Alıcı, D. (2013). Okula yönelik tutum ölçeği'nin geliştirilmesi: Güvenirlik ve geçerlik çalışması. *Eğitim ve Bilim*, 38(268), 318-331.

Allison, P. D. (2001). *Missing data.* Thousand Oaks, CA: Sage.

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, *112*(4), 545-557.

Asparouhov, T., & Muthen, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 1-14.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464-504.

Chen, S. F., Wang, S., & Chen, C. Y. (2012). A simulation study using EFA and CFA programs based the impact of the missing data on test dimensionality. *Expert Systems with Applications*, *39*(4), 4026-4031.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233-255.

Chung, H., Kim, J., Park, R., Bamer, A. M., Bocell, F. D., & Amtmann, D. (2016). Testing the measurement invariance of the University of Washington Self-Efficacy Scale short form across four diagnostic subgroups. *Qual Life Res.*, *25*(10), 2559-2564.

Çüm, S. & Gelbal, S. (2015). Kayıp veriler yerine yaklaşık değer atamada kullanılan farklı yöntemlerin model veri uyumu üzerindeki etkisi. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi,* 35, 87-111.

Demir, E. (2013). Kayıp verilerin varlığında çoktan seçmeli testlerde madde ve test parametrelerinin kestirilmesi: SBS örneği. *Eğitim Bilimleri Araştırmaları Dergisi, 3*(2), 47-68.

Downey, R. G. & King, C. V. (1998). Missing data in likert ratings: A comparison of replacement methods. *The Journal of General Psychology*, *125*(2), 175-191.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin*, *95*(1), 134-135.

Enders, C. K. (2010). *Applied missing data analysis.* New York: The Guilford Press.

Harrington, D. (2009). *Confirmatory factor analysis*. New York: Oxford University Press.

Köse, A. (2014). The effect of missing data handling methods on goodness of fit indices in confirmatory factor analysis. *Educational Research and Reviews, 9*(8), 208-215.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical analysis with missing data. (2nd edition)*. New York: Wiley

OECD (Organization for Economic Cooperation and Development) (2016). *PISA 2015 results in focus*. Retrieved February 12, 2017, from https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

Olinsky, A., Chen, S., & Harlow, L. (2003). The comparative efficacy of imputation methods for missing data in structural equation modeling. *European Journal of Operational Research*, *151*(1), 53-79.

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

Schafer, J. L. & Graham, J. W. (2002) Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147-177.

Schnabel, D. B. L., Kelava, A., Vijver, F. J. R., & Seifert, L. (2015). Examining psychometric properties, measurement invariance, and construct validity of a short version of the Test to Measure Intercultural Competence (TMIC-S) in Germany and Brazil. *International Journal of Intercultural Relations, 49*, 137-155.

Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, 9(4), 486-492.

Selvi, H., Alıcı, D., & Uzun, N. B. (2020). Investigating measurement invariance under different missing value reduction methods. *Asian Journal of Education and Training, 6*(2), 237-245.

Tabachnick, B. G. & Fidell, L. S. (2013). *Using multivariate statistics*. (6th edition). Boston: Pearson.

Wang, M., Willett, J. B., & Eccles, J. S. (2011). The assessment of school engagement: Examining dimensionality and measurement invariance by gender and race/ethnicity. *Journal of School Psychology*, *49*(4), 465-480.

Whitaker, B. G., & Mckinney, J. L. (2007). Assessing the measurement invariance of latent job satisfaction ratings across survey administration modes for respondent subgroups: A MIMIC modeling approach. *Behavior Research Methods, 39*(3), 502-509.

Xu, H., & Tracey, T. J. G. (2017). Use of multi-group confirmatory factor analysis in examining measurenet invariance in counseling psychology research. *The European Journal of Counselling Psychology*, *6*(1), 75-82.

Çok Boyutlu Yapılarda Kayıp Veri ile Baş Etme Yöntemlerinin Ölçme Değişmezliğine Etkisi Açısından Karşılaştırılması

Giriş

Ölçme araçları, eğitim sistemleri içerisinde büyük öneme sahiptir. İhtiyaca uygun nitelikli insan gücü yetiştirmek için bireylerin eğitim kurumlarına yerleştirilmesi, eğitim sistemlerinde ihtiyaca uygun olarak değişiklikler ve geliştirmeler yapılabilmesi ölçme araçlarından elde edilen bulgular sonucunda yapılabilmektedir. Ulusal ve uluslararası düzeyde yapılan ölçme ve değerlendirme çalışmaları sonucunda, ülkeler eğitim politikalarında dahi değişikliklere gidebilmektedir. Özellikle uluslararası karşılaştırmaların yapılmasına olanak sağlayan büyük ölçekli ölçme ve değerlendirme çalışmalarının sonuçları, eğitimin tüm paydaşları tarafından ilgiyle takip edilmektedir.

Uluslararası düzeyde uygulanan PISA (Programme for International Student Assessment) ve TIMSS (Trends in International Mathematics and Science Study) projeleri, ülkeler arası karşılaştırma yapılmasını amaçlamaktadır. Özellikle uluslararası düzeyde yapılan ve sonucunda karşılaştırma yapılan sınavlarda, elde edilen sonuçların karşılaştırılabilir olması, ülkeler arası değerlendirmelerde büyük önem taşımaktadır. Aynı ölçme aracının uygulandığı, özellikleri birbirinden farklı gruplardan elde edilen bulguların yorumlanabilmesi için, ölçme aracının bütün gruplar için aynı anlama gelmesi gerekmektedir. Bu bağlamda karşımıza ölçme değişmezliği kavramı ortaya çıkmaktadır. Drasgow (1984) ölçme değişmezliğini "gözlenen test puanları ile gizil özelliklerin arasındaki ilişkinin tüm alt gruplar arasında benzer olması" şeklinde tanımlamıştır.

Ölçme araçları tarafından elde edilen veriler her zaman eksiksiz şekilde elde edilememektedir. Yanıtlayıcıdan, ölçme aracından veya uygulayıcıdan kaynaklanan sebeplerden dolayı veri setlerinde, bazı değişkenlerde kayıp veriler bulunabilmektedir. Kayıp veriler, veri setleri üzerinde yapılan istatistiksel işlemlerin sonuçlarını doğrudan etkileyen önemli bir problemdir.

Bu araştırmanın amacı, kayıp veri ile baş etme yöntemlerinin ölçme değişmezliğine etkisi açısından karşılaştırılmasıdır. Yapılan ölçme değişmezliği çalışmalarında kayıp veri durumu, analizlere başlamadan önce kontrol edilmesi ve çözülmesi gereken bir problemdir. Kayıp verilerin varlığı, doğrulayıcı faktör analizi de dahil olmak üzere birçok analizin sonucunu etkileyebilir. Kayıp verilerin veri setinden çıkarılması örneklemi küçülteceğinden, elde edilen sonuçların evrene genellenebilme gücü azalır. Ayrıca kayıp verilerin varlığı, analizlerden elde edilen anlamlılık değerlerini etkileyerek tip I ve tip II hataların oluşmasına sebep olabilir. Kayıp veri problemini çözmek için kullanılan yöntemlerin farklılığı bile, analizlerden farklı bulgular elde edilmesine neden olabilir (Harrington, 2009).

Kültür, etnik köken, dil, cinsiyet gibi farklı gruplardan bireylerin karşılaştırılmasında kullanılan testlerin öncelikle ölçme değişmezliğinin sağlanması gerekmektedir. Ölçme değişmezliği analizlerinden elde edilen bulguların doğru bir şekilde yorumlanabilmesi için ise kayıp veri probleminin çözülmesi gerekmektedir. Farklı kayıp veri oranlarına bağlı olarak, kayıp veri ile baş etme yöntemlerinin karşılaştırıldığı bu araştırmadan elde edilen bulgularla, yapılacak olan ölçme değişmezliği çalışmalarında veri setine uygun olan kayıp veri ile baş etme yönteminin seçilebilmesi amaçlanmıştır. Bu bağlamda "Kayıp veriler ile baş etmede kullanılan dizin silme (DS), seri ortalaması atama (SO), regresyon atama (RA), beklenti maksimizasyonu (BM) ve çoklu atama (ÇA) yöntemlerinin, farklı oranlarda kayıp içeren veri setlerinde ölçme değişmezliğine etkisi ne düzeydedir?" problemine yanıt aranmaktadır.

Alan yazın incelendiğinde, Reise, Widaman ve Pugh (1993), doğrulayıcı faktör analizi ve madde tepki kuramı modellerinin, psikolojik ölçmelerin değişmezliğine etkilerini araştırmışlardır. Minesota ve Çin'den toplanan gerçek psikolojik veriler her iki yöntemle de incelenmiş ve yöntemlerin avantaj ve dezavantajları araştırılmıştır. Cheung ve Rensvold (2002), ölçe değişmezliği çalışmalarında genellikle

kullanılan çok gruplu doğrulayıcı faktör analizinde, GFI uyum iyiliği istatistiğinin ne şekilde değiştiğini araştırmışlardır. İki gruptan oluşan simülasyon verisinde gerçekleştirilen değişmezlik çalışmasının sonucunda, GFI indeksini temel alan 20 farklı uyum indeksinden, ΔCFI, ΔGamma ve ΔMcDonald's indekslerinin kullanılması önerilmiştir. Chen, Wang ve Chen (2012), açımlayıcı ve doğrulayıcı faktör analizinde kayıp veri yöntemlerini karşılaştırmak amacıyla, farklı oranlarda kayıp içeren veri setleri üzerinde simülasyon çalışması yapmıştır. Altı farklı yöntemin incelendiği çalışmada, tüm yöntemler açımlayıcı ve doğrulayıcı faktör analizinde modele uygun sonuçlar üretmiştir. Açımlayıcı faktör analizi için en uygun yöntemin beklenti maksimizasyonu olduğu sonucuna ulaşılmıştır. %20'nin altında kayıp olması durumunda yöntemler arasında istatistiksel olarak anlamlı bir fark bulunamamıştır. Ancak eksik veriler %30'dan fazla olduğunda, seri ortalaması atama yöntemi ve doğrusal eğim yöntemi kullanılması önerilmiştir.

Dünyada ve Türkiye'de ölçme değişmezliği ile ilgili yapılan çalışmalara genel olarak bakıldığında, çalışmaların genelinde gerçek veriler kullanılarak cinsiyet ve kültür gibi farklı gruplar arasında ölçme değişmezliğinin sağlanıp sağlanmadığıyla ilgili olduğu görülmektedir (Schnabel, Kelava, Vijver ve Seifert, 2015; Wang, Willett ve Eccles, 2011;). Bir kısım araştırmaların da, ölçme değişmezliği incelenirken kullanılan uyum iyiliği katsayılarının karşılaştırılmasına yönelik olduğu görülmüştür (Chen, 2007; Cheung ve Rensvold, 2002;).

Kayıp veri atama yöntemleri ile ilgili olan çalışmalara bakıldığında ise, çalışmalar genellikle kayıp verilerin tamamlanmasında hangi yöntemin daha başarılı olduğunu belirlemeye yöneliktir (Allison, 2003; Chen, Wang & Chen, 2012; Downey & King, 1998; Olinsky, Chen & Harlow, 2003). Kullanılan veri setleri genellikle simülasyon verileri olup, başarılı yöntemlerin, farklı örneklem büyüklüklerinde ve farklı oranlarda kayıp içeren veri setlerinde değiştiği görülmektedir. Kayıp veri çalışmaları son zamanlarda artmıştır. Kayıp veri sorunu, artık göz ardı edilmeyerek, problemin çözümüne yönelik çalışmalar yapılmaktadır. Bu araştırmada, kayıp veri atama yöntemlerinin ölçme değişmezliğine etkisi araştırılmaktadır. Alan yazında yapılacak ölçme değişmezliği çalışmalarında, farklı örneklem büyüklüklerinde ve farklı oranlardaki kayıp verilerde, kayıp veri probleminin çözümüne yönelik öneriler getirmek amaçlanmaktadır.

Yöntem

Bu araştırmada, farklı kayıp veri ile baş etme yöntemleri ile tamamlanmış veri setlerinde ölçme değişmezliği çalışması yapılmıştır. Çalışmanın amacı, DS, SO, RA, BM ve ÇA yöntemlerinin çok boyutlu yapılarda ölçme değişmezliğine etkisini incelemektir.

Araştırmanın örneklemini PISA 2015 uygulamasına Türkiye ve Singapur'dan katılmış 12010 (Türkiye=5895, Singapur=6115) öğrenciden, fen okuryazarlığına ilişkin duyuşsal özellikler ile ilgili maddelere eksiksiz yanıt vermiş 10857 (Türkiye=5109, Singapur=5748) öğrenci oluşturmaktadır.

Araştırmada, 5496 kişilik eksiksiz veri setinde ülkeler arası ölçme değişmezliği çalışmaları yapılmıştır.

Oluşturulan eksiksiz veri setinden, hücre bazında %5, %10 ve %20 oranında rastgele değerler R programı yardımıyla silinmiş ve kayıp veriler oluşturulmuştur. Oluşturulan veri setlerinde bulunan kayıp verilerin mekanizmasının belirlenebilmesi için, her veri setinde Little'ın TROK testi gerçekleştirilmiştir. Türkiye örneklemi için %5 kayıp içeren veri setinde p= 0,864 (ki-kare=3474,455), %10 kayıp içeren veri setinde p= 0,909 (ki-kare=8279,206) ve %20 kayıp içeren veri setinde p= 0, 921 (ki-kare=21341,920) bulunmuştur. Singapur için %5 kayıp içeren veri setinde p= 0, 976 (ki-kare=3458,673), %10 kayıp içeren veri setinde p= 0, 990 (ki-kare=8840,290) ve %20 kayıp içeren veri setinde p= 0, 645 (ki-kare=23308,247) bulunmuştur. Buna göre tüm veri setindeki kayıp verilerin TROK mekanizmasına sahip olduğu söylenebilir.

Daha sonra, her bir veri setinde DS, SO, RA, BM ve ÇA yöntemleri uygulanmış ve ülkeler arası ölçme değişmezliği çalışmaları çok gruplu doğrulayıcı faktör analizi (ÇGDFA) yaklaşımı ile incelenmiştir.

Sonuç ve Tartışma

Araştırmada kayıp veri içeren veri setlerinin, DS, SO, RA, BM ve ÇA yöntemleriyle tamamlanmasının, ölçme değişmezliğine etkisi araştırılmıştır. Tüm oranlarda, farklı yöntemlerle tamamlanmış veri setlerinde yapılan ülkeler arası ölçme değişmezliği çalışmalarının sonucunda, eksiksiz veri setine uygun olarak katı değişmezlik dışındaki tüm değişmezlik aşamalarının sağlandığı görülmüştür. Veri setlerinde, farklı kayıp veri ile baş etme yöntemleri ile tamamlansa da, ülkeler arası ölçme değişmezliğini referans veri setinden farklı gösterecek bir sonuç bulunmamıştır.

Araştırma, kayıp veri ile baş etme yöntemleri, kayıp veri mekanizması ve ölçme değişmezliği yaklaşımı açısından sınırlandırılmıştır. Kayıp veri ile baş etme yöntemlerinden dizin silme (DS), seri ortalaması atama (SO), regresyon atama (RA), beklenti maksimizasyonu (BM) ve çoklu atama (ÇA) yöntemlerine yer verilmiştir. Veri setleri, kayıp veri mekanizmalarından tamamen rassal olarak kayıp (TROK) mekanizmasına sahiptir. Çalışmada birden fazla faktöre sahip veri setleri kullanılmıştır. Ölçme değişmezliği çok gruplu doğrulayıcı faktör analizi yaklaşımıyla ele alınmıştır.

Her bir veri setinden elde edilen uyum katsayılarının, eksiksiz veriden elde edilen uyum katsayıları ile karşılaştırılması sonucunda, %5 kayıp içeren veri setinde RA ve ÇA ile tamamlanan veri setleri, referans değerlere daha yakın sonuçlar vermiştir. %10 kayıp içeren veri setinde, BM yönteminden diğer yöntemlere göre daha yakın sonuçlar elde edilmiştir. %20 kayıp içeren veri setinde ise referans değere en yakın sonuç veren kayıp veri ile baş etme yöntemi ÇA olmuştur.

Araştırma sonucunda uygulamaya yönelik olarak öneriler şu şekildedir:

Çok boyutlu veri setlerinde yapılacak olan ölçme değişmezliği çalışmalarında, %5 civarında kayıp içeren veri setleri RA veya ÇA yöntemi ile tamamlanabilir. %10 civarında kayıp veri bulunuyorsa BM yöntemi diğer yöntemlere göre daha iyi sonuç vermektedir. Kayıp veri miktarı %20 civarında ise, ÇA yöntemi kayıp verileri tamamlamak için kullanılabilir.