Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309-6575 *Yaz* 2023 *Summer* 2023 *Cilt: 14-Sayı: 2* *Volume: 14-Issue: 2*

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 yayımlanan hakemli uluslararası bir dergidir. Yayımlanan yazıların tüm sorumluğu ilgili yazarlara aittir.

Journal of Measurement and Evaluation in Education and Psychology (JMEEP) is a international 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), EBSCO

Hakem Kurulu / Referee Board

Hikmet ŞEVGİN (Van Yüzüncü Yıl Üni.)
Hülya KELECİOĞLU (Hacettepe Üni.)
Hülya YÜREKLI (Yıldız Teknik Üni.)
İbrahim Alper KÖSE (Bolu Abant İzzet Baysal Üni.)
İbrahim YILDIRIM (Gaziantep Üni.)
İbrahim UYSAL (Bolu 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 ERTUNA (Sakarya Üni.)
Levent YAKAR (Kahramanmaraş Sütçü İmam Üni.)
Mehmet KAPLAN (MEB)
Mehmet ŞATA (Ağrı İbrahim Çeçen Üni.)
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.)
Özen YILDIRIM (Pamukkale Üni.)
Özge ALTINTAS (Ankara Üni.)
Özge BIKMAZ BİLGEN (Adnan Menderes Üni.)
Özlem ULAŞ (Giresun Üni.)
Recep GÜR (Erzincan Üni.)
Ragıp TERZİ (Harran Üni.)
Sedat ŞEN (Harran Üni.)
Recep Serkan ARIK (Dumlupınar Üni.)
Safiye BİLİCAN DEMİR (Kocaeli Üni.)
Selahattin GELBAL (Hacettepe Üni.)
Seher YALÇIN (Ankara Üni.)
Selen DEMİRTAŞ ZORBAZ (Ordu Üni.)
Selma ŞENEL (Balıkesir Üni.)
Seçil ÖMÜR SÜNBÜL (Mersin Üni.)
Sait Çüm (MEB)
Sakine GÖÇER ŞAHİN (University of Wisconsin Madison)
Sema SULAK (Bartın Üni.)
Semirhan GÖKÇE (Niğde Ömer Halisdemir Üni.)
Serkan ARIKAN (Boğaziçi Üni.)
Seval KIZILDAĞ ŞAHİN (Adıyaman Üni.)
Sevda ÇETİN (Hacettepe Üni.)
Sevilay KİLMEN (Abant İzzet Baysal Üni.)
Sinem Evin AKBAY (Mersin Üni.)
Sungur GÜREL (Siirt Üni.)
Süleyman DEMİR (Sakarya Üni.)
Sümeyra SOYSAL (Necmettin Erbakan Üni.)
Şeref TAN (Gazi Üni.)
Şeyma UYAR (Mehmet Akif Ersoy Üni.)
Tahsin Oğuz BAŞOKÇU (Ege Üni.)
Terry A. ACKERMAN (University of Iowa)
Tuğba KARADAVUT (İzmir Demokrasi Üni.)
Tuncay ÖĞRETMEN (Ege Üni.)
Tülin ACAR (Parantez Eğitim)
Türkan DOĞAN (Hacettepe Üni.)
Ufuk AKBAŞ (Hasan Kalyoncu Üni.)
Wenchao MA (University of Alabama)
Yavuz AKPINAR (Boğaziçi Üni.)
Yeşim ÖZER ÖZKAN (Gaziantep Üni.)
Yusuf KARA (Southern Methodist University)
Zekeriya NARTGÜN (Bolu Abant İzzet Baysal Üni.)
Zeynep ŞEN AKÇAY (Hacettepe Üni.)

*Ada göre alfabetik sıralanmıştır. / Names listed in alphabetical order.

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Haziran 2023, Sayı: 14-2)

Journal of Measurement and Evaluation in Education and Psychology (June 2023, Issue: 14-2)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(2); 106-117

A Comparison of Different Designs in Scoring of PISA 2009 Reading Open Ended Items According to Generalizability Theory

Meral ALKAN* Nuri DOĞAN**

Abstract

This study compares the different designs obtained through four raters' scoring the open-ended items used in PISA 2009 reading literacy altogether or alternately according to the Generalizability Theory. The sample of the research was composed of 362 students (out of 4996 students participating in PISA 2009) who responded to the items of reading skills and who were scored by more than one rater. Two designs were created so as to be used in generalizability theory in the study. One of them was the crossed design symbolized as "s x i x r" (student x item x rater), in which students are scored by each rater in terms of the same skills. The second was the nested design symbolized as "(r:s) x i", where each rater scored only a group of students and raters are nested in students and the items were crossed with these variables. On comparing the s x i x r design with (r:s) x i design, it was found that the relative and absolute error variances estimated for (r:s) x i design were smaller than those for s x i x r design and that therefore the G and Phi coefficients took on bigger values. On increasing the number of raters in both designs, the G and Phi coefficients also increased in the D study. While acceptable values of G and Phi coefficients were reached on reducing the number of raters by half in Booklet 2, raising the number of raters seemed more appropriate in Booklet 8.

Keywords: Generalizability theory, reliability, G study, D study, PISA 2009

Introduction

At the beginning of the 21st century, social, economic, and technological developments have caused rapid change in every field. The desire of societies to keep up with this change has brought the issue of the quality of education to the fore. The quality of education is the most important factor in equipping new generations with new skills and competencies to keep up with this change. In today's world, where knowledge is accepted as a power and spreads rapidly, raising individuals who think critically, question, are responsible for their own learning, creative, and ready for life has become the most important goal of education systems. This situation affected assessment practices as well as educational practices. If subject knowledge alone is not a sufficient criterion, tests based on choosing the correct answer among the given options are not sufficient on their own. This understanding brought to learning has revealed the necessity of organizing the tests in a form in which the individual can structure their own answers and the curriculum from low-level thinking to an understanding that requires high-level thinking; teaching methods and techniques from a teacher-centered structure to a student-centered structure; assessment and evaluation approaches, on the other hand, have transformed from a structure that measures the extent to which information is acquired, to a structure that measures how information can be used in new situations or in real life (Biemer, 1993). This situation has been the trigger for turning to different approaches in the teaching process. OECD PISA (Programme for International Student Assessment) tests are designed to assess how well students, at the end of compulsory education, can

* Lect. PhD., Gazi University, Rectorate, Ankara-Türkiye, meralalkan@gazi.edu.tr, ORCID ID: 0000-0001-9497-3660
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Türkiye, nurid@hacettepe.edu.tr, ORCID ID: 0000-0001-6274-2016

To cite this article:
Alkan, M., & Doğan N. (2023). A comparison of different designs in scoring of PISA 2009 reading open ended items according to generalizability theory. *Journal of Measurement and Evaluation in Education and Psychology, 14*(2), 106-117. https://doi.org/10.21031/epod.1210917

Received: 28.11.2022
Accepted: 12.06.2023

apply their knowledge to real-life situations and can, therefore, fully participate in society (OECD, 2017; EARGED, 2010), and Turkey has been taking part in PISA and other international studies such as TIMSS and PIRLS assessing students' achievement comparatively on an international scale since the early 2000s. Turkish students cannot attain the desired success level in open-ended items in those studies (Balbağ, Leblebicier, Karaer, Sarıkahya & Erkan, 2016). The type of questions in which Turkish students attained the highest percentage of success level was multiple choice items in the sub-fields of reading skills and science literacy in PISA (Demir, 2010). Yet, open-ended items are considered more appropriate for measuring students' upper-level thinking skills (Ministry of National Education, 2017). For this reason, in parallel to the developments in the world, there has recently been a tendency to use open-ended items in selection and placement tests administered in Turkey in the transition both into secondary education and higher education. The Measuring, Selection and Placement Center (ÖSYM) made an announcement about the intention of using open-ended items in the Undergraduate Placement Exam (LYS) and published sample open-ended items regarding those would be used in the exam (ÖSYM, 2017). However, it is observed that multiple-choice tests are used more frequently than other types of tests in most institutions and establishments of education, in teachers' in-class activities, in student selection examinations, and especially in assessments involving a great number of students. The major reason for this is that using multiple-choice tests has certain advantages. Probably, one of the most important advantages is that scoring the correct answers does not change from rater to rater since answering the items requires only choosing one of the given options and there is no need to determine the degree of accuracy in scoring. In these tests, scoring consists of counting the correct answers and is purely objective (Özçelik, 2010). The greatest restriction of multiple choice tests is that they cannot measure high levels of performance, abilities, or skills (Konak, 2010; Güler, 2013). Classical testing methods such as multiple-choice, short answer, True-False, matching, and fill-in-the-blank used in assessing students' behaviours are incapable of determining upper-level mental processes such as problem-solving, reading comprehension, critical thinking, analytical thinking, empathising, researching, decision-making, understanding the importance of social history, and creativity (Kutlu, 2006). Open-ended items, on the other hand, enable students to create their own answers, give different personal answers, and answer the items from their own perspective.

When more than one rater is used in the assessment process, they can be the source of variance; they can cause errors and thus reduce the reliability of the assessment. Error components in individuals' scores are due to a variety of factors that introduce measurement error into the scores, such as intraindividual factors, characteristics of the measure itself, administration factors, scoring errors, and so on (Goodwin, 2001). The crucial concern about performance assessment and scoring of open-ended items is the objectivity of scoring as it is not easy to assess performance objectively, unlike traditional assessments (e.g., fixed response items) (Romagnano, 2001). And in the scoring of the open-ended items, the different factors interfere with scoring and reduce reliability (Atılgan et all, 2011). Rater reliability is the consistency between the scores given to a certain property. There are biases in scoring arising from raters. Scullen et al., (2000) describe raters' influence as "a broad category of effects which are not related to students' real performance but are related directly to raters who cause systematic errors in performance evaluation". It may be said that raters' influence arises from such psychological states as motivation, anxiety, achievement, and self-efficacy (Bernardin & Villanova, 2005), from personal traits (Wexley & Youtz, 1985), from their former beliefs, demographic properties such as gender and age, and raters' experience in scoring (Weigle, 1998). Raters' interaction with other sources of variability mingled in measurement is also important in reliability (Brennan, 2001). Therefore, errors arising from several sources of variability should also be taken into consideration in determining reliability. Error and error sources involved in measurement results should be well defined and methods should be found to estimate the amount of error (Turgut, 1992). The accuracy of the measurement results is very important as it affects the decisions to be made based on these results. Reliability can be defined as the degree to which measurement results are free from random errors (Baykul, 2000). One of the methods capable of analysing different sources of variability and the interactions between those sources altogether is Generalizability Theory (G Theory) (Shavelson & Webb, 1991). Due to the fact that G Theory considers more than one source of errors at the same time, it is thought that analysing the international studies in which Turkey also takes part from this perspective would be beneficial. In the

study conducted by Goodwin (2001), 3 approaches used in the inter-rater concordance and reliability study were compared: a) Percent of Agreement and Kappa b) Simple Correlation Methods c) G Theory techniques. In the study, 10 students were scored by 2 raters for the quality of their physical activities on 6 different days. Each rater evaluated the students over 7 points (1-lowest, 7-highest). As a result of the comparison of the advantages and disadvantages of different approaches, it is emphasized that the G Theory techniques are the most comprehensive, most flexible ones and allow the isolation of measurement errors caused by different sources in a study. It has been stated that the G Theory is the approach that gives the most information about the generalizability or reliability of the scores. Lee (2005), in his study, investigated the effect of the change in the number of tasks and raters in the TOEFL test on generalizability and tried to determine the most appropriate number of tasks for maximum reliability. As a result of the study, it was seen that increasing the number of tasks was more effective than increasing the number of raters. Therefore, it was concluded that using fewer raters in performance evaluation is appropriate for an acceptable level of generalizability. In a study on the generalizability of TIMMS open-ended items, Smith (1997) analyses the effects of raters. The researcher analyses the answers given by 150 students to each item with 50 booklets in each of seven English-speaking countries. It is stated that the number of raters should be raised to 15 from 5 for a generalizability level of 0.80 in all items and that this situation can cause problems in countries where the number of raters is small. Brennan (2001) states that the proficiency criteria of the reliability coefficients vary voluntarily, but some researchers may consider it "high" if the G and Phi coefficients are greater than 0.80. Sharma and Weathers (2003) investigated the generalizability of scales used in international research projects in all participating countries. It was concluded that the scale had the same meaning in all countries and that it was not specific to a certain country. It was also concluded that if the level of generalizability was 0.90, using 11 items out of 17 was sufficient for that level. Using a minimum number of items decreases examination/questionnaire time, eliminates the effects of tiredness and lowers costs. Thus, it is thought that knowing the adequate number of raters will be used in PISA would help to reduce labour and costs.

The number of raters to fulfill the task of scoring open-ended items as well as scoring time increases depending on the number of students participating and the number of items to be scored. For instance, 16 Turkish / Turkish Language and Literature teachers were needed only in scoring the reading literacy items of PISA 2009 (OECD, 2012). Appointing teachers to the task of scoring causes problems since schools are still open during this period and teachers have teaching tasks in their schools as well.

Countries can form a combination of differing designs in PISA and a design can be applied if it is accepted by the PISA consortium (OECD, 2012). Therefore, it is expected that determining which of the designs would be more appropriate for use and determining the minimum number of raters to attain the desired level of generalizability in each booklet will contribute to such examinations in terms of time, costs, and labour.

This study is believed to set a model in determining inter-rater reliability for open-ended items in such international studies as TIMSS, PIRLS, and ICILS and in terms of performing decision studies and to shed light on studies to be conducted in the future.

Assessment, Selection and Placement Centre (ÖSYM) started to make use of open-ended items in the selection and placement examinations (ÖSYM, 2013, 2017). Under the title of *"Information about Open-Ended Items and Examples"* details were given. Although it was expressed as an open-ended item, it was seen that the question type mentioned was short-answer questions and it was stated that the answer will consist of a word, a number, or a sentence (ÖSYM, 2017). However, there are no reliability and decision studies for such items; besides, such issues as how many items would be adequate in those examinations in which there is a great number of participants and how many raters should score and in what design they should perform the task are of great importance. For this reason, it is believed that this study will function as a guide. This study compares the results of the G study obtained from the designs created through more than one rater's scoring students' reading skills in PISA 2009 altogether or alternately according to the G Theory with the results of the decision study conducted with those designs.

Method

This is a case study since it determines the properties of scoring PISA 2009 reading skills, and it is descriptive. Descriptive studies are the studies describing an existing event or the properties of an individual or a group as it is/as they are and describing a current state quantitatively or qualitatively (Karasar, 1998).

Population and Sample

The sample of Turkey in PISA 2009 was composed of 4996 students from 170 schools who were randomly chosen by PISA international center by stratifying them according to 12 statistical region classification (IBBS, NUTS) and school types. 362 students who answered the reading skills items in PISA 2009 and whose booklets were exposed to multiple scoring constituted the sub-sample.

For the main survey, it was recommended to have 16 coders to code reading, 8 coders to code mathematics, and an additional 8 coders to code science items. Other possible coding designs were 16 reading and 8 mathematics and science coders or 16 reading, 4 mathematics, and 4 science coders. These numbers of coders were considered to be adequate for countries testing between 4 500 (the minimum number required) and 6 000 students to meet the timeline of submitting their data within 3 months of testing (OECD, 2012).

National Project Managers (NPMs) were responsible for recruiting appropriately qualified people to carry out the single and multiple coding of the test booklets. It was not necessary for coders to have high-level academic qualifications, but they needed to have a good understanding of the language of the test and to be familiar with ways in which secondary-level students express themselves.

In Turkey, the population of raters was composed of Turkish or literature teachers who had experience in international projects as a rater before or who had been teaching 15-year-old students (from 7^{th} graders to 10^{th}). An important factor in recruiting coders was that they could commit their time to the project for the duration of the coding, which was expected to take up to one month. An official letter has been sent to schools to identify potential teachers with these qualifications. As a result of this process, 16 teachers were selected to take part in the scoring.

From 13 booklets used in PISA 2009, booklets 2 and 8 which contained reading skills and scored by four raters were used in this study. Booklet 2 contained six items whereas booklet 8 contained eight items.

Research Data

The data collected from multiple raters scoring of reading skills in PISA 2009 constituted the data of this study. The data were provided by the Educational Research and Development Directorate (EARGED).

Two designs were created so as to be used in G Theory in the study. One of them was the crossed design symbolized as "s x i x r" (student x item x rater), in which students were scored by each rater in terms of the same skills. The second was the nested design symbolized as "(r:s) x i", where each rater scored only a group of students and raters nested in students, and the items were crossed with these variables.

Data Analysis

This study aims to compare the results of the generalizability (G) and decision (D) studies of the scores of reading literacy items in PISA 2009 according to the crossed (s x i x r) and nested ((r:s) x i) designs and to compare the G and Phi coefficients as estimated by increasing or decreasing the number of raters in these designs. The data were analysed on the basis of these designs.

EDUG 6 programme was used in estimating variance components of the designs through G Theory, in calculating the rates of explaining the total variance of variables, and in performing the decision study for each design. EDUG 6 programme was developed for G Theory analyses, and it enables researchers to perform G and D studies for sources of variability they describe and for the designs they form with those sources of variability.

Findings and Interpretations

In this section, the variance components and percentages explaining the total variance for crossed (s x i x r) and nested ((r:s) x i) designs and Generalizability levels and the results for D Study performed by changing the number of raters in these designs in Booklet 2 and Booklet 8 will be given.

Table 1

Variance Components for "s x i x r" and "(r:s) x i" Designs and Percentages Explaining the Total Variance in Booklet 2 and Booklet 8

	Crossed Design						Nested Design					
	Sources of variance	Squares total	Degrees of freedom	Squares average	Variance	%	Sources of variance	Squares total	Degrees of freedom	Squares average	Variance	%
Booklet 2	Students	267.59	99	2.70	0.10577	16.0	Students	287.68	99	2.90	0.11076	18.5
	Items	137.14	5	27.42	-0.00095	0.0	Items	78.79	5	15.75	0.03899	6.5
	Raters	39.53	3	13.17	-0.02470	0.0	r:s	156.45	300	0.52	0.01439	2.4
	si	60.85	495	0.122	-0.03699	0.0	si	79.91	495	0.16	-0.06844	0.0
	sr	92.80	297	0.312	0.00692	1.0						
	ir	419.35	15	27.95	0.27686	41.9						
	sir, e	402.31	1485	0.27	0.27092	41.0	ir:s, e	652.79	1500	0.43	0.43519	72.6
	Total	1419.59	2399			100%	Total	1255.64	2399			100%
Booklet 8	Student	398.14	99	4.02	0.09072	19.8	Students	434.02	99	4.38	0.07488	13.7
	Items	14.25	7	2.03	0.00039	0.1	Items	14.99	7	2.14	0.00484	0.9
	Raters	60.97	3	20.32	0.02181	4.8	r:s	607.90	300	2.02	0.22295	40.9
	si	114.08	693	0.16	-0.01120	0.0	si	141.53	693	0.20	-0.00964	0.0
	sr	345.49	297	1.16	0.11923	26.0						
	ir	40.41	21	1.92	0.01715	3.7						
	sir, e	435.36	2079	0.20	0.20941	45.7	ir:s, e	509.84	2100	0.24	0.24278	44.5
	Total	1408.73	3199			100%	Total	1708.30	3199			100%

Variances estimated through G study and percentages explaining the total variance in Booklet 2 and Booklet 8 are given in Table 1. The variance component of the variable of students in Booklet 2 explains 16% of the total variance for crossed design while it explains 18.5% for nested design. We can see almost the same pattern in Booklet 8 and the variance component of the variable of students explains 19.8% of the total variance in crossed design whereas it explains 13.7% in nested design in this booklet. The variance component of students indicates that students differ in terms of reading skills. This pattern is similar in both designs and in both Booklets. In generalizability studies, variance due to students is considered as a universe score and this variance shows the difference between students in terms of characteristic which was measured (Brennan, 2001; Shavelson & Webb, 1991).

Accordingly, the percentage of the variance components of items explain the total variance is 0% for crossed design while it is 6.5% for nested design in Booklet 2 and the percentage of the variance components of items explain 0.1% of total variance in crossed design while it explains 0.9% in nested design in Booklet 8. It is clear in this case that items do not differ in terms of difficulty in crossed design but that they differ in nested design in Booklet 2. The fact that variance components are bigger in nested design is indicative of the fact that tasks are discriminated better and items do not differ in terms of difficulty in both designs in Booklet 8.

In Booklet 2, it may be stated that the variance component of raters' influence is quite small in crossed design and that therefore students are scored consistently in this booklet. In Booklet 8, it may be said that the value is high and that students' scores differed from one rater to another in crossed design. Since raters are nested within students, it is impossible to separate the raters' main effect from the interaction between students and raters. We interpret the substantial variance component for those combined effects (r:s=0.01439; 2.4 % of the total variance) in Booklet 2, and (r:s=0.22295; 40.9% of the total variance) in Booklet 8 as indicating student behavior differed from one rater to another. We do not know whether one rater produced more behavior than another (rater main effect), whether the relative standing of the student differed from one rater to another (student-by-rater interaction), or both (Shavelson & Webb, 1991. On examining the joint item rater variance component in Booklet 2, it is clear that the variance component is small in value and that raters do not differ in scoring from one item to another.

The residual variance was found to be high in both designs in Booklet 2 (sir,e=0.27092; 41.0% of the total variance; ir:s,e=0.43519; 72.6% of the total variance) but it was higher in the nested design. And also in Booklet 8, the residual variance was found to be high in both designs (sir,e=0.20941; 45.7% of the total variance; ir:s,e=0.24278; 44.5% of the total variance). For the variance component obtained from the interaction of three sources of variability to be zero (0) is a desired situation. The large residual component indicates that a substantial amount of variation is due to these confounded sources of variation (Shavelson & Webb, 1991).

Table 2

Generalizability Levels for s x i x r and (r:s) x i Designs

	Crossed Design			Nested Design	
Booklet 2	G coefficient	0.89	Booklet 2	G coefficient	0.99
	Phi coefficient	0.81		Phi coefficient	0.98
Booklet 8	G coefficient	0.71	Booklet 8	G coefficient	0.95
	Phi coefficient	0.68		Phi coefficient	0.85

On comparing the "s x i x r" and "(r:s) x i" designs, it was found that the G and Phi coefficients obtained from the (r:s) x i design were higher than those obtained for the s x i x r design. Having a generalizability coefficient of >.80 is a desirable situation (Mushquash & O' Connor, 2006). It was found in crossed design results, especially for booklet 8, that the G coefficient was not at an acceptable level and that the Phi coefficient was also below that level. However, the G and Phi coefficients for the same booklet were 0.95 and 0.85 respectively in the nested design and thus they were above the acceptable level.

Table 3

Results for D Study Performed by Changing the Number of Raters in the "s x i x r" and "(r:s)xi" Designs

	Design	Crossed Design					Nested Design				
	Rater	2	3	4	5	6	2	3	4	5	6
Booklet 2	G	0.80	0.86	0.89	0.91	0.92	0.95	0.96	0.97	0.98	0.98
	Phi	0.68	0.76	0.81	0.84	0.87	0.92	0.94	0.95	0.95	0.95
Booklet 8	G	0.56	0.65	0.71	0.76	0.79	0.80	0.86	0.89	0.91	0.92
	Phi	0.52	0.62	0.68	0.73	0.76	0.61	0.68	0.72	0.74	0.76

According to Table 4, the G coefficient obtained through scoring 100 students by 4 raters in terms of 6 items in booklet 2 in the crossed design is 0.89 and the Phi coefficient is 0.81. In nested design, on the other hand, the G coefficient in scoring 100 students by 4 raters is 0.97 and the Phi coefficient is 0.95- which are high values. On reducing the number of raters to 2 for scoring 100 students the G coefficient is found to be 0.80 and the Phi coefficient is found to be 0.68 in crossed design. However, the G and Phi coefficient were 0.95 and 0.92 respectively in the nested design.

Accordingly, the G coefficient found by scoring 100 students by 4 raters in terms of the 8 items in booklet 8 is 0.71 and the Phi coefficient is 0.68 in crossed design. This level of generalizability is well below 0.80- which is the acceptable level. The G coefficient with the same number of raters scoring the same number of students is 0.89 and the Phi coefficient is 0.72 in nested design.

On reducing the number of raters to 2 in scoring 100 students in crossed design, the G and Phi coefficients were found as 0.56 and 0.52, respectively. Yet, the G coefficient was 0.80 and the Phi coefficient was 0.61 in the nested design.

Discussion and Conclusions

Discussion

In the literature, G Thoery techniques are considered the most comprehensive and flexible in the estimation of interrater agreement and reliability (Goodwin, 2001) and G Theory is one of the methods capable of analysing different sources of variability and interactions between those sources altogether. In this study, the aim was to compare different designs (crossed and nested) according to G Theory and to find out the most effective way of scoring PISA open-ended items. Almost 5000 students participated in PISA 2009 assessment and the scoring process took almost a month to complete with the participation of 16 teachers as the raters. Reducing the cost and labour was the starting point of this study.

Conclusion

On examining the results for Generalizability study in the s x i x r design for booklets 2 and 8 in PISA 2009, it was found that students differed in terms of reading skills in booklet 2, that items did not differ in terms of difficulty, that students' performance did not differ from item to item, that raters made a consistent assessment, that the student-rater interaction was very low and that raters' assessment did not differ from student to student. Yet, the joint effect of items and rater was very high. In this case, raters' scoring could be said to change from item to item.

It was found that raters differed in booklet 8 and that the joint effect of students and raters explained 26% of the total variance. Thus, raters' scoring changed from student to student. This situation manifested itself in the generalizability coefficient and the level of generalizability remained at 0.71.

However, the variance components for the joint effect of student item and rater explained the total variance at a high percentage. For the variance component obtained from the interaction of three sources of variability to be zero (0) is a desired situation. Having this ratio high can indicate that the student, item, rater interaction, and/or sources of random error can be big. On comparing the s x i x r and (r:s) x i designs, it was found that the relative and the absolute error variances estimated in the (r:s) x i design was smaller than in the s x i x r design, and thus the G and the Phi coefficients took on bigger values.

On examining the decision studies performed in both designs, it was found that increasing the number of raters provided an increase in the G and Phi coefficients in both designs but that the increase in the G coefficient obtained in this way was not big enough to bring advantages in terms of being economical. It was found that it was possible to reach acceptable levels of G coefficient by reducing the number of raters by half in Booklet 2.

The G and Phi coefficients were calculated as 0.89 and 0.81 respectively in booklet 2. When the number of raters is 5 and the number of students is kept constant, the G coefficient was calculated as 0.91 and the Phi coefficient as 0.84. When the number of raters is 6, the G coefficient was 0.92 and the Phi coefficient was 0.86. On increasing the number of raters, there was an increase in the G and Phi coefficients. But it was not substantial. When the number of raters was reduced to 3, the G and Phi coefficients were found to be 0.85 and 0.76 respectively whereas the G coefficient fell down to 0.80 and the Phi coefficient to 0.68 on reducing the number of raters to 2. In this situation, the Phi coefficient fell below the acceptable level while the generalizability coefficient remained at an acceptable level.

In booklet 8, the G and Phi coefficients were calculated as 0.71 and 0.68, respectively. This was below 0.80- which is the acceptable level for the G and the Phi coefficients (Mushquash & O' Connor, 2006). The Generalizability coefficient falls down to 0.56 and the Phi coefficient to 0.52 on reducing the number of raters to 2 from 4 and keeping the number of students constant. When the number of raters is 5, the G coefficient is 0.76 and the Phi coefficient is 0.73. When the number of raters is 6, the G coefficient is 0.79 and the Phi coefficient is 0.76. An increase occurred in the G and the Phi coefficients when we increased the number of raters. Yet, at least 6 raters are required to get a G coefficient at an acceptable level. It was found that the results obtained in earlier studies concerning G Theory were supportive of the ones obtained in this study. Different designs of G Theory were compared and the most appropriate number of items and the most appropriate number of raters were considered. Increasing the number of raters led to an increase in the G coefficient, but the effects of the increase diminished after a certain number of raters. Atılgan (2008), in a study concerning the generalizability of tests for selecting students in Music Department at Inönü University, found that it would be more appropriate to continue with the initial number of raters due to the fact that the increase in the G coefficient was not very effective. Smith (1997), in a generalizability study concerning the effect of the number of raters in the scoring of the open-ended items in TIMSS, found that the effect of the increase in the number of raters differs from one item to another and it would be more appropriate to raise the number of raters to 15 from 5 for the desired level of generalizability in all items and this shows that there may be a problem in countries where the number of raters is low. In some studies, it was found that increasing the number of tasks rather than raters was more efficient to maximize the score reliability. In a study concerning

generalizability of a performance assessment measuring achievement in eighth-grade Mathematics, Mcbee and Barnes (1998) investigated the effect of task similarity to generalize the results of the assessments. The Generalizability study results showed that the number of tasks required to reach acceptable levels of generalizability would be prohibitively high, even using only highly similar tasks. Schoonen (2005) found out that the generalizability of writing scores and the effects of raters and topics are very much dependent on the way the essays are scored and the trait that is scored. The overall picture is that writing tasks contribute more to the score variance than raters do. Lee (2005), in a generalizability study concerning the effect of number of raters and the number of tasks in the scoring of TOEFL writing assessment, reported that increasing the number of tasks rather than the number of raters per task would be more efficient to maximize the score reliability for writing. In a study of generalizability of students writing across multiple tasks, Hathcoat and Penn (2012) found that 77% of error variance may be attributable to differences within people across multiple writing assignments. D studies indicated that substantive improvements in reliability may be gained by increasing the number of assignments, as opposed to increasing the number of raters. Therefore, it was concluded that using fewer raters in performance evaluation is appropriate for an acceptable level of generalizability. In a study called the comparison of different designs in accordance with the generalizability theory in communication skills, Nalbantoğlu and Gelbal (2011) did G and D studies and compared crossed (s x t x r) and nested ((s:r) x t) designs and observed that the variance that were estimated for variables in both designs were similar to each other and also D studies yielded the similar results and it was found that the scoring of certain number of students alternately (nested design) is much more convenient in time, labor, and cost. Polat and Turhan (2021) compared crossed and nested designs in language testing. G and Phi values and the variance associated with the student's main effect were higher, while the variance value of the residual effect was lower in crossed design. This study revealed that crossed designs could generate more reliable results in speaking exams. Zorba (2020) compared the result of a written exam used in personnel recruitment with different patterns in the generalizability theory. In this study, G and Phi coefficients were calculated as 0,33 and 0,29 for (p x i x r) (person x item x rater) design, and 0,76 and 0,64 for (p x (i : r)) (person x (item : rater) design, respectively. According to the results of the D study, it was observed that increasing the number of raters in crossed (p x i x r) design and increasing the number of items in nested (p x (i: r)) design increased the reliability. Khodi (2021), in a study of G-theory analysis of rater, task, and scoring method contribution, found that at least four raters (with G-coefficient = 0.80) were necessary for a valid and reliable assessment and he suggested student performance should be rated on at least two scoring methods by at least four raters.

Therefore, it is believed that determining the minimum number of raters by considering the degree to which an increase in the G coefficient is influential in results will help to reduce labour and costs in making decisions about determining the number of raters.

Recommendations

Using the designs in which a group of raters scores a group of students alternately instead of having all raters score all students will be more economical in terms of time and labour if there is consistency between raters in performance determining examinations where there are a great number of students and more than one rater score them.

It was observed that a certain amount of decrease occurred in inter-rater consistency and in generalizability coefficients in partial scoring in the form of partial credit as 2-1-0 in booklets. Therefore, it should be made sure that a greater number of examples is given in training raters for booklets which are scored partially, the number of local examples should be increased, and scoring should be done on the item level not on the unit level. This means that all the items in a unit should be coded one by one and a new unit is started only after completing all the items in the previous unit in all booklets.

It is important that the group to function as raters in international activities was described beforehand. Assigning teachers for scoring from schools during school time and especially at the end of a semester for such activities lowers the teachers' motivation. Teachers should be able to perform the task of scoring with no fear of wasting time and disrupting their school work at the end of a semester.

Using open-ended items in a test for the transition from elementary education into secondary education and in a university entrance exam has been on the agenda. Using open-ended items in those examinations-which are extremely important in shaping students' future- is an issue that should be carefully worked on. It should not be forgotten that rater reliability is very important in assessing open-ended items. Considering the situations where the scoring of a booklet by 4 raters in PISA is inadequate, the number of staff to make assessment meeting the standards and the length of time required for the assessment of 2 million students and the reflections into the system should be calculated carefully.

Studies comparing different sources of variability (such as booklets, modules, etc.) and different designs in international performance assessment examinations such as TIMSS, PIRLS, PISA, and ICILS, in which Turkey takes part, could be performed.

Studies considering all the sub-fields such as mathematics literacy and science literacy in international performance assessment examinations such as TIMSS and PISA and analysing the correlations between them from the perspective of different scoring designs could be performed.

Declarations

Author Contribution: Meral Alkan: Conceptualization, methodology, analysis, writing & editing, visualization. Nuri Doğan: Methodology, editing, and supervision.

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

Ethical Approval: This research study complies with research publishing ethics. Secondary data were used in this study. Therefore, ethical approval is not required.

References

Atılgan, H. (2008). Using generalizability theory to assess the score realibility of the special ability selection examinations for music education programmes in higher education. *International Journal of Research and Method Education*, *31*(1), 63-76. https://doi.org/10.1080/17437270801919925.

Atılgan, H., Kan, A. & Doğan, N. (2011). *Eğitimde ölçme ve değerlendirme.* (5. Baskı). Anı Yayıncılık.

Balbağ, M., Leblebicier, K., Karaer G., Sarıkahya E. & Erkan Ö. (2016). Türkiye'de fen eğitimi ve öğretimi sorunları. *Eğitim ve Öğretim Araştırmaları Dergisi, 5*(3), 1-12. http://www.jret.org/FileUpload/ks281142/File/02.m._zafer_balbag.pdf

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması.* ÖSYM

Bernardin, H. J. & Villanova, P. (2005). Research streams in rater self-efficacy. *Group and Organizational Management*, 30, 61-88. https://doi.org/10.1177/1059601104267675

Biemer, L. (1993). Trends-social studies /authentic assessment. *Educational Leadership*, *50* (8). https://www.ascd.org/el/articles/-authentic-assessment

Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag Publishing. https://doi.org/10.1007/978-1-4757-3456-0

Demir, E. (2010). *Uluslararası öğrenci değerlendirme programı (PISA) bilişsel alan testlerinde yer alan soru tiplerine göre Türkiye'de öğrenci başarıları* (Yayınlanmamış yüksek lisans tezi). Hacettepe Üniversitesi.

EARGED (2010). *PISA 2009 projesi, ulusal ön raporu*. 15 Mart 2011 tarihinde http://earged.meb.gov.tr/pdf/pisa2009rapor.pdf adresinden erişilmiştir.

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercises Science, 5*(1), 13-34. https://doi.org/10.1207/S15327841MPEE0501_2

Güler, N. (2013). *Eğitimde ölçme ve değerlendirme* (5. Baskı). Pegem Akademi.

Hathcoat, J. D., & Penn, J. D. (2012). Generalizability of student writing across multiple tasks: A challenge for authentic assessment. *Research & Practice in Assessment*, 7, 16-28. https://files.eric.ed.gov/fulltext/EJ1062689.pdf

Karasar, N. (1998). *Araştırmalarda rapor hazırlama yöntemi*. Pars Matbaacılık

Khodi, A. (2021). The affectability of writing assessment scores: A G-theory analysis of rater, task and scoring method contribution. *Language Testing in Asia 11*, Article 30 https://doi.org/10.1186/s40468-021-00134-5

Konak, Ö. A. (2010). Eğitim ve öğretim etkinlikleri üzerine. *Cito Eğitim: Kuram ve Uygulama Dergisi,* 10, 4-5.

Kutlu, Ö. (2006). Üst düzey zihinsel süreçleri belirleme yolları: Yeni durum belirleme yaklaşımları. *Çağdaş Eğitim Dergisi, 31(*335), 15-21. https://search.trdizin.gov.tr/tr/yayin/detay/74516/

Lee, Y. W. (2005). *Dependability of scores for a new ESL speaking test: Evaluating prototype tasks*. ETS. http://www.ets.org/Media/Research/pdf/RM-04-07.pdf

Mcbee, M., & Barnes, L. (1998), The generalizability of a performance assessment measuring achievement in eighth-grade mathematics. *Applied Measurement in Education, 11*(2), 179-194. https://doi.org/10.1207/s15324818ame1102_4

MEB (2017). *Akademik becerilerin izlenmesi ve değerlendirilmesi (ABİDE) projesi.* 1 Eylül 2022 tarihinde http://abide.meb.gov.tr/proje-hakkinda.asp adresinden erişilmiştir.

Mushquash, C., & O'Connor, B.P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods 38*, 542–547 https://doi.org/10.3758/BF03192810

Nalbantoğlu, F. & Gelbal, S. (2011). İletişim becerileri istasyonu örneğinde genellenebilirlik kuramıyla farklı desenlerin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 41*, 509-518. http://www.efdergi.hacettepe.edu.tr/shw_artcl-718.html

OECD (2012). *PISA 2009 technical report*, PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264167872-en

OECD (2017), OECD (2017), *PISA 2015 assessment and analytical framework: science, reading, mathematic, financial literacy and collaborative problem solving*, PISA, OECD Publishing http://dx.doi.org/10.1787/9789264281820-en

ÖSYM (2013). *Açık uçlu sorularla deneme sınavı: Soru/cevap kitapçığının yayımlanması* www.osym.gov.tr/belge/1-19413/acik-uclu-sorularla-deneme-sinavi-sorucevap-kitapcigini-.html adresinden erişim sağlanmıştır.

ÖSYM. (2017). *Açık uçlu sorular hakkında bilgilendirme ve açık uçlu soru örnekleri.* https://www.osym.gov.tr/TR,12909/2017-lisans-yerlestirme-sinavlari-2017-lys-acik-uclu-sorular-hakkinda-bilgilendirme-ve-acik-uclu-soru-ornekleri-05012017.html adresinden erişim sağlanmıştır.

Özçelik, D. A. (2010). *Ölçme ve değerlendirme*. Pegem Akademi.

Polat, M. & Turhan, N. (2021) Applying generalizability theory in language testing: Comparing nested and crossed scoring designs in the assessment of speaking skills, *International Journal of Curriculum and Instruction*,*13*(3), 3344–3358. https://ijci.globets.org/index.php/IJCI/article/view/825/409

Romagnano, L. (2001). The myth of objectivity in mathematics assessment. *Mathematics Teacher*, *94* (1), 31-37. https://doi.org/10.5951/MT.94.1.0031

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing 22*(1) 1-30. https://doi.org/10.1191/0265532205lt295oa

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*(6), 956–970 https://doi.org/10.1037/0021-9010.85.6.956

Sharma, F. & Weathers, D. (2003). Assessing generalizability of scales used in cross-national research. *International Journal of Research in Marketing*, 20, 287-295. http://dx.doi.org/10.1016/S0167-8116(03)00038-7

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage Publications

Smith, Teresa A. (1997 March 24-28). *The Generalizability of Scoring TIMSS Open-Ended Items. (Report).* Annual Meeting of the American Educational Research Association, Chicago, USA

Turgut, F. M. (1992) *Eğitimde ölçme ve değerlendirme metotları. (*9. Baskı). Saydam Matbaacılık.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287. https://doi.org/10.1177/026553229801500205

Wexley, K. N. & Youtz, M. A. (1985). Rater beliefs about others: Their effect on rating errors and rater accuracy. *Journal of Occupational Psychology*, 58, 265-275. https://psycnet.apa.org/doi/10.1111/j.2044-8325.1985.tb00200.x

Zorba, İ. (2020). *Personel alımında kullanılan bir yazılı sınav sonucunun genellenebilirlik kuramındaki farklı desenlerle karşılaştırılması* (Yayımlanmamış yüksek lisans tezi). Ankara Üniversitesi.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(2); 118-127

The Relation of Item Difficulty Between Classical Test Theory and Item Response Theory: Computerized Adaptive Test Perspective

Eren Can AYBEK*

Abstract

This study aims to transform the calculated item difficulty statistics according to Classical Test Theory (CTT) into the item difficulty parameter of Item Response Theory (IRT) by utilizing the normal distribution curve and to analyze the effectiveness of this transformation based on Rasch model. In this regard, 36 different data sets created with catR package were studied. For each data set, item difficulty parameters and transformed item difficulty parameters were calculated and the correlation coefficients between these parameters were analyzed. Then, Computerized Adaptive Test (CAT) simulations were performed using these parameters. According to the simulation results, the correlation coefficients between the estimated theta values with both methods were high. Furthermore, in CAT simulations in which both parameters were used, especially in the samples which were over 250, it was found to have similar bias, RMSE values, and the average number of administered items.

Keywords: Item difficulty, classical test theory, item response theory, Rasch model

Introduction

A measurement tool can be developed based on Classical Test Theory (CTT) or Item Response Theory (IRT) (de Ayala, 2009). Tests are easy to develop under the CTT, yet it has some limitations. For example, a single standard error value for the entire test score can be calculated by using CTT; the item statistics depend on the examinees, and the true score estimates are based on the item set (Hambleton & Swaminathan, 1985). The studies show that an item that should be removed from the test according to CTT should also be taken out of the test according to IRT, which reveals the fact that CTT and IRT estimates are similar when deciding whether an item is good or bad (Çelen & Aybek, 2013). On the other hand, IRT comes to the fore for studies such as Computerized Adaptive Test (CAT), test equation and linking, and Differential Item Functioning (DIF), but loses its practicality for classroom assessment.

IRT models can be classified in different ways according to the dimension that is measured and the number of response categories. In addition to unidimensional IRT models in which an item measures one single dimension, there are also multidimensional IRT models in which an item can measure multiple dimensions (Reckase, 2009). In addition, there are some models such as Rasch, 1 Parameter Logistic (1PL), 2PL, 3PL, and 4PL models for dichotomous items (Hambleton et al., 1991); Nominal Response Model (NRM) (Bock, 1972); Partial Credit Model (PCM) (Masters, 1982); Generalized Partial Credit Model (GPCM) (Muraki, 1992); and Graded Response Model (GRM) for polytomous items (Samejima, 1996).

In the Rasch model, the probability of responding to an item correctly depends only on the item difficulty, *b*, parameter of that item, while the item discrimination, *a*, parameter is considered to be 1.00 for all the items. The Rasch and 1PL models are similar in that item discrimination is considered the

* Assoc. Prof., Pamukkale University, Faculty of Education, Denizli-Türkiye, erencan@aybek.net, ORCID ID: 0000-0003-3040-2337

To cite this article:
Aybek, E. C. (2023). The relation of item difficulty between Classical Test Theory and Item Response Theory: Computerized adaptive test perspective. *Journal of Measurement and Evaluation in Education and Psychology, 14*(2), 118-127. https://doi.org/10.21031/epod.1209284

Received: 23.11.2022
Accepted: 28.06.2023

same for all items; however, *a* parameter can take different values than 1.00 in 1PL model (de Ayala, 2009).

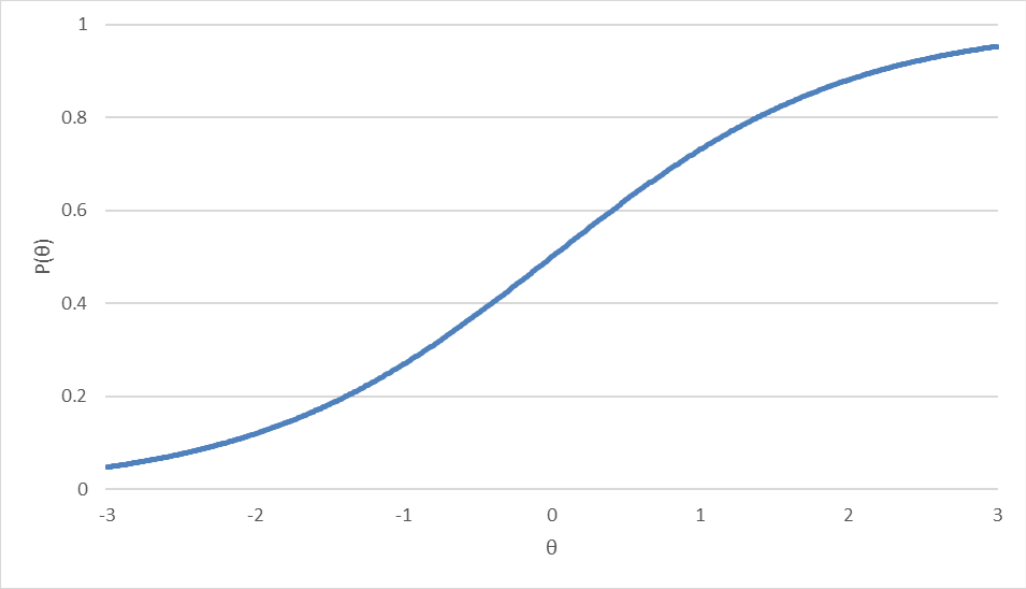
According to the Rasch model, the probability for an individual with a given (θ) ability level to respond correctly to an item whose difficulty parameter is *b* is calculated with the equation below (1) (Rasch, 1961):

$$p(\theta) = \frac{e^{(\theta-b)}}{1+e^{(\theta-b)}} \qquad (1)$$

The *b* parameter represents item difficulty and refers to the θ level at which the item is correctly answered with 50% probability. Although the theoretical limits for θ are between (-∞, ∞), they usually work within ranges such as [-3, 3] or [-4, 4]. When Equation 1 is applied, the probability of responding to an item correctly with $b = 0.00$ for all θ levels within the range [-3, 3] with an increment of 0.01 creates a curve as shown in Figure 1, and this curve is called the item characteristic curve.

When Equation 1 and Figure 1 are analyzed, another superiority of IRT can be recognized. Item parameters and the examinee's ability level are described on the same scale. As stated earlier, the *b* parameter for the item shown in Figure 1 is 0.00, which means that an examinee whose θ level is 0.00 responds to this item correctly with a probability of %50. In addition, when Figure 1 is carefully analyzed, it can be recognized that, as the θ level decreases, the probability of responding to an item correctly also decreases, and as it increases, the probability of responding to the item correctly increases, as well. In this context, the *b* parameter has similar limits as θ.

Figure 1.
A sample item characteristic curve for b=0 parameter in Rasch model



On the other hand, normal distribution is described as "*a theoretical distribution for a continuous variable measured for an infinite population"* (Crocker & Algina, 1986, p.24) and defined using the Equation 2 (Pitman, 1993):

$$Y = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}z^2} \qquad (2)$$

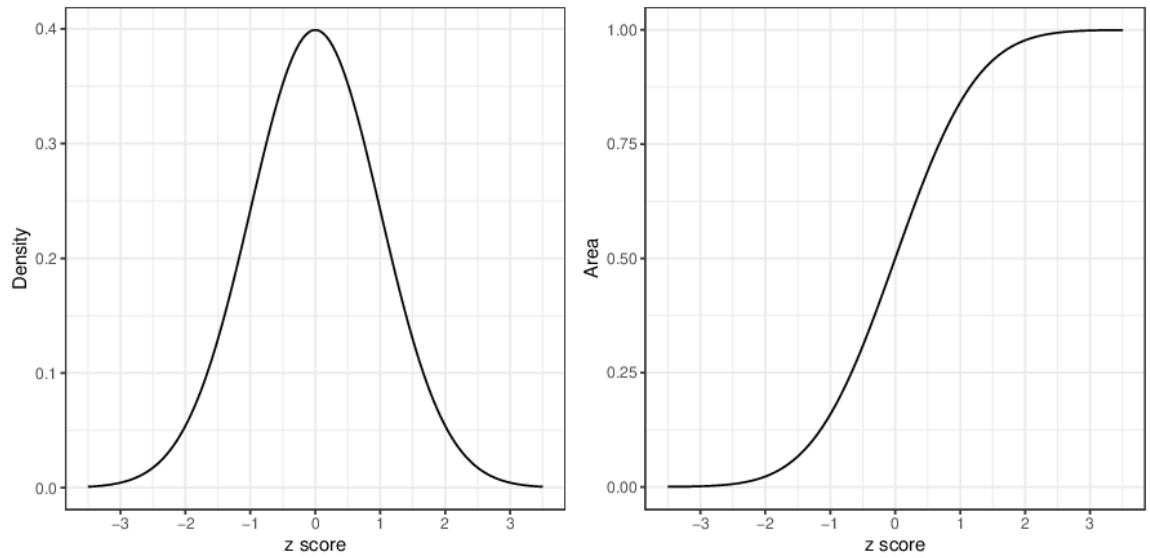
In this Equation, z stands for the standard z score and is obtained as $z = (x - \mu)/\sigma$ In addition, the area under the normal distribution curve can be obtained approximately by Equation 3 (Pitman, 1993):

$$\phi(z) \approx 1 - \frac{1}{2}(1 + c_1 z + c_2 z^2 + c_3 z^3 + c_4 z^4)^{-4} \quad (z \geq 0) \qquad (3)$$

c values in this equation as follows: c_1 = 0.196854, c_2 = 0.115194, c_3 = 0.000344, and c_4 = 0.019527. When z is below 0, $\phi(-z) = 1 - \phi(z)$ relation can be used by utilizing the symmetric characteristic of normal distribution curve. Accordingly, when the Equation 3 is used, the area under normal distribution curve for z = 1 constitutes approximately %84,3 of the whole area. Based on all this information, a normal distribution curve and the area under the normal distribution curve are given in Figure 2.

Figure 2.

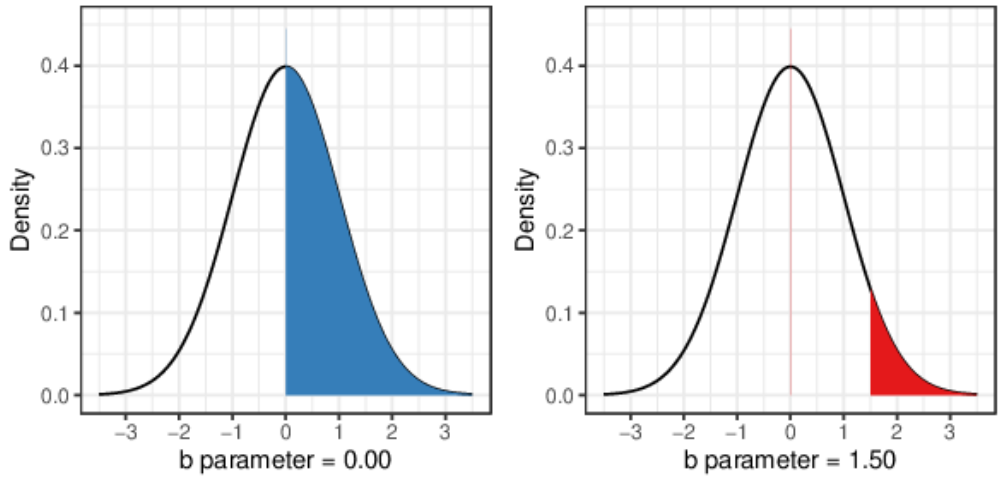
Normal distribution curve and the area under the normal distribution curve



The right side of Figure 2 shows the area under the normal distribution curve for different z scores. When this plot is analysed, its similarity with the item characteristic curve in Figure 1 is significant. Therefore, is it possible to interpret that an item with b = 0.00 is responded correctly by half of the group (p = 0.50 according to CTT) and an item with b = 1.50 is responded correctly by %6.5 (p = 0.065 according to CTT) of the group? Both cases are shown in Figure 3.

Figure 3.

Two sample b parameters on the normal distribution curve



As shown in the graphic in Figure 2, the view that the normal distribution curve can be used for an IRT-based conversion is not unrealistic. In fact, Lord (1980) focuses on the relation between CTT and IRT and he indicates that "*We can see from the figure that the item response function … is equal to a standardized normal curve area*" (p.31), and also adds that this approximation is "*not for practical use but rather to give an idea of the nature of the item discrimination parameter"* (p.33-34). Lord (1980) also describes the relationship between CTT item difficulty and discrimination statistics and IRT *a* and *b* parameters with mathematical proving. When all the items have the same discrimination (e.g., Rasch model), $b_j \approx \phi_j$ while b_j represents the IRT item difficulty parameter for *item j* and ϕ_j represents the area under the normal distribution curve at the point of CTT item difficulty statistics, p_j. On the other hand, if the items have different item discrimination, then $b_j \approx \phi_j \,/\, r_{jx}$ while r_{jx} represents the item-total biserial correlation or CTT item discrimination statistics. Lord (1980) also describes the relationship between IRT *a* parameter and CTT item discrimination statistics r_j as in Equation 4:

$$a_j = \frac{r_{jx}}{\sqrt{1 - r_{jx}^2}} \qquad (4)$$

In addition to that, even the equation of the area under the normal distribution curve looks very complicated, a simple function (e.g., NORM.DIST) in a spreadsheet software (Microsoft Excel, LibreOffice Calc, Google Sheet, etc.) can make the calculations.

Recent studies that support this perspective can be found in the literature. Kohli et.al. (2014) discussed the comparability of CTT and IRT based item parameters with underlying normal variable assumption. They found extremely high correlations between IRT and CTT based parameters, and these correlations are affected more by sample size rather than item pool size. Raykov and Marcoulides (2016) also shows the relationship and equivalence of CTT & IRT. They also recommend researchers combine the benefits of both test theories. A recent study by van der Ark and Smits (2023) suggests a new CAT method without using IRT, and they call it FlexCAT. Yet, their method is based on Latent Class Analysis (LCA), and it is still not very feasible for non-technical researchers.

Beyond the relationship between CTT and IRT in the manner of item parameters, how practical is the CTT to IRT transformation using this relationship for CAT applications? Due to the nature of CAT, we can estimate the trait level of the examinee with items that give more information about the examinee. In a typical CAT application, the next item to administer is selected based on the previous responses of the examinee, and the trait level can be estimated with much fewer items in contrast to conventional linear tests. This feature of the CAT makes it more convenient to schedule the test-taking time and place since not every examinee takes the same item set (van der Linden & Glas, 2002). Since CAT applications need a calibrated item bank, and the calibration process needs a large sample size, developing an item bank is not very feasible for a small-scale application. IRT calibration also needs expertise and cannot easily be implemented in the testing process for unfamiliar researchers to the IRT. The conversion of the item difficulty from CTT to IRT using the normal distribution curve mention above has potential for not only the development of CAT forms but also other applications based on IRT.

In this context, the research aims to evaluate the effectiveness of the transformation from the CTT-based *p* statistic to IRT-based *b* parameter using the normal distribution curve in terms of a CAT simulation. And due to fact that the focus of this study is on converting CTT item difficulty statistic to the *b* parameter, the present study is limited with the Rasch model since all the parameters but *b* are constant.

Methods

Data

In R (R Core Team, 2020), using the *genDichoMatrix* function of the catR package (Magis & Barrada, 2017; Magis & Raiche, 2012), 10, 50, 100, 250, 500 and 1000-item pools were created sequentially. The item pool was created according to the Rasch model, accordingly, the item discrimination parameter *a* was accepted as *a = 1.00*, the pseudo-guessing parameter as *c = 0.00*, and the asymptote parameter as *d = 1.00*. Therefore, only *b* parameters were generated using *genDichoMatrix*. Then, 10, 50, 100, 250, 500, and 1000 response patterns were generated for each item pool using the *genPattern* function included in the catR package. Therefore, 36 different response patterns have been studied, including a total of six item pools and six response patterns for each item pool. The rationale behind choosing these conditions, is due to test the performance of the conversion on the data from different sample sizes and item pools. For instance, a teacher may want to convert the item statistics calculated from the data obtained from a classroom as small as 10 and item pool as small as 10. But it is also important to see the performance of the conversion from the data from a larger sample and item pool. While generating response patterns, theta values and item parameters in the item pool were used. Theta values were obtained from a normal distribution whose mean score is 0 and standard deviation is 1.

Data Analysis

Since item parameters were generated according to IRT, item difficulties were first obtained according to classical test theory for data analysis. In this regard, item difficulty values were calculated by finding the means of each item in 36 response patterns. Then, those item difficulties were converted to standard z score using the following function below, and these scores were accepted as *b* parameter according to IRT. The item difficulty parameters obtained according to classical test theory are demonstrated with *p*; item difficulty parameters converted from classical test theory to item response theory with b_p, and the item difficulty parameters obtained according to the item response theory were indicated with *b*. The following function is used to obtain b_p:

$$b_p = 0 - qnorm(colMeans(var)) \qquad (4)$$

This function simply takes the *p* parameter as a percentage of the area under normal distribution and the z value corresponding to the percentage indicated by the *p* parameter as the *b* parameter according to IRT. In this function, *colMeans (var*) calculates the means of columns. In other words, the item difficulty value of each item is calculated. For example, in this function, if 0.065 is used instead of *colMeans (var),* 1.51 is obtained, and this value is in accordance with the example given after the Equation 4. Similarly, when 0.50 is written instead of *colMeans (var),* the function gives the output as 0. In other words, for $p = 0.50$, $b_p = 0$ is obtained.

Following the parameter transformations, *b*, *p*, and b_p parameters were obtained sequentiallyly, for each response pattern. At this stage, *Inf* and *-Inf* values were obtained during the b_p conversion, especially when the sample size was 10. To avoid errors in simulations, *Inf* values were changed into 6 while *-Inf* values were changed into *-6*.

A CAT simulation was conducted with both *b* and b_p parameters. In the simulation, Maximum a Posteriori (MAP) was used as an ability estimation method and Maximum Fisher Information (MFI) as item selection method. In the first item selection, theta was assumed as 0.00 and the simulation was terminated when the standard error value was below .40. The simulations were carried out via the *simulateRespondents* function included in the *catR* package.

According to the simulation results, when *b* and b_p were used, the average number of items used in the simulation, the correlation coefficients between the full-item estimated theta and theta levels estimated by CAT, and bias and RMSE values were compared and the seed value set as 26 for the item and response generation, and CAT simulations.

Results

According to the results of a total of 72 CAT simulations using the *b* and b_p parameters for a total of 36 data sets, correlations between theta values estimated using *b* and b_p parameters are shown in Figure 4.

Figure 4.

Correlation coefficients between theta levels estimated from CAT simulations using b and b_p

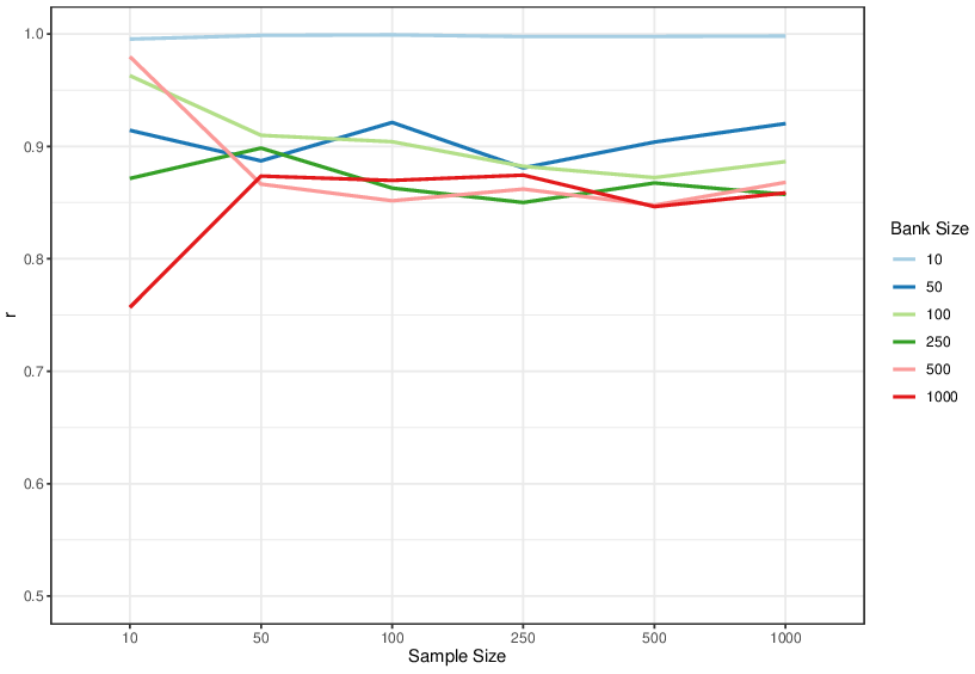
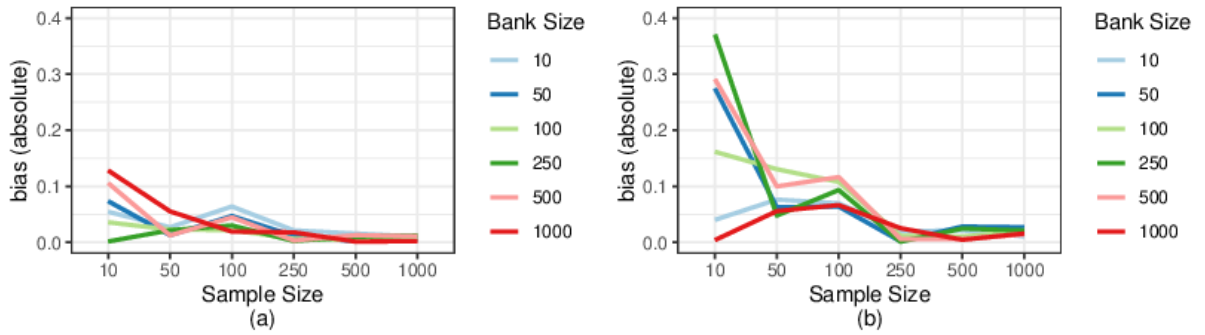


Figure 4 shows that the correlation coefficients were above .80 for almost all cases. When the item pool size was 10, the correlation coefficients for all sample sizes were close to 1.00. This is because the simulation cannot reach the .40 standard error used for the termination rule below 10 items. While the highest correlation coefficient was obtained with a pool of 50 items, other item pools were found to have around .85 correlation coefficients, especially in samples of 50 respondents and above. This indicates that the IRT-based *b* parameter or CTT-based b_p parameter can perform similar theta estimation in CAT simulation.

The bias values of theta estimates were analyzed for both *b* and b_p, and the values obtained for all item pools and sample sizes are presented in Figure 5. For clarity, bias values are analyzed as absolute values.

Figure 5.

Bias values from CAT simulations using b and b_p



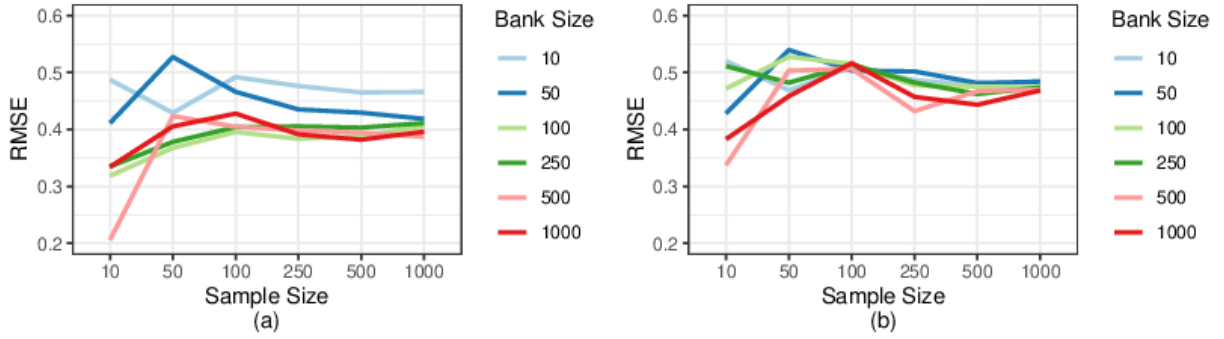
(a) with b parameter; (b) with bp parameter

Although it is seen that both methods have high bias values in small sample sizes, it is understood that the IRT-based parameter estimates with lower bias. Especially when the sample size is 250 and above, the bias value approximates to zero for the theta levels estimated by IRT-based parameters. A decrease in bias value because the sample size increased was also observed in the difficulty parameter obtained by CTT conversion. Similarly, when the sample size is 250 and above, the bias value drops below 0.05.

RMSE values of ability estimates were also analyzed for the whole item pool and sample sizes (see the plots in Figure 6).

Figure 6.

RMSE values from CAT simulations using b and b_p



(a) with b parameter; (b) with bp parameter

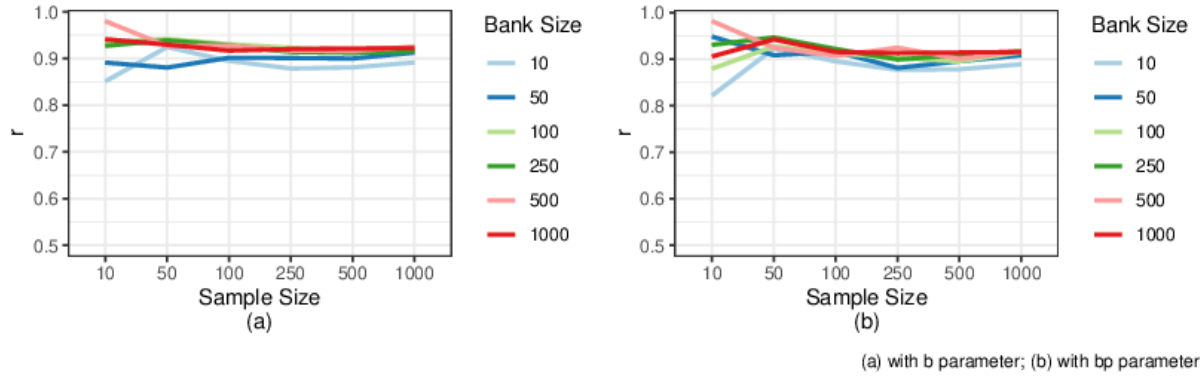
When RMSE values are analyzed, it is seen that, similar to bias, RMSE values of the ability level estimated by IRT-based difficulty parameter are lower. However, in cases where the sample size is 250 and above, it is seen that the RMSE value goes below .50 in both methods.

The correlation coefficients between the ability levels estimated by CAT simulations using both *b* and b_p parameters and the ability levels estimated from all items were analyzed and shown in Figure 7.

Figure 7 shows that the CAT simulations using IRT-based *b* and b_p converted from CTT have similar correlation coefficients between full-theta estimates and CAT estimates. Although the correlation coefficients obtained for 10 respondents vary according to the item pool sizes, it is seen that the correlation coefficients between all theta and estimated theta values with the sample sizes above 50 are around .90.

Figure 7.

Correlation coefficients between CAT simulations using b and b_p and theta levels estimated from all items

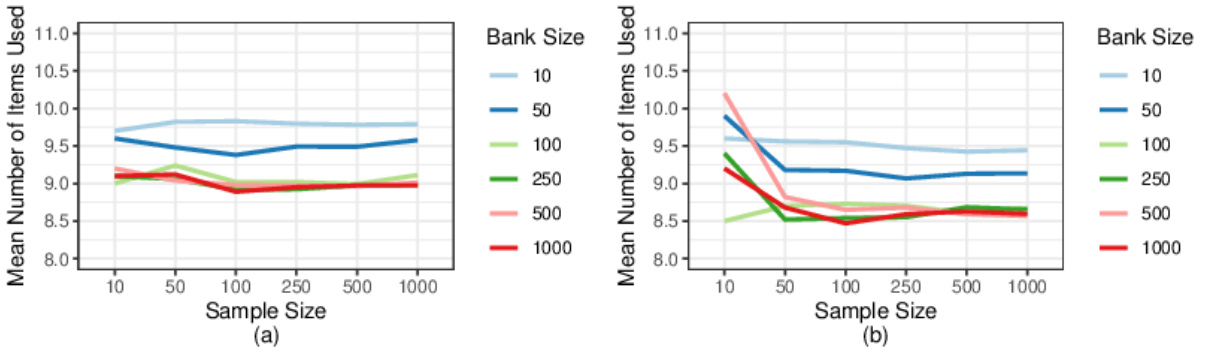


(a) with b parameter; (b) with bp parameter

The average numbers of items in which CAT simulations are terminated for both parameters are given in Figure 8.

Figure 8.

Average numbers of items in which CAT simulations are terminated using b and b_p parameters



(a) with b parameter; (b) with bp parameter

It is seen in Figure 8 that in the CAT simulation using the *b* parameter, the entire item pool is used when the item pool size is 10. On the other hand, in cases where the item pool size is 100 and above, it is seen that the simulation terminates with a similar number of items for both *b* and b_p parameters.

Discussion

This study aims to investigate the effectiveness of transforming CTT-based *p* statistic into IRT-based *b* parameter by utilizing the area under the normal distribution curve in the manner of CAT simulation.

It is seen that the converted b_p parameter has a higher bias and RMSE value than the *b* parameter in CAT simulation. However, it was found that bias and RMSE values in the simulations using b_p also decreased, especially when the sample size was 250 and above. On the other hand, while the correlation coefficients between the estimates were found to be around .85, the correlation coefficients between the ability levels estimated by CAT and the ability levels estimated from all items were found to be around .90 when both *b* and b_p parameters were used. In both cases, the simulation terminated with less than 10 items.

All these findings reveal the potential of b_p converted from CTT into IRT in IRT-based studies and supported by previous studies (Kohli et al. 2014; Raykov & Marcoulides, 2016). Simulation results are more effected by sample size rather than item pool size (except item size was 10) which is matched with the Kohli, Koran, & Henn (2014). Although the findings show that the b_p parameter is not as effective as the *b* parameter, the similarity of CAT simulation results is promising. Especially due to COVID-19 pandemic, the practicality of measurement and assessment processes in distance education has become even more important. In this process, tailored test solutions such as CAT are beyond being available to educators who are not particularly familiar with IRT.

In this context, it is expected that CAT applications can be developed by easily converting parameters from CTT to IRT with the proposed conversion. Practically, a teacher who applied a test to 250 students can convert the *p* statistic to *b* parameter and use the items in a CAT form. In addition to that, converted *b* parameters can be used to kickstart an operational CAT application, then make the IRT based calibrations as data grows. On the other hand, the data used in the research were produced in accordance with IRT assumptions with catR package. Investigating the performance of the b_p parameter where IRT assumptions are not met, as well as applying real data-based post-hoc CAT simulations, will provide a deeper understanding to see how effective the transformation is. In addition, the transformation applied in the research assumed that student ability is normally distributed. Further studies are required to be conducted on how violating this assumption may affect the b_p parameter and the results of the analysis.

Acknowledgement

This study is supported by Pamukkale University Scientific Research Projects Committee. Project No: 2021BSP008. The preliminary results of this study were presented in IACAT 2022, Frankfurt, Germany.

Declarations

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

Ethical Approval: The data was simulated; thus ethical approval is not required.

References

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29-51. https://doi.org/10.1007/BF02291411

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* Cengage Learning.

Çelen, Ü., & Aybek, E. C. (2013). Öğrenci başarısının öğretmen yapımı bir test ile klasik test kuramı ve madde tepki kuramı yöntemleriyle elde edilen puanlara göre karşılaştırılması. *Journal of Measurement and Evaluation in Education and Psychology*, *4*(2), 64-75. https://dergipark.org.tr/tr/pub/epod/issue/5800/77213

De Ayala, R.J. (2009). *The theory and practice of item response theory.* The Guilford Press.

Hambleton, R., & Swaminathan, R. (1985). *Fundementals of Item Response Theory.* Sage Pub.

Hambleton, R., Swaminathan, R., & Rogers, H.J. (1991). *Fundementals of Item Response Theory*. Sage Pub.

Kohli, N., Koran, J., & Henn, L. (2015). Relationships among classical test theory and item response theory frameworks via factor analytic models. *Educational and psychological measurement*, *75*(3), 389–405. https://doi.org/10.1177/0013164414559071

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Routledge.

Magis, D. & Barrada, J.R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. *Journal of Statistical Software, Code Snippets*, *76*(1), 1-19. https://doi.org/10.18637/jss.v076.c01

Magis, D. & Raiche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, *48*(8), 1-31. https://doi.org/10.18637/jss.v048.i08

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. https://doi.org/10.1007/BF02296272

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series, 1992*(1), i-30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Pitman, J. (1993). *Probability (6th Edition).* Springer.

R Core Team (2020). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. URL https://www.R-project.org/.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 321-333). University of California Press.

Raykov, T., & Marcoulides, G. A. (2016). On the relationship between classical test theory and item response theory: From one to the other and back. *Educational and psychological measurement*, *76*(2), 325–338. https://doi.org/10.1177/0013164415576958

Reckase, D. (2009). *Multidimensional Item Response Theory.* Springer.

Samejima, F. (1996). *Polychotomous responses and the test score*. The University of Tennessee.

van der Linden, W. J. & Glas, G.A.W. (2022). *Computerized Adaptive Testing: Theory and Practice.* Kluwer Academic Publishers.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(2); 128-153

Examination of Differential Item Functioning in PISA 2018 Mathematics Literacy Test with Different Methods

Emre KUCAM* Hamide Deniz GÜLLEROĞLU**

Abstract

This study aims to determine whether the PISA 2018 Mathematical Literacy test items show differential item functioning (DIF) according to gender and parental education level. The sample of the study consisted of a total of 521 students who participated in the practice in Turkey and answered the booklets numbered 1 and 7. The research was conducted on a total of 45 items in these booklets. In this study, the Mantel-Haenszel (MH), Logistic Regression (LR), and Rasch Tree (RT) methods were applied to determine the items showing DIF regarding the gender variable. As a result of the analyses, it was determined that two items in the 1st booklet showed DIF in favour of girls, and an item in the 7th booklet that was common with the 1st booklet showed DIF. This item showed DIF in common for all three methods according to the DIF analyses performed separately by the Mantel Haenszel, Logistic Regression, and Rasch Tree methods. As a result, an item showing DIF in favour of girls was determined with both the MH and LR methods in the 1st and 7th booklets. In addition, when the items in booklets 1 and 7 were examined to see whether they show DIF according to parental education level, it was concluded that an item in booklet 1 was easy for students whose mother's education level was high school, university, and above, but difficult for students whose mother's education level was high school or below.

Keywords: DIF, Logistic Regression, Mantel-Haenszel, PISA, Rasch Tree

Introduction

International research on effective schools and quality research in education regarding developing countries are of great value as sources of information for creating an effective education system (Karip & Köksal, 1996). PISA (Programme for International Student Assessment), which is expressed as the largest international organization that includes all this research, aims to establish the sustainable development of the participating countries with the feedback it gives based on the comparison of the educational statuses of the countries. In this way, a reliable system that is constantly developed, dynamic, effective, and efficient is created. One of the most important stages of this system is the test development. In addition to including important steps to be carried out, the main purpose the test development process is the estimation of validity and reliability. Cronbach (1990) defined the concept of validity as the process of collecting evidence in order to determine the situation of measuring the structure that a measurement tool aims to measure. In line with this, it can be stated that if difficulties are encountered and/or errors are observed in measuring the structure that a measurement tool aims to measure, suspicious situations will arise regarding the quality of the evidence collected. In other words, the error involved in the measurement reduces the validity. If this error is produced systematically and if this error produces results in favour of or against the group/groups taking the item/test, it can be said that this situation creates bias. These results are expressed using two different concepts: test bias and item bias. If the probability of a group answering an item correctly is less than that of another group taking the test due to some characteristics of the item or the test conditions unrelated to the purpose of the test, it is called item bias (Zumbo, 1999). Bias can be defined as a systematic error in test scores depending on a group of individuals (Camilli & Shepard, 1994). To rephrase, in both cases, not all

* PhD student., Ankara University, Faculty of Educational Sciences, Ankara-Turkey, emrekucam@gmail.com , ORCID ID: 0000-0002-4283-7103
** Associate Professor, Ankara University, Faculty of Educational Sciences, Ankara-Turkey, denizgulleroglu@yahoo.com , ORCID ID: 0000-0001-6995-8223

To cite this article:
Kucam, E., & Gülleroğlu, H. D. (2023). Examination of differential item functioning in PISA 2018 mathematics literacy test with different methods. *Journal of Measurement and Evaluation in Education and Psychology, 14*(2), 128-153. https://doi.org/10.21031/epod.1122857

Received: 29.05.2022
Accepted: 22.06.2023

individuals taking the item/test are equal on that item/test, which causes the expected measurement results to change against or in favour of a particular group. To reveal this situation, it is important to determine the bias of the measurement tools. While doing this, Differential Item Functioning (DIF) must first be determined by statistical means. DIF is the differentiation of the probability of answering an item correctly among individuals at the same ability level but in a different group. This possible difference should arise from the properties of the items, not from the properties of the subgroups. If an item contains DIF, there is a possibility of bias, however, if an item is biased, it definitely contains DIF. In other words, DIF is necessary but not sufficient for the item bias (Zumbo, 1999). For this reason, it is determined whether an item shows DIF first and then it continues the bias analysis. There are several methods for determining DIF. These methods are summarized below as IRT and CTT-based methods.

Mantel-Haenszel (MH), Simultaneous Item Bias (SIBTEST), and Logistic Regression (LR) methods are examined under the most widely used Classical Test Theory (CTT). On the other hand Lord's Chi-square, Raju's Area Measures and Likelihood-Ratio methods are most used under the Item Response Theory (IRT) (Ellis & Raju, 2003). In the methods examined under CTT and IRT, two types of groups are referred to as reference and focal groups. The focal group is considered to be the disadvantaged group, and the reference group is the group that is advantageous over the focal group. The differentiation of these two groups with respect to each other is determined with statistical methods. DIF is examined under two headings as uniform and non-uniform differential item functioning. Many of the DIF detection methods are designed to reveal the uniform DIF (Jodoin & Gierl, 2001). The uniform DIF is the consistently high level of answering the examined item correctly at all ability levels in a particular group. On the other hand, the non-uniform DIF is the case in which the examined item works in favour of one group in a certain ability level range, while it works in favour of the other group in another ability range (Osterlind & Everson, 2009). Regarding the commonly used DIF determination methods, both Potenza and Dorans (1995) and Alatlı and Şenel (2020) state the theory they are affiliated with, the possibilities of determining uniform or non-uniform DIF, and the number of groups that can be compared in the method as shown in Table 1.

Table 1
Methods of Determining DIF according to Theory, Number of Groups, and Type

Theory	DIF Determination Method	Number of Groups	Uniform/Non-uniform
CTT	Breslow-Day chi-square	2	Non-uniform
	Mantel-Haenszel	2	Uniform
	Simultaneous Item Bias Test-SIBTEST	2	Uniform
	Standardization	2	Uniform
	Transformed Item Difficulties	2	Uniform
	Logistic regression	2	Both
	Generalized logistic regression	>2	Both
	Generalized Mantel Haenszel	>2	Uniform
IRT	Likelihood-Rate Test	2	Both
	Lord's chi-square Test	2	Both
	Raju's Area Measures	2	Both
	Generalized Lord's chi-square Test	>2	Both

Different DIF determination methods have also emerged with the studies conducted after the methods specified in Table 1 were applied. Since the focal and reference groups are predetermined for the methods in the table, these methods are insufficient in determining other potential variables (Zhang, 2009). In addition, the methods in the table focus on only one variable in each implementation, which has the limitation of not being able to focus on the related variables together, especially in large-scale evaluations. The Rasch Tree (RT) method developed for this limitation is one of the new IRT-based methods. The RT method has distinguished among the DIF determination methods because it focuses

on multiple variables together. From this point of view, it can be said that the RT method, which focuses on more than one variable, is more useful than the MH method, which focuses on a single variable. In determining DIF with the Mantel-Haenszel (MH) method, which is one of the methods that deal with only one variable in each application, focal and reference groups are divided into skill or competence layers based on the total test scores. Then, a chi-square probability table is prepared for each skill layer. In the table, the frequencies of correct and incorrect answers are expressed for the groups in each skill layer. The information generated for each skill layer is given in Table 2.

Table 2

Chi-Square Table for Each Skill Layer

Group	Correct Answer	Incorrect Answer	Total
Reference Group	A_j	B_j	n_{Rj}
Focal Group	C_j	D_j	n_{oj}
Total	m_{1j}	m_{0j}	T_j

The ΔMH value is obtained as a result of multiplying the logarithm of the likelihood ratio (αMH) reached with the (∑j Aj Dj / Tj) / (∑j Bj Cj / Tj) operation by -2.35. The DIF levels for these values provided by Zieky (1993) are presented in Table 3.

Table 3

The equivalent of DIF Levels for ΔMH Values

Level of DIF	Condition	Explanation
A	|ΔMH|<1	No or negligible level of DIF
B	1≤|ΔMH|<1.5	Medium Level
C	|ΔMH|≥1.5	High Level

When ΔMH is positive, it is accepted that the items work in favour of the reference group, and when it is negative, it is considered that the items work in favour of the focal group. Another method also used in this study is the Logistic Regression method. Zumbo and Thomas (1997) stated that the 2-degrees-of-freedom chi-square test in the logistic regression should be considered together with the effect size in order to determine DIF. When DIF is determined in large samples without effect size, even insignificant effects may seem statistically significant. In this context, it is recommended to use the ΔR^2 effect size measurement, which is defined as the R^2 difference between the regression models created (Zumbo, 1999). The DIF levels regarding the ΔR^2 effect size values are suggested by Jodoin and Gierl (2001) as follows:

Table 4

The equivalent of DIF Levels for ΔR^2 Values

Level of DIF	Condition	Explanation
A	ΔR^2<.035	No or negligible level of DIF
B	.035≤ ΔR^2 <.070	Medium Level
C	ΔR^2≥.070	Significant Level

When the studies using the MH and LR methods are examined, it is seen that especially large-scale evaluations are studied and different results can be obtained in the same samples (Arslan, 2020; Ayan, 2011; Doğan & Öğretmen, 2008; Gök, Kelecioğlu & Doğan, 2010; Ozarkan, Kucam & Demir, 2017; Schwabe et al., 2014; Şenferah, 2015). It can be said that one of the reasons why different results can be

obtained with the same sample under different methods is the sample size. In these studies, DIF levels are determined based on several variables. The DIF levels are determined according to gender, ethnicity, disability, item type, socioeconomic level, mother tongue, country, content of tests, and affective characteristics (motivation, etc.). Test lengths and sample sizes may also be effective on these variables.

Another method also used in this study is the RT method. In the Rasch model, some of the methods used to determine DIF are for determining DIF in the items, and some are for determining general fit statistics. These methods are designed to compare the parameters of the predefined focal and reference groups. With these methods, it is determined which items may be difficult or easy to answer in which groups, and an opportunity is created to make inferences about what precautions can be taken in these cases. Latent class methods, which have a different understanding from these methods, enable DIF to be determined in groups that have not been defined beforehand and have not been determined to be a possible source of DIF (gender, ethnicity, etc.). Such methods are used in the first stage of the analysis as it is difficult to interpret the groups showing DIF with these methods. Then, the latent classes are tried to be defined. The RT method, on the other hand, combines these two types of DIF determination approaches and reveals a DIF determination method based on the iterative separation technique. In this way, by identifying the groups showing DIF that have not been identified before, direct comments can be made about these groups. It also provides a wide range of opportunities regarding the identification of the DIF sources. The following steps are followed in the RT method (Strobl, Kopf & Zeileis, 2015):

1. First, the item parameters are estimated by including the entire sample.
2. It is statistically tested whether the item parameters differ by considering each covariant.
3. If there are significant instabilities in the covariates at the item parameters, the sample is separated along the covariant with the strongest indecision, and the cut-off point is determined.
4. The process mentioned above is repeated until there is no significant indecision.

In the study of Altıntaş and Kutlu (2019), in which this method was also used, the DIF status according to the country and gender variable was examined by using the data of 615 (Azerbaijan, Bulgaria, and Syria) out of 2476 individuals who took the Ankara University Foreign Student Exam in 2017. In this study, in which the analyses were carried out using the RT method, DIF was determined in 16 items according to the countries. In addition, it was concluded that the exam did not include DIF according to gender. Similarly, the RT method and LR and Rasch methods among the traditional methods were compared regarding the identification of DIF according to gender, ethnicity, socioeconomic level, and mother tongue in Liu's (2017) study with a data set of 731 students studying at the eighth grade of the 2011 TIMSS mathematics subtest in the USA sample. It was determined that 6 items showed DIF in favour of girls with the LR method, 4 items showed DIF in favour of girls and 1 item in favour of boys with the Rasch method, and 2 items showed DIF in favour of girls, and 3 items in favour of boys with the RT method. While 2 of these items for which DIF was determined according to gender were common in all three methods, the results were obtained in favour of girls with the LR method and in favour of boys with other methods in 1 item. In addition, DIF was determined in 7 items related to ethnicity with the RT method. As a result, it was stated that the RT method generated similar results with the LR and Rasch methods in determining the items containing DIF.

Karami, Gramipour, and Minaei (2021), investigated the factors that reveal the differentiation in test items using the Rasch tree method in their study. Data from a special test of the Amin University of Law and Applied Sciences were used to answer the research questions. The data of this simulation study, in which 2414 people participated, were analysed with the DIFtree package in R software, in which the Rasch tree method was used. In the special examination of Amin University of Law and Applied Sciences, it was observed that 9 items showed DIF and most of these items were in the mathematics group, and these items showed DIF according to the age of 18 (second category) and 19 (first category). This study shows that the Rasch tree method is effective in determining the differentiation in test questions.

Asamoah (2020), administered the 10-item, 5-point Likert-type Perceived Stress Scale to 500 participants through a platform called MTurk, which matches practitioners and participants, in his

master's thesis. The data were analysed according to the age, gender, marital status, employment status, social media use, and race variables.According to these data, DIF for gender, ethnic group, employment and social media variables was determined in one item. It was determined that DIF could not be found for the variables of age, marital status and number of children. It was found that the number of items for which DIF was detected by the MH (Mantel-Haenszel) and LR (likelihood ratio test) methods were equal to each other.

In her doctoral thesis, Başman (2017) examined the interactions of the variables of motivation, self-efficacy, and anxiety on the mathematics test items within the scope of changing item function in order to understand the sources of the differences in mathematics achievement of the students participating in the PISA 2012 application. The sample of the research consists of 1084 students who participated in the practice in Turkey. Data were analysed using the Rasch Tree Method (RAY) in the Psychotree package in the R program and the Logistic Regression Likelihood Ratio Method (LROOY) in the Lordif package program. It was determined which items showed DIF according to gender. It was also observed that items showing DIF according to gender determined by RAY showed DIF according to the interaction between gender and intrinsic motivation. It was observed that the DIF status of the items changed both according to a certain threshold value of the girls' intrinsic motivation score and according to the interaction between gender and self-efficacy of mathematics items. It was observed that the DIF status of the items changed according to a certain threshold value of the self-efficacy score of the girls.

In their study, Strobl, Kopf, and Zeileis (2015) suggested the use of the newly named Rasch Tree Method to determine DIF in samples showing DIF but whose group could not be determined beforehand. With this method, DIF in a numerical covariate cannot be overlooked because the numerical covariates (like age) have lots of cutpoints. The exact cutpoint does not need to be pre-specified, the decision is made from the data. This is an advantage of the Rasch tree method.

When all these studies are considered, it is seen that the DIF analyses for large-scale evaluations are mostly made separately on the basis of a single variable and the items containing DIF are determined accordingly. In this case, when the error included in the DIF analysis for each variable in a test is considered, it can be said that the determination of all the variables to be examined whether they are the source of DIF in a single analysis and with a single error will contain statistically fewer errors. In addition, the presence of DIF is the most important threat that may reduce test validity. This type of data obtained from the large-scale exams is thought to be important in terms of identifying the possible sources of DIF.

When the literature is examined, it is seen that there is evidence for the presence of many items showing DIF in the large-scale tests (PISA, TIMSS, PIRLS, etc.) as a result of the analyses made on these tests (Ayan, 2011; Liu, 2017; Schwabe et al., 2014). The presence of the items with DIF even in these applications that fully comply with the test development stages, or more accurately, the presence of items that may constitute bias in these tests arouses suspicion and curiosity about the situation in the national exams prepared without following the test development stages. This is clearly observed in the analyses of the exams held within the scope of the national exams. The methods used are of great importance at the point of questioning the validity of these analyses. In addition, the MH method is frequently used, because it is easy to use and understand, and also because it allows testing the null hypothesis and provides an index showing the size of the DIF (Millsap & Everson, 1993). On the other hand, the LR method can be applied to items that fall into more than one group and ranking scale, and can diagnose regular and irregular DIF (Agresti, 2012). In the RA method, on the other hand, groups showing unidentified DIF can be identified, and direct comments can be made about these groups (Strobl, Kopf & Zeileis, 2015). Therefore, in this study, the DIF level of the items was compared using the LR and RT methods, in addition to the frequently preferred MH method. In this respect, it is expected that this research will contribute to the literature in terms of revealing the weaknesses and strengths of these three methods, determining the items with DIF using these methods in the national exams, and promote studies to be conducted on bias.

In addition, it is seen that DIF determination methods based on CTT and IRT for large-scale evaluations are used extensively in the literature (Altıntaş and Kutlu, 2019; Chen and Thissen, 1997; Doğan and Öğretmen, 2008; Gök, Kelecioğlu and Doğan, 2010), however, the RT method is used relatively less

(Başman, 2017; Liu, 2017; Strobl, Kopf and Zeileis, 2015). In this study, the Rasch Tree method was used, since it handles multiple variables together and the number of subgroups of the parent education level variable is more than 2. It is of great importance to reveal the validity of the measurement tools of PISA, which is one of the large-scale tests, and to realize this with the least amount of error. By comparing the methods based on both the observed score and the IRT, the differences and similarities of the methods were tried to be determined. Since the studies comparing these three methods mentioned above are very few in the literature, the study is important in this respect. The purpose of this research is to determine the differential item functioning (DIF), which varies according to gender and education level, of the PISA 2018 mathematical literacy test items with various methods, in the Turkish sample. For this purpose, the following questions were answered:

1. Do the items in the PISA 2018 mathematics subtest show DIF in the analyses made with the MH, LR, and RT methods according to gender?
2. Do the items in the PISA 2018 mathematics subtest show DIF in the analyses made with the RT method according to the education level of the parents?
3. Are the results regarding DIF coherent in the analyses conducted with the MH, LR and RT methods according to gender?

Method

Research Model

This study aims to determine whether the items in the Turkey sample of the PISA 2018 Mathematical Literacy test show differential item functioning (DIF) according to gender and parental education level and compare the DIF determining methods LR, RT, and MH. In this respect, the research is suitable for the descriptive research as it aims to describe the existing situation. Descriptive research is a research approach that aims to describe a situation as it is (Karasar, 2017).

Population and Sample

The population of the research consists of a total of 521 students, 255 of whom answered booklet number 1 and 266 students who answered booklet number 7 in the PISA 2018 Turkey sample consisting of 6890 people. Booklets 1 and 7 were chosen, because they contain the most common items compared to other booklets. The descriptive statistics regarding the population and sample of the PISA 2018 Turkey application are presented in Table 5.

Table 5
Distribution of PISA 2018 Turkey Population and Sample according to Gender, Class, and School Type

Variable	Group	Population		Sample	
		f	%	f	%
Gender	Boy	3494	50.7	262	50.3
	Girl	3396	49.3	259	49.7
	Total	**6890**	**100**	**521**	**100**
Class	7th Grade	3	0.05	1	0.15
	8th Grade	19	0.3	1	0.15
	9th Grade	1295	18.75	101	19.4
	10th Grade	5360	77.8	401	77
	11th Grade	207	3	17	3.3
	12th Grade	6	0.1	0	0
	Total	**6890**	**100**	**521**	**100**

School Type	Middle school	22	0.2	2	0.3
	General High School (Anatolian High School, Imam Hatip High School, Sports/Fine Arts High School and General High School)	3998	58	307	59
	Science High School	226	3.4	17	3.3
	Social Sciences High School	228	3.4	17	3.3
	Vocational Technical High School	2416	35	178	34.1
	Total	**6890**	**100**	**521**	**100**
Mother's Education Level	Primary School Dropout	7704	10.2	63	12.1
	Primary School	1936	28.1	155	29.8
	Middle School	1519	22	111	21.3
	High School	1079	15.8	86	16.6
	Undergraduate and Above	1580	22.9	100	19.2
	Missing	72	1	6	1
	Total	**890**	**100**	**521**	**100**
Father's Education Level	Primary School Dropout	72	3.9	21	4.1
	Primary School	506	21.8	109	21
	Middle School	1887	27.4	162	31.2
	High School	1492	21.7	100	19.2
	Undergraduate and Above	1653	24	121	23.3
	Missing	80	1.2	8	1.2
	Total	**6890**	**100**	**521**	**100**

Data

In the PISA 2018 Mathematical Literacy test for Turkey sample, 82 items applied in the computer environment were distributed into 36 booklets and used. While preparing the data, twenty-three questions were used in the booklet number 1, and twenty-two questions were used in the booklet number 7. These questions measure mathematical literacy and 11 of the questions in these two booklets are common. The DIF analyses were conducted on these items. The reason for considering booklets 1 and 7 is that the number of common items is the highest compared to other booklets. Dichotomous items were scored as 1-0, while partially scored items were scored as 1 for fully correct answers; it was converted to 0 points for partially correct, incorrect, and blank answers. The 1st and 7th booklets in the PISA 2018 Mathematical Literacy Turkey sample consist of items that are common, partially scored, and scored as dichotomous (1-0). The numbers of common and non-common items selected from these booklets are presented in Table 6.

Table 6

Distribution of Booklets Selected from the PISA 2018 Mathematics Subtest according to Common and Non-Common Items

Booklet Number	Number of Common Items	Number of Non-Common Items	Total
1	12 (3ps*, 9 ds**)	11	23
7	11 (2ps*, 9 ds**)	11	22

**ps: partial scoring*
***ds: dichotomous scoring (1-0)*

When Table 6 is examined, it is seen that 3 of the 12 common items selected from booklets 1 and 7 are scored as partial (ps) and 9 of them are scored as 1-0 (ds). On the other hand, 11 non-common items are scored as 1-0. In addition, the descriptive statistics of the booklets 1 and 7 used in the study are given in Table 7.

Table 7
Descriptive Statistics of Booklets 1 and 7

Descriptive Statistics regarding the Booklets	Booklet 1		Booklet 7	
	Girl	Boy	Girl	Boy
Number of items	23	23	22	22
Number of students	127	1.28	134	132
Mean score	8.18	8.19	9.13	9.05
Median	8	8	8.5	8
Peak Value	8	9	8	5.10
Standard Deviation	4.67	4.47	4.75	4.94
Skewness	.71	.47	.50	.27
Kurtosis	3.09	2.46	2.52	2.05
Lowest score	0	1	2	0
Highest score	21	20	21	21

As presented in Table 7, the mean scores of the girls in the booklets 1 and 7 were 8.18 and 9.13, respectively, while the mean scores of the boys were calculated as 8.19 and 9.05. The fact that the skewness coefficients were positive in both groups indicates that the distribution of scores is slightly skewed to the right. When the distribution of the mean, mode, and median is examined, it is seen that the values are very close to each other, which indicates that the distribution is very close to the normal distribution. When the mean scores of the girls and the boys in the booklets are examined, it can be stated that the values are very close to each other, in other words, the difference in achievement between the girls and the boys in the PISA 2018 Mathematics subtest for booklets 1 and 7 is almost non-existent. Before proceeding to the DIF analysis, the data set was examined in terms of missing values and outliers.

For the DIF analyses to be performed with IRT, it was appointed whether the data obtained from the booklets met the IRT assumptions. These assumptions are unidimensionality, local independence, and model-data fit (Lord, 1980).

The unidimensionality of the mathematical literacy items was examined with the Exploratory Factor Analysis (EFA). For this, the assumptions of EFA were tested first. In this context, the outlier, multivariate normal distribution, linearity, and single-multi-collinearity were examined. However, since it is not possible to directly examine multivariate normality, univariate normality and outliers were examined. The fact that there is not a violation of univariate normality also supports multi-variability (Sharma, 1995).

In this study, Shapiro Wilks test was applied to determine whether the data set meets the normal distribution assumption for booklets 1 and 7 and it was concluded that none of the 23 and 22 items in these booklets respectively showed a normal distribution ($p<.05$). The outliers were obtained by examining the Mahalanobis distances ($p<.001$) and multivariate normality. According to Tabachnick and Fidell (2007), the Mahalanobis Distance value should be compared with the $\chi 2$ table value, which accepts the number of independent variables as the degree of freedom. When the Mahalanobis Distance values are examined, it is seen that there is no value exceeding the critical values of $\chi 2(23)=49.72$ for the 1st booklet and $\chi 2(22)=48.26$ for the 7th booklet. This shows that there is no violation of the outlier and the multivariate normality. When the scatter plots in Figure 1 and Figure 2 are examined, it is seen that the data are clustered on a straight line.

Figure 1

Booklet 1 Mahalanobis Distance Values Scatterplot

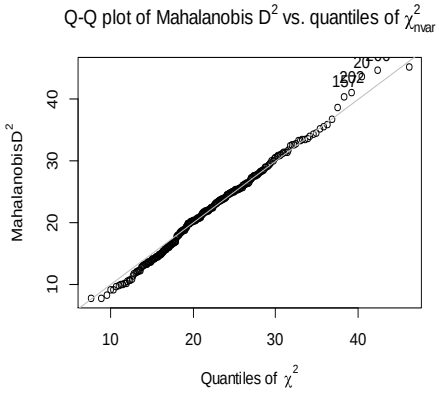
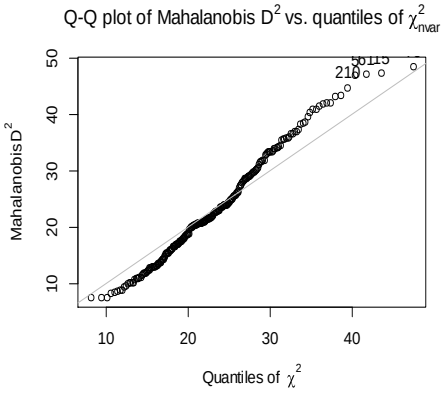


Figure 2

Booklet 7 Mahalanobis Distance Values Scatterplot



Tabachnick and Fidel (2007) stated that the sample size should be 300 or more in order to use factor analytical techniques, but if there is a very strong structure and the representativeness of the group is high, a sample size of up to 150 is acceptable. On the other hand, it is stated by different sources (DeVellis, 2017; Nunnally, 1978; Tavşancıl, 2018) that a sample size of 8-10 times the number of variables/items is sufficient. As the third option, the Kaiser Meyer Olkin (KMO) sample size adequacy test can be applied. In this study, it is noteworthy that the number of students who took the booklets 1 and 7 as the test, namely the sample size, is close to the suggestion of Tabachnick and Fidel (2007) (N1=255, N7=266). (On the other hand, 8-10 times of 23 items makes 184-230, and 8-10 times of 22 items makes 176-220, which shows that this recommendation is more than fulfilled.). As the third option, the KMO test was applied. Since the univariate normal distribution could not be achieved, the KMO test value calculated using the Spearman Rank Differences Correlation matrix was found to be .84 for the 1st booklet and .87 for the 7th booklet. As these values are over .70, it can be stated that the sample size is sufficient for the factor analytical studies.

For the assumption of multicollinearity, the correlation among the variables should be examined and most of the bilateral correlations should be significant (Andy Field, 2012) or Bartlett's Sphericity test can be used. A rough look at the correlation matrix obtained with the Spearman Rank Differences Correlation calculation shows that the pairwise correlations are low but factorable. Bartlett's Test of Sphericity was applied as statistical evidence. As a result, it was determined that the multiple correlations among the variables were statistically significant (Bartlett test of sphericity for the booklet 1; Chi-square=1100.647; df=253 and $p<.05$; Bartlett test of sphericity for the booklet 7; Chi-square=1215.448; df=231 and $p<.05$). In this context, when the correlations among 23 items detected for the booklet 1 in PISA 2018 Mathematical Literacy test are examined, it is seen that the correlations vary between .02 and .26, and when the correlations among 22 items determined for the booklet 7 are examined, it is seen that the correlations vary between .03 and .31. These results indicate that there is no problem of single or multi-collinearity for both booklets.

When multivariate normality, sample size, and the significance of multiple correlations between variables/items were examined, no serious violations were observed that would prevent the use of exploratory factor analysis, provided that the rank-difference coefficient of correlation was used. Thus, it appears that the data met the assumptions of the EFA. In this scope, the EFA was conducted for the 1st booklet (255 people) and the 7th booklet (266 people). Since the univariate normal distribution could not be achieved, the EFA was conducted using the Spearman Rank Correlation Coefficients matrix (Spearman, 1905). For this assumption, it is recommended to examine the eigenvalues and the scree plots of the factors obtained consequently the factor analysis (Cattell, 1966). In this context, the

eigenvalues obtained from the EFA for the 1st and 7th booklet items are shown in Table 8, and the scree plot is presented in Figure 3.

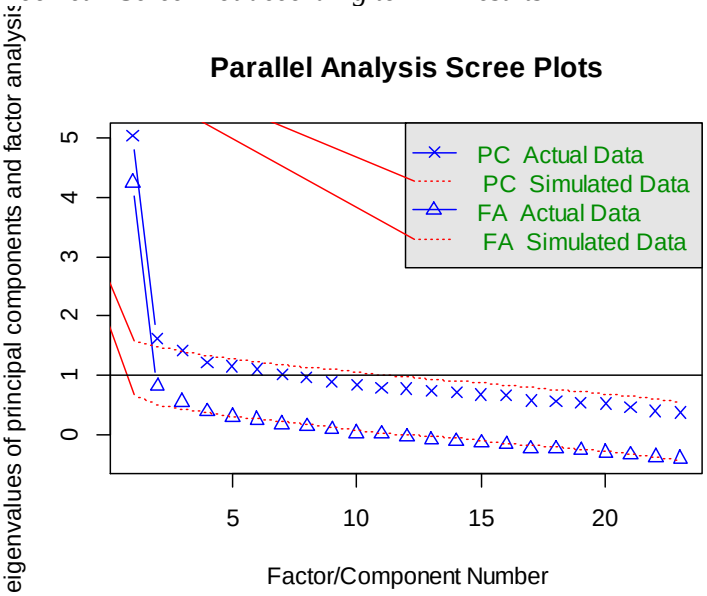
Table 8

Booklet 1 and Booklet 7 Eigenvalues according to EFA Results

Number of factors	Eigenvalues		Variance Explained (%)		Total Variance Explained (%)	
	Booklet 1	Booklet 7	Booklet 1	Booklet 7	Booklet 1	Booklet 7
1	5.03	5.54	22	25	22	25
2	1.63	1.40	7	6	29	31
3	1.41	1.26	6	6	35	37
4	1.21		5		40	

Figure 3

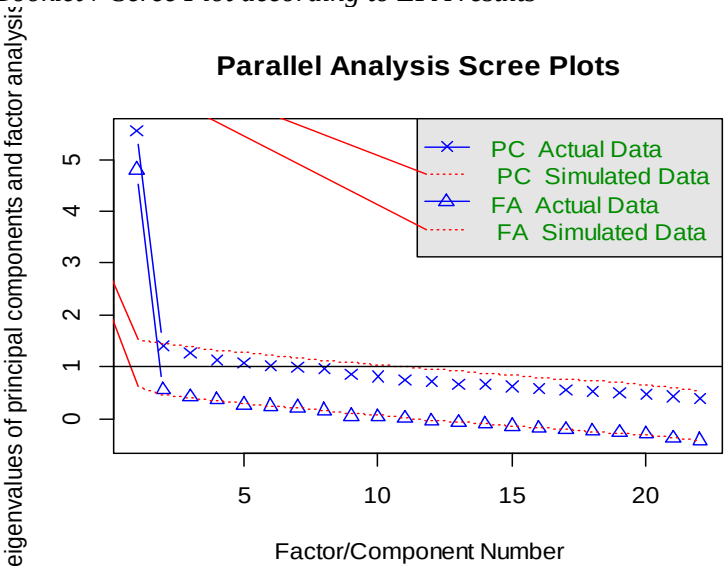
Booklet 1 Scree Plot according to EFA results



When Table 8 is examined, it is seen that the difference in the eigenvalues obtained with the EFA between the first factor of the items in the 1st booklet and the other factors is very large. This shows that the unidimensionality assumption is met (Hambleton & Swaminathan, 1989). When Figure 3 is examined, it is seen that a sharp bend is formed with the decrease after the first factor, which indicates that the contributions of the other factors after the first factor to the variance are close to each other and lower than that of the first factor. Local independence, which is one of the other assumptions, is the situation that the answer given to each item is independent of the answers given to the other items (Crocker & Algina, 1986). To test the local independence, Yen's Q3 statistic was calculated. Accordingly, it can be stated that the Q3 values among all the items in the 1st booklet do not exceed .20 (Chen & Thissen, 1997), and thus the local independence assumption is also met.

Figure 4

Booklet 7 Scree Plot according to EFA results



When Table 8 is examined, it is seen that the difference in the eigenvalues obtained with the EFA between the first factor of the items in the 7th booklet and the other factors is very large. This shows that the unidimensionality assumption is met (Hambleton & Swaminathan, 1989). When Figure 4 is examined, it is seen that a sharp bend is formed with the decrease after the first factor, which indicates that the contributions of the other factors after the first factor to the variance are close to each other and lower than that of the first factor. Accordingly, it can be stated that the Q3 values among all the items in the 7th booklet do not exceed .20 (Chen & Thissen, 1997), and thus the local independence assumption is also met.

To determine the model-data fit and to carry out the analysis based on IRT, it should be determined which of the 1, 2, and 3 parameter logistic models the data set is compatible with. In the 1st booklet, the Log-Likelihood values obtained for each model and the number of compatible items are presented in Table 10.

Table 9

Booklet 1 Log Likelihood Values of IRT Models and Number of Compatible Items

	1PL	2PL	3PL
Log-Likelihood (LL)	3108.266	3041.361	3022.439
Number of Compatible Items	13	19	17

The fact that the items with a p-value greater than 0.05 are compatible with the model also means the acceptance of the null hypothesis. From this point of view, 13 items are compatible with 1PL, 19 items are compatible with 2PL, and 17 items are compatible with 3PL. The difference between the Log-Likelihood values of the models is taken into account in the evaluation of the model data fit. These difference values are given below:

$$LL_{2PL}-LL_{3PL}=18.922$$

$$LL_{1PL}-LL_{2PL}=60.905$$

LL values showing the chi-square distribution were compared with the critical chi-square value according to the number of items for model-data fit. Since there are 23 items in the 1st booklet, the critical chi-square value is $\chi 2=13.09$, and when compared with the differences above, it is seen that the

difference values are greater than the critical value. In this case, it can be said that the test is compatible with the 3PL model. However, when the number of items compatible with the model is examined, it can be said that the test is coherent with the 2PL model since the number of items compatible with the 2PL model is higher.

In the 7th booklet, the Log-Likelihood values obtained for each model and the number of compatible items are presented in Table 10.

Table 10
Booklet 7 Log-Likelihood Values of IRT Models and Number of Compatible Items

	1PLM	2PLM	3PLM
Log-Likelihood (LL)	2999.812	2955.51	2939.772
Number of Compatible Items	16	21	20

The fact that the items with a p-value greater than 0.05 are compatible with the model also means the acceptance of the null hypothesis. From this point of view, 16 items are compatible with 1PLM, 21 items are compatible with 2PLM, and 20 items are compatible with 3PLM. The difference between the Log Likelihood values of the models is taken into account in the evaluation of the model data fit. These difference values are given below:

$$LL_{2PL}\text{-}LL_{3PL}=15.738$$

$$LL_{1PL}\text{-}LL_{2PL}=44.302$$

LL values showing the chi-square distribution were compared with the critical chi-square value according to the number of items for model-data fit. Since there are 22 items in the 7th booklet, the critical chi-square value is $\chi 2=12.33$, and when compared with the differences above, it is seen that the difference values are greater than the critical value. In this case, it can be said that the test is compatible with the 3PL model. However, when the number of items compatible with the model is examined, it can be said that the test is compatible with the 2PL model since the number of items compatible with the 2PL model is higher. In this case, it can be stated that it is appropriate to choose the 2PLM, in which the majority of the items are compatible, as the IRT model for both booklets.

Data Analysis

To obtain the findings for the first and second research questions, the DIF analyses of the items in the 1st and 7th booklets in the PISA 2018 Mathematics subtest were conducted using the MH, LR, and RT methods. The reference and focal groups required for the analyses were created according to the variables of gender, mother's education level, and father's education level. For MH, the "difMH" command in the "difR" package within the R program was used, and the "raschtree" command in the "psychotree" package within the R program was used for RT. The DIF levels of the items showing DIF for MH and the group in favour of which they showed DIF were determined, and the classification system organized by Zieky (1993) was used for these items.

Results

Findings Regarding Differential Item Functioning According to Gender

Whether the PISA 2018 Mathematics subtest showed DIF according to gender was analysed by the MH, LR, and RT methods, respectively. For this purpose, the items in the 1st booklet and then the ones in the 7th booklet were analysed.

DIF Analysis with Mantel Haenszel Method

The analysis results of the items in the 1st booklet obtained with the MH method are presented in Table 11.

Table 11

Booklet 1 Mantel Haenszel Method Results

Item	Chi-Square	Alpha	Delta	p
CM564Q02S	1.027	0.715	0.785	0.310
CM564Q01S	0.000	1.042	-0.098	0.994
CM571Q01S	0.392	1.283	-0.586	0.530
CM603Q01S	1.200	1.463	-0.895	0.273
DM406Q02C	1.123	0.119	5.003	0.289
DM406Q01C	0.006	0.869	0.329	0.938
CM192Q01S	0.715	1.330	-0.671	0.397
CM423Q01S	0.180	1.263	-0.549	0.671
CM496Q02S	0.055	0.882	0.295	0.814
CM496Q01S	0.402	1.335	-0.679	0.525
CM305Q01S	0.001	0.972	0.064	0.974
CM034Q01S	0.015	0.898	0.250	0.900
DM462Q01C	3.703	0.442	1.916	0.054
CM442Q02S	0.003	0.953	0.112	0.951
CM803Q01S	1.070	0.561	1.356	0.300
CM411Q02S	0.117	0.854	0.370	0.731
CM411Q01S	1.509	0.636	1.060	0.219
CM155Q04S	2.789	1.644	-1.168	0.094
DM155Q03C	4.455	0.331	2.596	0.034*
CM155Q01S	0.015	1.010	-0.023	0.902
DM155Q02C	0.141	0.840	0.407	0.706
CM474Q01S	1.911	1.581	-1.077	0.166
CM033Q01S	0.983	1.382	-0.760	0.321

*p<.05

** Bold item codes refer to the same items in booklets 1 and 7.

When Table 11 is examined, it is seen that only the p-value of the item "DM155Q03C" is significant (p<.05). The ΔMH value of this item was compared with the ΔMH threshold values and it was detected at what level the item showed DIF. Negative values of ΔMH may provide an advantage for the reference group and positive values may provide an advantage for the focal group. In this context, it was determined that the item "DM155Q03C" showed DIF at the C level in favour of the girls forming the focal group. In more general terms, only one of the 5 partially scored items in booklet 1 showed DIF. It is necessary to be careful when generalizing that only one item shows DIF. The finding that female students outperform male students on open-ended items is fitted with this situation (Schwabe et al., 2014; Koğar & Koğar, 2019). The analysis results of the items in the 7th booklet obtained with the MH method are presented in Table 12.

When Table 12 is examined, it is seen that the p-value of none of the items is significant. Negative values of ΔMH may provide an advantage for the reference group, and positive values may provide an advantage for the focal group. However, since negative or positive ΔMH was not significant for any item, it was concluded that none of the items in the 7th booklet showed DIF according to gender.

Table 12

Booklet 7 Mantel Haenszel Method Results

Item	Chi-Square	Alpha	Delta	p
CM034Q01S	0.018	0.989	0.024	0.892
DM462Q01C	1.350	1.616	-1.128	0.245
CM803Q01S	0.346	0.750	0.673	0.555
CM411Q02S	0.000	0.964	0.084	0.982
CM411Q01S	0.019	1.009	-0.022	0.890
CM155Q04S	0.034	0.913	0.212	0.853
DM155Q03C	2.053	1.989	-1.616	0.151
CM155Q01S	0.144	0.861	0.351	0.703
DM155Q02C	0.026	1.117	-0.260	0.871
CM474Q01S	1.362	0.682	0.899	0.243
CM033Q01S	0.796	0.739	0.708	0.372
CM447Q01S	0.000	1.045	-0.105	0.996
CM273Q01S	2.252	1.633	-1.152	0.133
CM408Q01S	0.238	0.807	0.502	0.625
CM420Q01S	0.325	1.246	-0.518	0.568
CM446Q01S	0.184	0.804	0.511	0.668
DM446Q02C	0.768	2.698	-2.333	0.380
CM559Q01S	0.000	0.962	0.090	0.985
DM828Q02C	0.009	0.930	0.170	0.923
CM828Q03S	0.312	1.277	-0.576	0.576
CM464Q01S	0.000	1.078	-0.178	0.982
CM800Q01S	1.311	0.543	1.433	0.252

*p<.05

** Bold item codes refer to the same items in booklets 1 and 7.

DIF Analysis Conducted with Logistic Regression Method

The analysis results of the items in the 1st booklet obtained with the LR method are presented in Table 13.

Table 13

Booklet 1 Logistic Regression Method Results

Item	Chi-Square	R^2	Jodoin&Gierl*	p
CM564Q02S	1.987	0.008	A	0.370
CM564Q01S	1.132	0.005	A	0.567
CM571Q01S	0.419	0.001	A	0.810
CM603Q01S	1.709	0.008	A	0.425
DM406Q02C	7.755	0.080	C	0.020**
DM406Q01C	1.870	0.011	A	0.392
CM192Q01S	6.319	0.026	A	0.042**
CM423Q01S	0.510	0.002	A	0.774
CM496Q02S	1.023	0.003	A	0.599
CM496Q01S	1.489	0.005	A	0.474
CM305Q01S	0.834	0.004	A	0.658
CM034Q01S	0.551	0.002	A	0.758
DM462Q01C	2.952	0.013	A	0.228
CM442Q02S	2.362	0.010	A	0.306
CM803Q01S	1.532	0.008	A	0.464
CM411Q02S	0.889	0.004	A	0.640
CM411Q01S	2.001	0.007	A	0.367
CM155Q04S	3.414	0.016	A	0.181
DM155Q03C	4.822	0.026	A	0.089
CM155Q01S	1.065	0.004	A	0.587
DM155Q02C	0.664	0.002	A	0.717
CM474Q01S	3.084	0.012	A	0,213
CM033Q01S	1.713	0.007	A	0,424

** According to the Jodoin and Gierl effect size, $R^2 \geq 0.070$ means there is a high level (C-level) DIF.*

**Bold item codes refer to the same items in booklets 1 and 7.

When Table 13 is examined, it is seen that only the p-values of the items "DM406Q02C" and "CM192Q01S" are significant (p<.05). The R2 value of these items was compared with the Jodoin and Gierl effect size values and it was determined at what level the items showed DIF.

Figure 5 and Figure 6 present the item characteristic curves of the girls and the boys for these items.

Figure 5
Item Characteristic Curve of the Item DM406Q02C

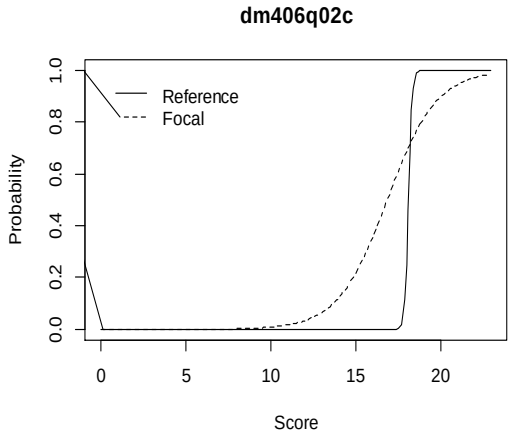
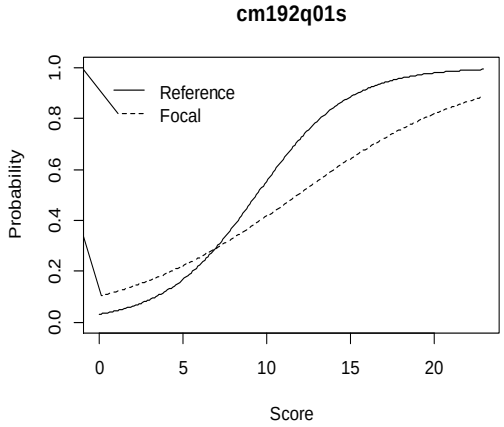


Figure 6
Item Characteristic Curve of the Item CM192Q01S



According to Figure 5, it is seen that the characteristic curve of the item "DM406Q02C" shows DIF at C level in favour of the girls who are in the focal group (R25=.08>.07). When the item characteristic curve is examined, it is seen that the probability of answering the item correctly after 18 points for the reference group boys (reference) increases, and after 10 points for the girls who are the focus group. When a significant DIF is detected for an item, researchers should question whether the DIF actually indicates a bias for the country concerned. That is, it must be decided whether DIF is related to structure (Robitzsch and Lüdtke, 2020). However, since this item cannot be reached, it can be cautiously stated that it is more difficult for men. In Figure 6, on the other hand, it is seen that while the characteristic curve of the item "CM192Q01S" works in favour of the girls (focal) up to about 7 skill levels, it shows DIF in favour of the boys (reference) in the skill group above 7, but the effect size of the DIF likelihood ratio test, which is significant, is at a negligible level (R27=0.026<0.035). The analysis results of the items in the 7th booklet obtained with the LR method are given in Table 14.

Table 14
Booklet 7 Logistic Regression Method Results

Item	Chi-Square	R^2	Jodoin&Gierl*	p
CM034Q01S	2.482	0.009	A	0.289
DM462Q01C	2.233	0.008	A	0.327
CM803Q01S	0.949	0.004	A	0.621
CM411Q02S	2.719	0.012	A	0.256
CM411Q01S	0.887	0.003	A	0.641
CM155Q04S	0.485	0.002	A	0.784
DM155Q03C	7.636	0.044	B	0.022*
CM155Q01S	0.255	0.001	A	0.880
DM155Q02C	0.789	0.002	A	0.674
CM474Q01S	2.515	0.009	A	0.284
CM033Q01S	1.051	0.004	A	0.591

CM447Q01S	1.522	0.005	A	0.467
CM273Q01S	2.796	0.011	A	0.247
CM408Q01S	0.161	0.000	A	0.922
CM420Q01S	0.220	0.000	A	0.895
CM446Q01S	2.711	0.010	A	0.257
DM446Q02C	1.988	0.018	A	0.370
CM559Q01S	0.514	0.001	A	0.773
DM828Q02C	0.182	0.000	A	0.912
CM828Q03S	2.907	0.013	A	0.233
CM464Q01S	0.234	0.000	A	0.889
CM800Q01S	1.910	0.013	A	0.384

***** *According to the Jodoin and Gierl effect size, $R^2 \geq 0.070$ means there is a high level (C-level) DIF.*

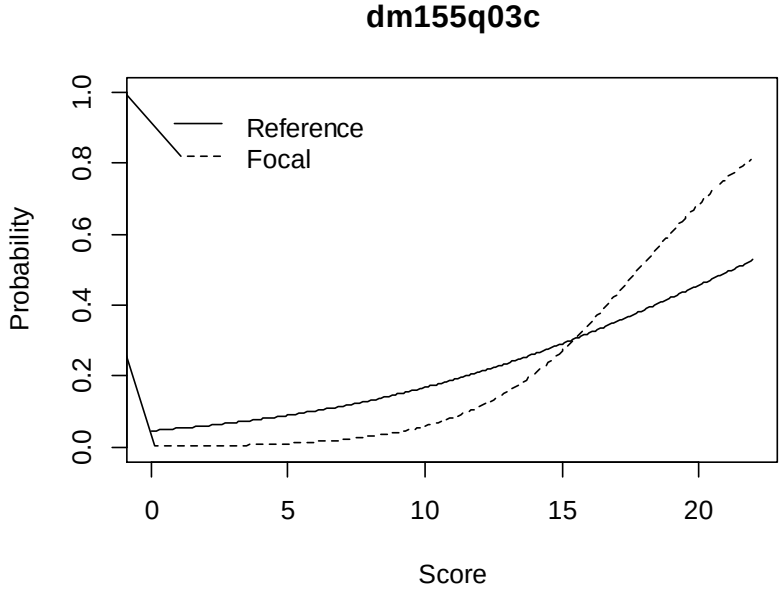
** Bold item codes refer to the same items in booklets 1 and 7.

When Table 14 is examined, it is seen that only the p-value of the item "DM155Q03C" is significant (p<.05). The R2 value of these items was compared with the Jodoin and Gierl effect size values and it was determined at what level the items showed DIF.

Figure 7 presents the item characteristic curves of the girls and the boys for these items.

Figure 7

Item Characteristic Curve of the item DM155Q03C



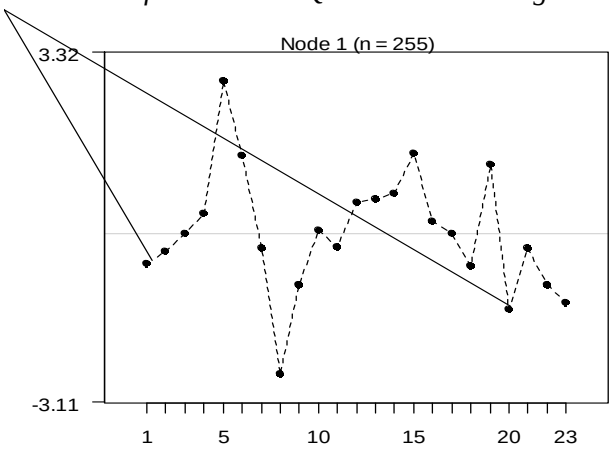
In Figure 7, it was determined that while the item "DM155Q03C" worked in favour of the boys (reference) up to a total score level of approximately 15 in the characteristic curve, it showed non-uniform DIF at the B level in favour of the girls (focal) in the total score group above approximately 15 (R27=0.044>0.035). It is predicted that this is due to the fact that the partially scored items mentioned above are easier for high-achieving girl groups.

DIF Analyses Conducted by Rasch Tree Method

The results of the DIF analysis of the items in the 1st booklet according to gender, obtained with the RT method, are presented in Figure 8.

Figure 8

Rasch Tree for Booklet 1 Questions according to Gender

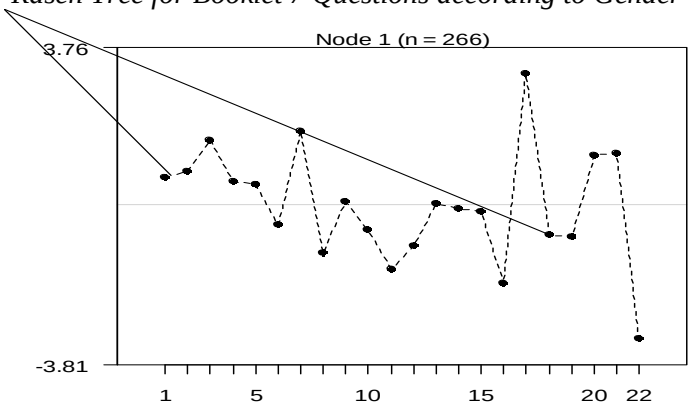


In Figure 8, the items above the horizontal line in the middle are difficult according to the subgroups in the related variable (here, it is gender), while the items on or below the horizontal line in the middle are easy according to the subgroups in the related variable (here, it is gender) (Strobl, Kopf, & Zeileis 2015). However, when Figure 8 is examined, it is inferred that there is no branching according to the subgroups, and 23 items in the 1st booklet, whose item difficulties range from -3.11 to 3.32, do not contain DIF according to gender. Appendix A more comprehensively shows what items showed DIF and what exactly the item difficulty parameters were.

The DIF analysis results of the items in the 7th Booklet obtained with the RT method according to gender are given in Figure 9.

Figure 9

Rasch Tree for Booklet 7 Questions according to Gender



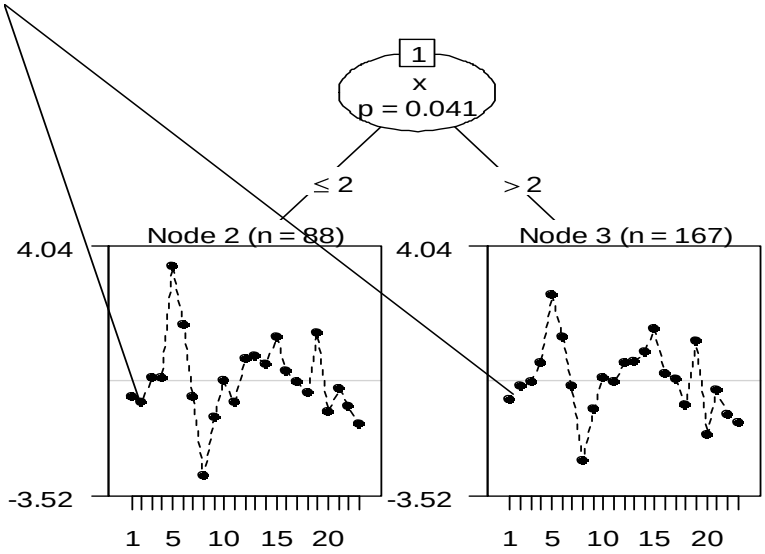
When Figure 9 is examined, it is observed that there is no branching according to the subgroups. In addition, it is determined that 22 items in the 7th booklet, whose item difficulties range from -3.81 to 3.76, do not contain any DIF according to gender. Appendix B more comprehensively shows what items showed DIF and what exactly the item difficulty parameters were.

Findings Regarding Differential Item Functioning According to Parental Education Level

Whether the PISA 2018 Mathematics subtest shows DIF according to the parental education level was analysed with the RT method. For this purpose, the items in the 1st booklet and then the 7th booklet were analysed. The DIF analysis results of the items in the 1st booklet according to the mother's education level and the father's education level obtained with the RT method are presented in Figure 10.

Figure 10

Rasch Tree for Booklet 1 Questions according to Mother's Education Level

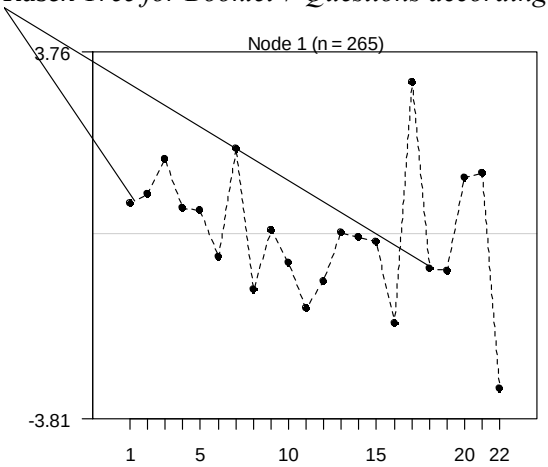


The variable "x" represents the mother's education level, "1" being a university graduate, "2" being a high school graduate, "3" being a secondary school graduate, "4" being a primary school graduate and "5" being a primary school dropout. Strobl, Kopf, and Zeileis (2015) state that item difficulty values below the zero line indicate that the items are easy, while items above the zero line indicate that the items are difficult. According to this statement, some items in PISA 2018 mathematics subtest Booklet 1 show DIF according to the mother's education level.

When Figure 10 is examined, item 11 (CM305Q01S) (ordered points in Figure 2), which is one of 23 items in the 1st booklet, whose item difficulty ranges from -3.52 to 4.04, seems easy for the students whose mother's education level is 2 or below, that is, high school graduate or university graduate, however, it seems more difficult for the students whose mother's education level is above 2, that is, middle school graduate, primary school graduate or primary school dropout. Appendix C more comprehensively shows what items showed DIF and what exactly the item difficulty parameters were. Considering that this item tests visuospatial ability, the significance of the difference between the spatial perceptions of the students whose mother's education level is high school or higher and the spatial perceptions of the students whose mother's education level is secondary school and below coincides with this situation (İrioğlu & Ertekin, 2011). The DIF analysis results of the items in the 7th booklet according to the parental education level obtained with the RT method are presented in Figure 11.

Figure 11

Rasch Tree for Booklet 7 Questions according to Mother's Education Level



In Figure 11, the items above the horizontal line in the middle are difficult according to the subgroups in the related variable, while the items on or below the horizontal line in the middle are easy according to the subgroups in the related variable. However, when Figure 11 is examined, it is seen that there is no branching according to the subgroups and 22 items in the 7th booklet, whose item difficulties range from -3.81 to 3.76, do not contain any DIF according to the parental education level. Appendix D more comprehensively shows what items showed DIF and what exactly the item difficulty parameters were.

Comparison of DIF Analyses Conducted with the MH, LR, and RT Methods According to Gender

The comparison of the DIF analyses according to gender in the PISA 2018 Mathematics subtest in the 1st and 7th booklets is presented in Table 15.

Table 15

Comparison of DIF Analyses according to Gender in Booklets 1 and 7

Booklet Number	Mantel Haenszel (MH)	MH DIF Direction	Logistic Regression (LR)	LR DIF Direction	Rasch Tree (RT)	RT DIF Direction
1	DM155Q03C	Girls	DM406Q02C	Girls	-	-
7	-	-	DM155Q03C	Girls	-	-

When Table 15 is examined, it is seen that DIF was determined in favour of the girls only for the item "DM155Q03C" in the 1st booklet using the MH method. It was also determined by Logistic Regression method that the same item in the 7th booklet (DM155Q03C) contained DIF in favour of the girls. In this respect, MH and LR DIF determination methods are compatible with each other, which is also consistent with the findings of Gök, Kelecioğlu, and Doğan (2010). In the DIF analyses conducted with LR, it was determined that the item "DM406Q02C" in the 1st booklet also contained DIF in favour of the girls. It was noticed that this item also contained DIF at C level in the findings obtained with the MH method, but it was not included in Table 16 because it was not significant. In the DIF analyses conducted with the Rasch Tree method, no DIF was determined for any of the items in the 1st and 7th booklets. This indicates that the RT method differs from the MH and LR methods. Considering the number of items determined to contain DIF, it is seen that the LR method is more sensitive than the RT method, which is in line with the findings of Liu (2017).

Discussion, Conclusion, and Suggestions

In this study, firstly, it was examined whether the items in the 1st and 7th booklets of the PISA 2018 mathematics subtest applied to the Turkish sample showed DIF according to gender. In the DIF analyses conducted with the MH, LR, and RT methods, it was concluded that the items "DM155Q03C" (MH) and "DM406Q02C" (LR) in the 1st booklet showed DIF at the C level in favour of the girls. In the 7th booklet, it was determined that the item "DM155Q03C" (LR), which is common with the 1st booklet, showed DIF at the B level. As a result, the item "DM155Q03C" showed DIF in favour of the girls in both MH and LR methods. It is noteworthy that the items showing DIF are open-ended, that is, partially scored items, regardless of methods applied, which is in line with the findings of Schwabe et al. (2014), Başman (2017), and Koğar and Koğar (2019). In addition, there are studies in the literature showing that the methods based on CTT and IRT are more compatible within themselves (Kan, Sünbül, & Ömür, 2013; Doğan & Öğretmen, 2008). It can be stated that the results obtained from this study are compatible with these studies.

It was examined whether the items in the 1st and 7th booklets of the PISA 2018 mathematics subtest applied to the Turkish sample showed DIF according to parental education level. This analysis was conducted with the RT method since the related variable had more than two categories. In these analyses, the item "CM305Q01S" in the 1st booklet was determined to be easy for the students whose mother's education level is high school graduate or university graduate, however, it is difficult for the students whose mother's education level is below high school level. Considering that the item is visuospatial, this finding coincides with the significant difference between the spatial perceptions of students whose mother's education level is high school or higher and the spatial perceptions of students whose mother's education level is lower than high school (İrioğlu & Ertekin, 2011).

When the literature is examined, it is noteworthy that while it is possible to come across many studies aiming to determine DIF, there are very few studies on determining bias regarding the evaluation of DIF together with the items. In this context, it can be suggested that examining the reasons behind DIF of the items in terms of both the technical and affective properties of the items may be beneficial in terms of increasing the quality of the items. In addition, the items showing DIF according to gender and parental education level were focused on within the scope of this research. However, there are also many different variables such as socioeconomic level and school type, which are thought to affect mathematics achievement. It may also be recommended to conduct studies that examine the underlying causes of the items showing DIF according to these variables.

Declarations

Author Contribution: Emre Kucam-Conceptualization, methodology, analysis, writing & editing, visualization. H.Deniz Gülleroğlu-Conceptualization, methodology, writing-review & editing, supervision.

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

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

References

Agresti, A. (2012). *Categorical data analysis (Vol. 792)*. John Wiley & Sons. https://doi.org/10.1002/0471249688

Alatlı, B. A., & Şenel, S. (2020). Değişen Madde Fonksiyonunun Belirlenmesinde "difR" R Paketinin Kullanımı: Ortaöğretime Geçiş Sınavı Fen Alt Testi. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, 1-37. https://doi.org/10.30964/auebfd.684727

Altıntaş, Ö., & Kutlu, Ö. (2019). Investigating Differential Item Functioning of Ankara University Examination for Foreign Students by Recursive Partitioning Analysis in the Rasch Model. *International Journal of Assessment Tools in Education, 6*(4), 602-616. https://dx.doi.org/10.21449/ijate.554212

Arslan, M. (2020). Teog Sınavının Yabancı Dil Alt Testine Ait Maddelerin Yanlılığının İncelenmesi. *Yüksek Lisans Tezi, Ankara: Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü*.

Asamoah, N. A. B. (2020). *Assessing Differential Item Functioning in the Perceived Stress Scale*. University of Arkansas. https://scholarworks.uark.edu/etd/3775

Ayan, C. (2011). PISA 2009 fen okuryazarlığı alt testinin değişen madde fonksiyonu açısından incelenmesi. *Yayınlanmamış Yüksek Lisans Tezi, Ankara: Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü*.

Başman, M. (2017). Matematik başarısında cinsiyet ve duyuşsal özelliklerin etkileşimine göre Rasch ağacı yöntemi ile değişen madde fonksiyonunun belirlenmesi. *Doktora Tezi, Ankara: Ankara Üniversitesi Eğitim Bilimleri Enstitüsü*.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items* (Vol. 4). Sage. https://doi.org/10.1177/109821409701800108

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, *1*(2), 245-276. https://doi.org/10.1207/s15327906mbr0102_10

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22(3), 265-289.* https://doi.org/10.2307/1165285

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

Cronbach, L. J. (1990). *Essentials of Psychological Testing*. Harper Collins Publishers, New York. https://doi.org/10.1002/sce.3730350432

DeVellis, R. F. (2017). *Ölçek geliştirme kuram ve uygulamalar. (T. Totan, Çev.).* Nobel Akademik Yayıncılık. https://doi.org/10.1177/109821409301400212

Doğan, N., & Öğretmen, T. (2008). Değişen madde fonksiyonunu belirlemede mantel-haenszel, ki-kare ve lojistik regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim, 33(148), 100-112.*

Ellis, B. B., & Raju, N. S. (2003). *Differential item and test functioning.* Jossey-Bass/Wiley.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage publications. https://doi.org/10.1111/insr.12011_21

Gök, B., Kelecioğlu, H., & Doğan, N. (2010). Değişen madde fonksiyonunu belirlemede Mantel–Haenszel ve Lojistik Regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim*, *35*(156).

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

İrioğlu, Z., & Ertekin, E. (2011). İlköğretim İkinci Kademe Öğrencilerinin Zihinsel Döndürme Becerilerinin Bazı Değişkenler Açısından İncelenmesi. *Journal of Educational and Instructional Studies in the World, 75.*

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

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.* https://doi.org/10.17860/efd.55452

Karami, H. R., Gramipour, M., & Minaei, A. (2021). *Application of The Rasch Tree Model In The Detection Of Differential Item Functioning* (Case Study: Recruitment Exams Of The Police Of The Islamic Republic Of Iran). https://doi.org/10.22054/jem.2021.61694.2190

Karasar, N. (2017). Bilimsel araştırma yöntemi (2. yazım, 32. Basım). *Nobel Yayın Dağıtım*.

Karip, E., & Köksal, K. (1996). Etkili eğitim sistemlerinin geliştirilmesi. *Kuram ve Uygulamada Egitim Yönetimi Dergisi*, *2*(2), 245-257.

Koğar, E. Y., & Koğar, H. (2019). Investigation of scientific literacy according to different item types: PISA 2015 Turkey sample. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, *19*(2), 695-709. https://doi.org/10.17240/aibuefd.2019.19.46660-467271

Liu, M. (2017). *Differential Item Functioning in Large-scale Mathematics Assessments: Comparing the Capabilities of the Rasch Trees Model to Traditional Approaches* (Doctoral dissertation, University of Toledo).

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge. https://doi.org/10.4324/9780203056615

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied psychological measurement*, *17*(4), 297-334. https://doi.org/10.1177/014662169301700401

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw Hill. https://doi.org/10.1177/014662169501900308

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (Vol. 161). Sage Publications. https://doi.org/10.4135/9781412993913

Ozarkan, H. B., Kucam, E., & Demir, E. (2017). Merkezi ortak sınav matematik alt testinde değişen madde fonksiyonunun görme engeli durumuna göre incelenmesi. *Current Research in Education*, *3*(1), 24-34.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied psychological measurement, 19(1), 23-37.*

Robitzsch, A.; Lüdtke, O. (2020). A review of different scaling approaches under full invariance, partial invariance, and noninvariance for cross-sectional country comparisons in large-scale assessments. *Psych. Test Assess. Model.* 2020, 62, 233–279. https://bit.ly/3ezBB05 (accessed on 12 June 2023).

Schwabe, F., McElvany, N., Trendtel, M., Gebauer, M. M., & Bos, W. (2014). Vertiefende Analysen zu migrationsbedingten Leistungsdifferenzen in Leseaufgaben. *Zeitschrift für Pädagogische Psychologie*.

Sharma, S. (1995). *Applied multivariate techniques*. John Wiley & Sons, Inc..

Spearman, C. (1905). Proof and disproof of correlation. *The American Journal of Psychology, 16*(2), 228-231.

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika, 80*(2), 289-316. https://doi.org/10.1007/s11336-013-9388-3

Şenferah, S. (2015). Seviye belirleme sınavı matematik alt testi için değişen madde fonksiyonlarının ve madde yanlılığının incelenmesi. *Yayınlanmamış Yüksek Lisans Tezi. Gazi Üniversitesi. Eğitim Bilimleri Enstitüsü. Ankara*.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5, pp. 481-498). Boston, MA: Pearson.

Tavşancıl, E. (2018). *Tutumların ölçülmesi ve SPSS ile veri analizi.* Nobel Akademik Yayıncılık.

Zhang, M. (2009). *Gender related differential item functioning in mathematics tests: A meta-analysis* (Doctoral dissertation, Washington State University).

Zieky, M. (1993). *Practical questions in the use of DIF statistics in test development.* Lawrence Erlbaum Associates, Inc.

Zumbo, B. D., & Thomas, D. R. (1997). A measure of effect size for a model-based approach for studying DIF. *Prince George, Canada: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science*.

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

APPENDICES

APPENDIX A

Item difficulties parameters for Booklet 1 (for gender)

Item	Female	Male
CM564Q02S	0.39	-0.06
CM564Q01S	0.39	0.02
CM571Q01S	0.70	0.41
CM603Q01S	0.94	0.58
DM406Q02C	4.40	1.17
DM406Q01C	2.27	0.74
CM192Q01S	0.39	0.18
CM423Q01S	-1.72	-0.69
CM496Q02S	-0.04	-0.10
CM496Q01S	0.70	0.48
CM305Q01S	0.58	-0.12
CM034Q01S	1.40	0.56
DM462Q01C	1.61	0.54
CM442Q02S	1.50	0.64
CM803Q01S	2.35	0.65
CM411Q02S	1.12	0.35
CM411Q01S	0.94	0.39
CM155Q04S	-0.01	-0.08
DM155Q03C	2.35	0.66
CM155Q01S	-0.52	-0.41
DM155Q02C	0.58	0.47
CM474Q01S	-0.33	0.003
CM033Q01S	-0.59	-0.21

APPENDIX B

Item difficulty parameters for Booklet 7 (for gender)

Item	Female	Male
CM034Q01S	1.01	0.85
DM462Q01C	1.01	1.07
CM803Q01S	1.84	1.45
CM411Q02S	0.93	0.79
CM411Q01S	0.93	0.69
CM155Q04S	0.11	-0.005
DM155Q03C	1.72	1.87
CM155Q01S	-0.41	-0.51
DM155Q02C	0.60	0.40
CM474Q01S	0.15	-0.21
CM033Q01S	-0.69	-0.86
CM447Q01S	-0.34	-0.32
CM273Q01S	0.32	0.56
CM408Q01S	0.49	0.25
CM420Q01S	0.32	0.28
CM446Q01S	-0.99	-1.08
DM446Q02C	2.97	2.77
CM559Q01S	-0.04	-0.21
DM828Q02C	-0.11	-0.21
CM828Q03S	1.41	1.27
CM464Q01S	1.56	1.23
CM800Q01S	-2.08	-2.14

APPENDIX C

Item difficulty parameters for Booklet 1 (for mother's education level)

Item	Primary School Dropout	Primary School	Middle School	High School	Undergraduate and Above
CM564Q02S	0.003	1.24	0.70	0.53	0.24
CM564Q01S	0.006	1.30	0.73	-0.19	0.54
CM571Q01S	0.006	1.25	0.81	1.09	0.65
CM603Q01S	0.019	1.45	0.81	1.09	0.87
DM406Q02C	0.36	2.32	0.97	3.05	24.88
DM406Q01C	0.02	1.69	0.85	3.05	1.97
CM192Q01S	0.01	1.31	0.70	0.66	0.04
CM423Q01S	-0.02	1.31	0.44	-2.00	-1.79
CM496Q02S	-0.001	1.62	0.63	-0.19	-0.15
CM496Q01S	0.008	1.47	0.72	1.24	0.44
CM305Q01S	0.004	1.41	0.76	0.41	0.13
CM034Q01S	0.01	1.47	0.84	2.02	1.10
DM462Q01C	0.01	1.47	0.79	1.59	1.35
CM442Q02S	0.01	1.50	0.90	1.41	1.10
CM803Q01S	0.02	1.68	0.87	1.79	2.16
CM411Q02S	0.007	1.38	0.74	1.41	1.10
CM411Q01S	0.01	1.41	0.72	1.24	0.54
CM155Q04S	0	1.30	0.72	0.41	0.54
DM155Q03C	0.02	1.52	0.85	2.28	1.97
CM155Q01S	-0.007	1.45	0.53	-0.19	0.04
DM155Q02C	0.007	1.34	0.65	0.41	0.65
CM474Q01S	-0.005	1.29	0.65	0.28	0.04
CM033Q01S	0.005	1.35	0.62	-0.44	-0.35

APPENDIX D

Item difficulty parameters for Booklet 7 (for mother's education level)

Item	Primary School Dropout	Primary School	Middle School	High School	Undergraduate and Above
CM034Q01S	0.02	0.27	0.64	1.46	0.81
DM462Q01C	0.02	0.24	0.77	1.21	0.81
CM803Q01S	0.04	0.33	0.97	1.46	2.35
CM411Q02S	0.02	0.29	0.94	0.79	0.54
CM411Q01S	0.03	0.25	0.52	0.89	0.46
CM155Q04S	0.02	0.22	0.26	-0.36	-0.41
DM155Q03C	0.07	0.38	0.94	2.28	1.98
CM155Q01S	-0.02	0.18	0.19	0.15	-1.43
DM155Q02C	0.01	0.26	0.53	0.69	0.30
CM474Q01S	-0.01	0.26	0.31	-0.19	-0.09
CM033Q01S	-0.003	0.04	0.01	-0.82	-1.11
CM447Q01S	0.008	0.17	0.05	-0.82	-0.41
CM273Q01S	0.01	0.25	0.43	0.99	0.06
CM408Q01S	-0.008	0.26	0.47	0.41	0.30
CM420Q01S	0.004	0.26	0.36	-0.10	0.38
CM446Q01S	-0.03	0.10	0.22	-1.01	-1.43
DM446Q02C	0.13	0.82	2.02	2.51	3.27
CM559Q01S	-0.003	0.23	0.10	-0.36	-0.49
DM828Q02C	-0.01	0.23	0.15	-0.02	-0.58
CM828Q03S	0.04	0.36	0.87	1.33	1.10
CM464Q01S	0.06	0.25	0.89	1.74	1.10
CM800Q01S	-0.09	-0.27	-0.10	-3.12	-2.32

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(2); 154-170

Investigation of Measurement Invariance of Turkish Subtest on ABIDE 2016 in Relation to Characteristics of Teachers: Sub-sampling Method

Süleyman ÜLKÜ* Burcu ATAR **

Abstract

The Ministry of National Education carried out the ABIDE (Monitoring and Evaluation of Academic Skills) in 2016 in order to test the knowledge and skills of 8th-grade students. Since the ABIDE 2016 study was implemented for the first time in our country, it is very important to prove measurement invariance for the validity of the results. Within the scope of this research, the measurement invariance of the success of the students in the Turkish test according to the education level and professional experience of the teachers was examined. In the research, data were obtained from the Ministry of National Education, Directorate-General of Measurement, Evaluation, and Examination Services. Responses of students to the multiple-choice items in the ABIDE 2016 Turkish test and teacher questionnaire data were used in the study. All the data were used in the investigation of measurement invariance according to professional experience. Investigation of measurement invariance according to education level was carried out both by using and not using the method of sub-sampling. Factor 10 and Mplus 7 programs were used in the analysis of the data. At the end of the study, the Turkish achievement model provided all levels of measurement invariance among the student groups formed according to the professional experience and education level of the teachers.

Keywords: Measurement invariance, ABIDE 2016, sub-sampling method

Introduction

Education has become one of the globally significant indicators for the attainment of development-focused strategic objectives of countries in recent years. It is possible to forecast the future of a given country based on the effectiveness of educational reforms and the actual student achievement rates. Therefore, standard measurement and evaluation systems are required to evaluate the quality of learning experiences and to provide stakeholders with feedback according to these evaluations.

Exams on an international scale such as PISA (Program for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study) are carried out by international organizations in order to provide feedback for the development and improvement of countries' education systems. Similarly, in Türkiye, it was aimed to develop ABIDE (Monitoring and Evaluation of Academic Skills) which is a standard measurement and evaluation tool by the Ministry of National Education (MEB in Turkish). The overall aim of ABIDE is to determine to what extent 8th-grade students have high-level mental skills in Turkish, mathematics, science, and social studies and to reveal the student, family, teacher, and school characteristics that affect the success of the students (MEB, 2017). ABIDE implemented in two-year periods was first implemented at the 8th grade level in 2016, and the second application was made at the 4th and 8th grade levels in 2018. ABIDE, which was planned to be held in 2020, was carried out in 2021 due to the covid-19 pandemic. However, final reports haven't been published by the MEB except for 2016.

The present study is a part of Master's Thesis conducted under the supervision of Prof. Dr. Burcu ATAR and prepared by Süleyman ÜLKÜ

* PhD. Student, Hacettepe University, Faculty of Education, Ankara-Türkiye, suleymanulku@ hacettepe.edu.tr, ORCID ID: 0000-0003-1965-0671
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Türkiye, burcua@hacettepe.edu.tr, ORCID ID: 0000-0003-3527-686X

To cite this article:
Ülkü, S., & Atar, B. (2023). Investigation of measurement invariance of Turkish subtest on ABIDE 2016 in relation to characteristics of teachers: Sub-sampling method. *Journal of Measurement and Evaluation in Education and Psychology, 14*(2), 154-170. https://doi.org/10.21031/epod.1084985

Received: 9.03.2022
Accepted: 5.06.2023

In the ABIDE 2017 report, comparisons of students' achievements are given according to teachers' educational backgrounds and professional experience. Of course, there are many factors such as school, family, teacher, and environment that underlie the success differences of students. The characteristics one must possess have started to differ gradually in recent years; education plays the most vital role in the indoctrination of such characteristics. Thus, teachers assume considerable responsibility in this process. ABIDE results allow for the interpretation of student achievement rates in terms of the characteristics of teachers enabling the institutions concerned and stakeholders in education to adopt measures and take decisions regarding the improvement of the education system (MEB, 2017). In this case, the report outcomes are expected to be valid and reliable among different groups.

In educational studies, comparisons between groups are frequently made to identify the qualities stemming from the individual, school, teacher, etc. affecting student achievement. However, before commenting that "differences between groups stem from variables originating from students, teachers or schools" based on such comparisons, it is necessary to examine whether these differences are caused by the measurement tool or not. In order to make comparisons according to groups using a measurement tool, measurement invariance must first be ensured in those groups, if measurement invariance is not ensured, the results of the comparison will lose their significance (Byrne Barbara, 2004). Measurement invariance denotes testing whether the measurement tool shows a similar structure among different groups to provide evidence for the validity of measurement tools (Van de Schoot, Lugtig & Hox, 2012). In other words, measurement invariance is to obtain similar results by applying the same scale to different groups that are similar in terms of measured characteristics (Cheug and Rensvold, 1998).

The ABIDE 2017 report frequently compares the achievement levels of students from different groups in terms of teacher qualities.

Individuals in different groups yet equivalent in terms of the attributes assessed must obtain the same score for the accuracy of the comparisons (Schmith & Kuljanin, 2008). This means that some evidence must be provided regarding the measurement of similar structures among groups assessed in terms of the attribute assessed. In other words, the measurement invariance of the tests in the groups determined must be established in order to make comparisons among varying groups using the observed variable scores (Vandenberg & Lance, 2000). Explaining the differences in the results obtained from a measurement tool solely based on individual properties in research studies, making comparisons among groups in terms of the variables to be assessed might not always be accurate. This is because the difference among individuals may also result from the measurement tool (Cheug ve Rensvold, 1998).

Obtaining information about the equivalence of the construct validity of the tests given within the scope of the ABIDE evaluation in 2016 among the student groups formed in terms of the education level and professional experience of teachers would contribute to provide the validity of the measurement results.

Measurement Invariance

The concept of validity is defined as supporting the outcomes based on the scores obtained from the measurement tool and the interpretations made with reference to these outcomes with experimental and theoretical evidence (AERA, APA & NCME, 2014). If one is to compare certain structures among various groups with a measurement tool, the theoretical structure must be the same and be interpreted in the same way by the sub-groups. Otherwise, test bias occurs (Kline, 2011). Based on this point, measurement invariance studies are conducted to identify whether a sub-group has an advantage over others or whether the measurement tools show the same structure as the sub-groups. Making comparisons among groups not displaying the same structure causes the measurement tool not to function. This leads to misinterpretations, which gives rise to misjudgments.

There are various measurement invariance analysis methods in the existing body of literature. The first group of methods examines differences in item and test functions based on the Item Response Theory, the second group consists of methods based on Latent Class Analysis and the final group includes the methods of multi-group confirmatory factor analysis (MGCFA) based on structural equation modeling and the invariance of mean and covariance structures (Kankaras et. al, 2011). MGCFA testing the

equivalence of covariance structures is frequently used in measurement invariance studies (Meredith, 1993).

Structural Equation Modeling (SEM) is a powerful and advanced statistical tool providing the researcher with a comprehensive method to assess and modify the model created through theoretical inferences (Dragan & Topolsek, 2014). According to Tabachnick & Fidel (2013), structural equation modeling (SEM) denotes a collection of statistical techniques allowing for the examination of the relationships between one or more independent variables, either continuous or discrete, and one or more dependent variables, either continuous or discrete. SEM analyses signify an expanded combination of factor analysis, multiple regression, and covariance analysis (Hoyle, 2012).

According to Kline (2011), SEM involves six steps: model specification, model identification, evaluation of model fit, measurement of fitness statistics, re-specification of the model where necessary, and reporting of the results. A frequently employed method in SEM analyses, MGCFA is a technique requiring the simultaneous application of CFA on two or more groups. This analysis tests whether the model created by the researcher for the measurement tool is the same for the sub-groups of the sample (Tabachnick & Fidell, 2013).

According to Vanderberg & Lance (2000), measurement invariance is handled by multi-group confirmatory factor analysis as follows: Let us assume the score obtained by the individual i within the group k for the assessed variable of j is X_{ijk}. In this case, the factor model for X_{ijk} is as follows.

$$X_{ijk} = \tau_{jk} + \Upsilon_{jk} W_{jk} + u_{jk} \quad (1)$$

τ_{jk} represents the coefficient factor between the observed and latent structure, Υ_{jk} signifies the factor loadings matrix of rx1 considering that r represents the number of items, W_{jk} shows the common factor loadings vector matrix for i individuals in the rx1 pattern, and u_{jk} shows the error vector of independently observed variables. Furthermore, j signifies the assessed variable, k the group, and i the individual. In this case, X_{ijk} is signified as the score of the individual i within the group k for the variable j. Additionally, it is assumed that the measurement errors are within themselves and the correlation with the common factor loadings is "0". Based on the assumption E(W_{jk}, u_{jk})=0, the covariance equation is as follows:

$$cov(X_{ijk}) = \Sigma_k = \Lambda_k \Phi_k \Lambda'_k + \theta_k \quad (2)$$

The expression Λ_k signifies the matrix of the pxr pattern consisting of Υ_{jk} while Φ_k indicates the variances and covariances in Υ_{jk}. θ_k signifies the diagonal matrix of measurement errors. Similarly, the average vector of X_{ik} can be expressed as follows:

$$E(X_{ik}) = \mu_k = \tau_k + \Lambda_k K_k \quad (3)$$

Based on the equations given above, the question of whether the parameters of $[\tau_k, \Lambda_k, \theta_k]$ are equal in k groups (Vandenberg & Lance, 2000, p. 10; Jöreskog & Sörborm, 1993; Millsap & Olivera-Aguilar, 2012, p. 381).

Measurement invariance is exhibited with multi-group confirmatory factor analysis through the testing of the four nested hierarchical levels or the hypothesis. These four levels are called configural, metric, scalar, and strict invariance, respectively (Meredith, 1993).

Configural Invariance

According to Wu, Li & Zumbo (2007), it denotes the initial level of measurement invariance analysis and constitutes a prerequisite for continuing with other levels. This level involves testing whether the model (factor structure) established based on the research hypothesis is the same among the groups. In other words, it means that the Λ_k the matrix in equation 3 has the same fixed and free factor loads for all groups (Widamann ve Reise, 1997).

$$\Lambda_k{}^{(1)} = \Lambda_k{}^{(2)} \quad (4)$$

If configural invariance is not ensured, in other words, if factor structure is the same among groups, the factor configuration among the groups does not differ and the items measure the same structure among different groups. If configural invariance is not ensured, there is no need to conduct the analyses to identify the differences among groups or test the remaining levels of measurement invariance as the measured configurations differ from one group to another (Vandenberg & Lance, 2000).

Metric Invariance

This invariance level is also called weak invariance (Meredith, 1993). In addition to configural invariance, metric invariance is based on the condition whereby the factor loadings of the items concerned must be equal among the groups.

$$\Upsilon_{jk}{}^{(1)} = \Upsilon_{jk}{}^{(2)} \quad (5)$$

Observed variables are connected to latent variables through factor loadings. Therefore, even a minute change in the latent variable affects the observed variable (Bollen, 1989). For this reason, factor loadings must be equivalent if one wants to measure the same configuration among different groups.

The fitness of the metric model is compared to that of the structural model using the difference between the chi-square tests or fit indices to identify whether the condition of metric invariance is fulfilled. If there are no significant differences in model fit or if the fit indices are within the desired range, one might argue that the factor loadings in the sub-groups subject to the comparison do not change. In this respect, this means that all individuals in the sub-groups interpreted the items similarly. The factor variances and covariances may be compared among the groups with the fulfillment of the condition of metric invariance. However, it is not possible to indicate exactly the source of the average difference among the groups.

If the condition of metric invariance is not fulfilled, one might indicate that factor loadings vary among the groups and people made different interpretations of the items concerned (Bialosiewicz, Murphy & Berry, 2013). The lack of metric invariance may signify that the meanings of the items are not the same for all groups, leading to item bias. Partial measurement invariance studies may be conducted if this is the case. If the condition of metric invariance is satisfied, one might move to the next level.

Scalar Invariance

In addition to the conditions required by metric invariance, scalar invariance is based on the equivalence of item threshold values for the sub-groups.

$$\tau_{jk}{}^{(1)} = \tau_{jk}{}^{(2)} \quad (6)$$

To assess scalar invariance, the fitness of the model established is compared with that of the metric model by using the difference between the chi-square difference tests or fit indices. If there are no significant differences in model fit or if the fit indices are within the desired range, one might argue that the factor threshold values do not vary among the sub-groups (Vandenberg & Lance, 2000).

The fulfillment of the condition of scalar invariance means that the averages of factors and observed variables may be compared. In other words, one might argue that there is no bias favoring any sub-group(s) and that the average differences in observed variables source from those in the latent variable (Başusta & Gelbal, 2015). Strict invariance is the next level following the fulfillment of the condition of scalar invariance.

Strict Invariance

At this level, the condition taken into consideration in addition to the conditions of scalar invariance is the equivalence of item error variances among the sub-groups.

$$\theta_k{}^{(1)} = \theta_k{}^{(2)} \quad (7)$$

To assess strict invariance, the fitness of the model established is compared with that of the model established at the level of scalar invariance by using the difference between the chi-square difference tests or fit indices. If there are no significant differences in model fit or if the fit indices are within the desired range, one might argue that the item error variances do not vary among the sub-groups (Bollen, 1989).

If the condition of strict invariance is fulfilled, one can compare observed variances and covariances in addition to the averages of latent and observed variables. However, one must also keep in mind that strict invariance is a quite limited model, and its conditions are rarely fulfilled in practice. This is because as the variance resulting from the latent variable increases, so do the residual variances of the items (Bialosiewicz, Murphy & Berry, 2013).

Purpose of Study

The present study examines whether measurement invariance is established among student groups created based on the education level and professional experience of teachers for the ABIDE 2016 Turkish test. In this respect, the following research questions were identified: (a) "Is measurement invariance established among student groups formed on the basis of the professional experience of teachers in the ABIDE 2016 Turkish test?", (b) "Is measurement invariance established among student groups formed on the basis of the education levels of teachers using the sub-sampling method in the ABIDE 2016 Turkish test?", and (c) "Is measurement invariance established among student groups formed on the basis of the education level of teachers without using the sub-sampling method in the ABIDE 2016 Turkish test?"

Method

The study is a descriptive research in order to illuminate a given situation and to determine the level of validity of the study, which aims to examine the measurement invariance of students' success in Turkish tests according to teachers' education level and professional experience. Studies that aim to reveal a situation without intervening are in the type of descriptive research (Fraenkel & Wallen, 2006; Karasar, 2011). Descriptive models are research models that aim to reveal the states of variables and to reveal the change between variables (Gall et. al, 1999).

Research Population and Sample

The population of the ABIDE assessment consists of 8th-grade students from Türkiye. The ABIDE 2016 assessment was applied in 16,118 schools and 48,091 classes. Conducted in all 81 provinces, the study took into consideration around 400 students from each province. The number of students to be included in the samples in metropolises was increased proportionately to the population to better reflect the overall population. Therefore, the assessment was given to a total of around 38,000 students. Furthermore, students were also classified into stratas through stratified sampling in order for the samples selected to better represent the province concerned (MEB, 2017).

For research purposes, the data on 7952 students using Form A of the Turkish test were obtained from the Directorate-General of Measurement, Evaluation, and Examination Services. 86 students not providing answers to any questions in Form A of the Turkish test were excluded from the study. As a result of the examination of missing values, data concerning a total of 365 students were excluded from the study. As a result, data from 7501 students were used in the analyses. Table 1 shows the information

about the student groups created based on the professional experience and education levels of teachers included in the research sample.

Table 1

Student Frequencies in Terms of the Professional Experience and Education Status of Teachers

Professional Experience of Teachers	N of Students	Education Status of Teachers	N of Students
0-5 (short)	2.183	Associate Degree	240
6-15 (medium)	3.790	Bachelor's Degree	6.997
16+ (extensive)	1.528	Master's Degree	264
Total	7.501	Total	7.501

Upon forming groups based on the educational background of teachers within the scope of the study, great differences seemed to emerge among the student groups. Within this scope, the study attempted to obtain information on the question of how the results vary by examining measurement invariance in terms of the educational backgrounds of teachers both using and not using the method of sub-sampling.

Sub-sampling method

Imbalanced sample sizes in groups may affect the outcome of measurement invariance studies. The difference between the observed model and the estimated model may be disregarded due to the relatively lower weight within the smaller group. Therefore, in case the sample sizes of the groups examined differ greatly, the outcomes of invariance studies may be misleading (Yoon & Lai, 2018).

Chen (2007) found that the power of detecting noninvariance led to a substantial drop when sample sizes in two groups were quite different. Although both of these studies noted potential problems of unbalanced sample sizes in testing factorial invariance, neither included a systematic investigation of unequal sample size conditions that would influence power in detecting violations of invariance (Yoon & Lai, 2018).

Yoon and Lai (2018) suggested that researchers use many random samples from the larger group in testing measurement invariance and report the summary of the results using many random samples. For example, the sub-samples of the larger group may be selected randomly 100 times and each sub-sample selected randomly and the smaller group may be used collectively for measurement invariance analysis. Thus, measurement invariance analysis is conducted 100 times for the different sub-samples of the larger group while using the same sample for the smaller group each time. The fit indices are recorded for each different instance and the average of the fit indices recorded for all 100 instances is calculated. If both the average values and the relevant percentage values for the fit indices are within the range of good fit, the measurement invariance model is supported (Yoon & Lai, 2018).

The R package software was used for creating sub-samples based on the educational background of teachers. The group consisting of teachers with associate degrees, whose size is the smallest (see Table 1) was taken into consideration for the Turkish test. The software output obtained was a file to be used for measurement invariance analysis on Mplus.

Table 2 shows the item averages of student groups created based on the professional experience and education status of teachers for the Turkish test in order to demonstrate the similarity of the averages for the sub-samples acquired using the sub-sampling method with sample averages. Furthermore, the table also features the averages of the sub-samples obtained with the sub-sampling method based on educational background.

Table 2

Item Averages Based on the Educational Background and Professional Experience of Teachers

	Professional Experience			Education Status			Education Status (Sub-sampling Methods)		
	Short	Medium	Extensive	Associate Degree	Bachelor's Degree	Master's Degree	Associate Degree	Bachelor's Degree	Master's Degree
M2	0,51	0,54	0,59	0,55	0,54	0,55	0,55	0,56	0,54
M3	0,31	0,34	0,38	0,38	0,34	0,31	0,38	0,32	0,31
M4	0,70	0,73	0,73	0,77	0,72	0,76	0,77	0,70	0,76
M5	0,47	0,51	0,54	0,61	0,50	0,45	0,61	0,53	0,46
M6	0,70	0,77	0,79	0,84	0,75	0,74	0,84	0,77	0,75
M7	0,65	0,73	0,74	0,79	0,71	0,67	0,79	0,73	0,67
M9	0,46	0,55	0,60	0,61	0,53	0,51	0,61	0,52	0,51

An assessment of Table 2 might lead to the conclusion that the items included in the Turkish test are generally of average difficulty. Additionally, the item averages based on educational background and the averages of the data originating from the educational background sub-samples seem to be close. In other words, the averages of the sub-samples were found to be similar to the average value for the original sample.

Data Collection Process

The open-ended and multiple-choice items included in the ABIDE 2016 assessment were developed by item writers, subject matter experts, measurement and evaluation specialists, and language experts. Then, a pilot scheme was conducted with around 5000 students. The tests were finalized using the item and test statistics at the end of the pilot scheme. Between April and May 2016, the main assessment scheme was put into action in 81 provinces (MEB, 2017). The research data were obtained from the Directorate-General of Measurement, Evaluation, and Examination Services within the MEB.

Data Collection Tools

The study was conducted on the basis of existing data containing the answers given by students to the multiple-choice questions of the Turkish test in the ABIDE 2016 evaluation as well as of teacher questionnaire data. No additional data collection tools were employed besides the ones indicated here. Table 3 shows the number of items per booklet for the ABIDE 2016 assessment.

Table 3

ABIDE 2016 Booklet Types and No. of Items

A Booklet	B Booklet	C Booklet
9 + 9 = 18 items	9 + 9 = 18 items	9 + 9 = 18 items
A1: 18+2 pilot=20 items	B1: 18+2 pilot=20 items	C1: 18+2 pilot=20 items
A2: 18+2 pilot=20 items	B2: 18+2 pilot=20 items	C2: 18+2 pilot=20 items
A3: 18+2 pilot=20 items	B3: 18+2 pilot=20 items	C3: 18+2 pilot=20 items
A4: 18+2 pilot=20 items	B4: 18+2 pilot=20 items	C4: 18+2 pilot=20 items

Source: ABIDE 2016 Report

The present study focuses on the items included in the Turkish test within Booklet A of the ABIDE 2016 assessment. Booklet A consists of nine multiple-choice and nine open-ended questions. The answers given for the open-ended items were scored as incorrect (0), partially correct (1), and correct (2). As for

the multiple-choice items, the items were scored as either correct (1) or incorrect (0). The nine multiple-choice items were included within the scope of the study.

Data Analysis

The research data were analyzed in three stages. The first stage involved examining the missing values, outliers, and the number of multicollinearity assumptions. The second stage concerned the establishment of the achievement model for the subject of Turkish and the attempts to verify the said model. As for the final stage, it was about testing the measurement invariance of the models established on the basis of the educational backgrounds and professional experience levels of teachers using the MGCFA method.

Examining Assumptions

Certain assumptions and requirements for the data obtained from the sample must be tested to minimize the problems that may arise prior to the SEM analyses. These can be listed as missing values, outliers, normality, and multicollinearity (Çokluk, Şekercioğlu & Büyüköztürk, 2010).

Missing values

The initial step before continuing with the analyses involved the examination of missing values. There are different approaches for dealing with missing values. The missing data must be completely random for these approaches to be used. 365 students were excluded from the present study because the data were categorical, the sample size was large, and the missing values accounted for less than 5% of the data and were distributed randomly. The missing values within the data used for the study ranged between 0.2% and 1.6%.

Outliers and Normality

The outliers and the assumption of normality were not examined as the data employed in the present study were categorical.

Multicollinearity

For this assumption, the relationships among the items in each factor must be analyzed. A correlation value exceeding 0.90 among the items gives rise to the issue of multicollinearity. A high correlation signifies that the items assess similar properties (Tabachnick & Fidell, 2013). Therefore, the question of whether the correlations were below 0.90 within the tetrachoric correlation matrix was examined. Table 4 shows the tetrachoric correlation matrices.

Table 4
Tetrachoric Correlation Matrix

	T0005	T0006	T0009	T0012	T0013	T0016	T0020
T0005	1						
T0006	0,193	1					
T0009	0,192	0,195	1				
T0012	0,173	0,214	0,192	1			
T0013	0,299	0,259	0,313	0,217	1		
T0016	0,236	0,214	0,242	0,167	0,356	1	
T0020	0,317	0,351	0,305	0,271	0,388	0,311	1

Based on the information given in Table 4, there is no multicollinearity among the items as all correlation values among them are below 0.90. Additionally, the tolerance values and variance inflation factors were examined in consideration of multicollinearity. The assumption is accepted if the tolerance value

is greater than 0.1 and the variance inflation factor value is lower than 10 (Tabachnick & Fidell, 2013). Table 5 features the values obtained at the end of the analyses.

Table 5
Tolerance and Variance Inflation Values

Items	VIF	Tolerance
T20160005	1,082	0,925
T20160006	1,079	0,927
T20160009	1,077	0,929
T20160013	1,059	0,945
T20160016	1,133	0,883
T20160017	1,092	0,916
T20160020	1,170	0,855

Table 5 proves that all tolerance values are greater than 0.1 and the variance inflation factor values are lower than 10, indicating the absence of multicollinearity.

All the assumptions were examined and the missing values were excluded from the study. Thus, the dataset was rendered suitable for MGCFA. The stage following these analyses involved the specification of the model. The dataset was subjected to EFA prior to the establishment of the model. Then, the model established was confirmed using CFA and modeled using a path diagram.

Exploratory Factor Analysis

EFA was calculated on the basis of 9 multiple-choice items covered within the scope of the study. It was conducted on the Factor10 software based on the tetrachoric correlation matrices since the data were categorical. The KMO value was calculated as 0.747>0.60 while Bartlett's Test of Sphericity revealed the value of $p<0.001$ for the Turkish test used for the study. In this regard, one might argue that the dataset is suitable for EFA.

The EFA results revealed that the items are collected under a single factor, which is an expected outcome according to the existing body of literature on achievement tests. However, the factor loadings for items no. T20160017 and T20160001 were calculated to be 0.109 and 0.257, respectively, leading to their exclusion from the study. The explained variance rate was 36.76% after the exclusion of the two items. Seven items in the Turkish test were collected under a single factor named "Achievement in Turkish". Table 6 shows the factor loadings for the items.

Table 6
Item Factor Loadings for the Tests

Items	Factor Loadings
T20160020	0,666
T20160013	0,622
T20160016	0,502
T20160009	0,467
T20160006	0,461
T20160005	0,458
T20160012	0,387

Table 6 shows the factor loadings for the items in the test range between 0.387 and 0.666.

Multi-Group Confirmatory Factor Analysis (MGCFA)

MGCFA provides the researcher with information on both the structural validity of the items and the invariance of this validity among groups (Gregorich, 2006). Therefore, MGCFA was used to examine whether the Achievement in Turkish model satisfies the condition of measurement invariance in terms of the professional experience and educational background of teachers within the scope of the study. The stages of the MGCFA were tested using the Mplus 7 analysis software, using the estimation method of WLSMV. Furthermore, the analyses were conducted based on the tetrachoric correlation matrix generated according to the data. The model created at each level was assessed based on the fit indices of χ^2, RMSEA, CFI, and TLI. In MGCFA, one variable is fixed, and other variable values are left to change. This variable is also called the reference variable (Jöreskog & Sörbom, 1993). Within this context, item no. T20160020 was set to be the reference variable in the study.

While identifying groups based on professional experience within the scope of the study, the amounts of time teachers spent in the profession were categorized as 0-5, 6-15, and 16+ years. Then, durations between 0 and 5 years were identified as "short experience" while those ranging between 6 and 15 years were called "medium-level experience" and periods exceeding 16 years were categorized as "extensive experience". In terms of education status, the groups identified were "Associate Degree", "Bachelor's Degree", and "Master's Degree".

Four hierarchical models or hypotheses are tested in measurement invariance through MGCFA. In each level, the differences between chi-squares and fit indices, which are the prerequisites for advancing into the next step, were examined. Table 7 shows the goodness-of-fit and acceptable fit levels of the fit indices. The difference between the models as far as the fit indices of CFI and TLI are concerned must be between -0.01 and 0.01. The studies by Cheung & Rensvold (2002) and Vandenberg & Lance (2000) indicated that the chi-square difference must be taken into consideration for measurement invariance. In models classified as such, it was also asserted that the use of the changes in the χ^2/Sd value and fit indices would produce more accurate and reliable results.

Results

This section discusses the findings regarding measurement invariance among the student groups formed based on the professional experience levels (short, medium-level, extensive) as well as on education levels (Associate Degree, Bachelor's Degree, Master's Degree) of their teachers both using and not using sub-sampling. The stages of measurement invariance examined through MGCFA were implemented in pairwise groups and the fit indices and the differences between these indices were reported in each invariance level. The order of these levels, as indicated previously, were as follows: configural, metric, scalar, and strict invariance. Table 7 features the fit indices obtained from the invariance tests regarding the model displaying the achievement in Turkish among the student groups formed based on the professional experience of teachers.

Table 7

Fit Indices for the Model Indicating the Success in Turkish Among Student Groups Formed Based on the Professional Experience of Teachers

	Levels of Invariance	χ^2	Sd	RMSEA	CFI	TLI	$\Delta\chi^2$	ΔSd	ΔCFI	ΔTLI
Short-Medium	Configural	67,238	28	0,022	0,987	0,981	-	-	-	-
	Metric	64,682	34	0,017	0,990	0,990	4,819 p=0,567	6	0,003	0,009
	Scalar	80,534	40	0,018	0,987	0,986	16,574 p=0,011	6	0,003	0,004
	Strict	69,843	33	0,019	0,988	0,985	12,828 p=0,076	7	0,001	0,001

Short-Extensive	Configural	56,844	28	0,024	0,984	0,976	-	-	-	-
	Metric	65,462	34	0,022	0,983	0,979	11,063 p=0,086	6	0,001	0,003
	Scalar	74,949	40	0,022	0,981	0,980	10,022 p=0,1237	7	0,002	0,001
	Strict	56,038	33	0,019	0,987	0,984	18,030 p=0,011	7	0,006	0,004
Medium-Extensive	Configural	48,734	28	0,017	0,993	0,989	-	-	-	-
	Metric	54,177	34	0,015	0,993	0,991	8,147 p=0,227	6	0,000	0,002
	Scalar	60,779	40	0,014	0,993	0,992	7,015 p=0,319	7	0,000	0,001
	Strict	53,107	33	0,015	0,993	0,991	9,170 p=0,240	7	0,000	0,001

According to Table 7, the values for short and medium-level experience in the structural equation model were calculated as RMSEA=0.022, CFI=0.987, and TLI=0.981. In the metric invariance model, index values were found to be RMSEA=0.017, CFI=0.990, and TLI=0.990, the chi-square difference (p=0.567) was insignificant, and the difference between ΔCFI=0.003 and ΔTLI=0.009 was within the desired range (-0.01 - +0.01). In the scalar invariance model, while the indices were calculated as RMSEA=0.018, CFI=0.987, and TLI=0.986 and the chi-square difference (p=0.011) was significant, the difference between ΔCFI=0.003 and ΔTLI=0.004 was within the desired range (-0.01 - +0.01). In the strict invariance model, index values were found to be RMSEA=0.019, CFI=0.988, and TLI=0.985, the chi-square difference (p=0.076) was insignificant, and the difference between ΔCFI=0.001 and ΔTLI=0.001 was within the desired range (-0.01 - +0.01).

The indices were found to be RMSEA=0.024, CFI=0.984, and TLI=0.976 in the structural equation model for short and extensive experience levels. In the metric invariance model, index values were found to be RMSEA=0.022, CFI=0.983, and TLI=0.979, the chi-square difference (p=0.086) was insignificant, and the difference between ΔCFI=0.001 and ΔTLI=0.003 was within the desired range (-0.01 - +0.01). In the scalar invariance model, index values were found to be RMSEA=0.022, CFI=0.981, and TLI=0.980, the chi-square difference (p=0.123) was insignificant, and the difference between ΔCFI=0.002 and ΔTLI=0.001 was within the desired range (-0.01 - +0.01). In the strict invariance mode, while the indices were calculated as RMSEA=0.019, CFI=0.987, and TLI=0.984 and the chi-square difference (p=0.011) was significant, the difference between ΔCFI=0.006 and ΔTLI=0.004 was within the desired range (-0.01 - +0.01).

The indices were found to be RMSEA=0.017, CFI=0.993, and TLI=0.989 in the structural equation model for the instances of medium-level and extensive experience. In the metric invariance model, index values were found to be RMSEA=0.015, CFI=0.993, and TLI=0.991, the chi-square difference (p=0.227) was insignificant, and the difference between ΔCFI=0.000 and ΔTLI=0.002 was within the desired range (-0.01 -+0.01). In the scalar invariance model, index values were found to be RMSEA=0.014, CFI=0.993, and TLI=0.992, the chi-square difference (p=0.319) was insignificant, and the difference between ΔCFI=0.000 and ΔTLI=0.001 was within the desired range (-0.01 - +0.01). In the strict invariance model, index values were found to be RMSEA=0.015, CFI=0.993, and TLI=0.991, the chi-square difference (p=0.240) was insignificant, and the difference between ΔCFI=0.000 and ΔTLI=0.001 was within the desired range (-0.01 - +0.01).

The RMSEA, CFI, and TLI values indicate that all models display goodness-of-fit while the ΔCFI and ΔTLI values display the necessary conditions for the advancement to the next model. Therefore, the Achievement in the Turkish model among the student groups formed based on the professional experience levels of teachers (i.e., short, medium-level, extensive) fulfilled all the levels of measurement invariance. Table 8 shows the fit indices obtained from the invariance tests regarding the model displaying the achievement in Turkish among the student groups formed based on the educational background of teachers without using the method of sub-sampling.

Table 8

Fit Indices for the Models Indicating the Achievement in Turkish Among Student Groups Formed Based on the Educational Background of Teachers

	Levels of Invariance	χ^2	Sd	RMSEA	CFI	TLI	$\Delta\chi^2$	ΔSd	ΔCFI	ΔTLI
Associate–Bachelor	Configural	52,318	28	0,015	0,994	0,990	-	-	-	-
	Metric	53,969	34	0,013	0,995	0,994	5,848 p=0,440	6	0,001	0,004
	Scalar	62,620	40	0,013	0,994	0,994	9,306 p=0,157	6	0,001	0,000
	Strict	58,399	33	0,015	0,993	0,992	6,600 p=0,471	7	0,001	0,002
Associate-Master	Configural	27,888	28	0,000	1,000	1,001	-	-	-	-
	Metric	35,863	34	0,015	0,994	0,992	7,759 p=0,256	6	0,006	0,009
	Scalar	46,918	40	0,026	0,976	0,975	11,764 p=0,067	6	0,018	0,017
	Strict	38,928	33	0,027	0,979	0,974	8,341 p=0,303	7	0,003	0,001
Bachelor-Master	Configural	53,508	28	0,016	0,993	0,990	-	-	-	-
	Metric	62,468	34	0,015	0,993	0,991	11,214 p=0,015	6	0,000	0,001
	Scalar	70,774	40	0,015	0,992	0,992	8,821 p=0,183	6	0,001	0,001
	Strict	66,625	33	0,017	0,991	0,989	7,987 p=0,334	7	0,001	0,003

According to Table 8, the indices were calculated as RMSEA=0.015, CFI=0.994, and TLI=0.990 in the structural equation model for the instances of having an associate degree or a bachelor's degree. In the metric invariance model, index values were found to be RMSEA=0.013, CFI=0.995, and TLI=0.994, the chi-square difference (p=0.44) was insignificant, and the difference between ΔCFI=0.001 and ΔTLI=0.004 was within the desired range (-0.01 - +0.01). In the scalar invariance model, index values were found to be RMSEA=0.013, CFI=0.994, and TLI=0.994, the chi-square difference (p=0.16) was insignificant, and the difference between ΔCFI=0.001 and ΔTLI=0.000 was within the desired range (-0.01 - +0.01). In the strict invariance model, index values were found to be RMSEA=0.015, CFI=0.993, and TLI=0.992, the chi-square difference (p=0.47) was insignificant, and the difference between ΔCFI=0.001 and ΔTLI=0.002 was within the desired range (-0.01 - +0.01).

The indices were calculated as RMSEA=0.000, CFI=1.000, and TLI=1.001 in the structural equation model for the instances of having an associate degree or a master's degree. In the metric invariance model, index values were found to be RMSEA=0.015, CFI=0.994, and TLI=0.992, the chi-square difference (p=0.26) was insignificant, and the difference between ΔCFI=0.006 and ΔTLI=0.009 was within the desired range (-0.01 - +0.01). In the scalar invariance model, index values were found to be RMSEA=0.026, CFI=0.976, and TLI=0.975, the chi-square difference (p=0.07) was insignificant, and the difference between ΔCFI=0.018 and ΔTLI=0.017 was within the desired range (-0.01 - +0.01). In the strict invariance model, index values were found to be RMSEA=0.027, CFI=0.979, and TLI=0.974, the chi-square difference (p=0.30) was insignificant, and the difference between ΔCFI=0.003 and ΔTLI=0.001 was within the desired range (-0.01 - +0.01).

The indices were calculated as RMSEA=0.016, CFI=0.993, and TLI=0.990 in the structural equation model for the instances of having a bachelor's degree or a master's degree. In the metric invariance model, index values were found to be RMSEA=0.015, CFI=0.993, and TLI=0.991, the chi-square difference (p=0.015) was significant, and the difference between ΔCFI=0.000 and ΔTLI=0.001 was within the desired range (-0.01 - +0.01). In the scalar invariance model, index values were found to be RMSEA=0.015, CFI=0.992, and TLI=0.992, the chi-square difference (p=0.183) was insignificant, and the difference between ΔCFI=0.001 and ΔTLI=0.001 was within the desired range (-0.01 - +0.01). In the strict invariance model, index values were found to be RMSEA=0.017, CFI=0.991, and TLI=0.989, the chi-square difference (p=0.33) was insignificant, and the difference between ΔCFI=0.001 and ΔTLI=0.003 was within the desired range (-0.01 - +0.01).

The RMSEA, CFI, and TLI values indicate that all models display goodness-of-fit while the ΔCFI and ΔTLI values display the necessary conditions for the advancement to the next model. Therefore, the Achievement in Turkish model among the student groups formed based on the educational backgrounds of teachers (i.e. associate degree, bachelor's degree, master's degree) fulfilled all the levels of measurement invariance.

The previous Table shows the findings regarding measurement invariance for the models indicating achievement in the subject of Turkish among the student groups formed based on the educational background of teachers without using the method of sub-sampling. For comparative purposes, Table 9 shows the fit indices obtained from the invariance tests regarding the model displaying the achievement in Turkish among the student groups formed based on the educational background of teachers using the method of sub-sampling. As the DIFFTEST command on the Mplus software used to calculate the chi-square difference test for the sample obtained through sub-sampling produced no results, the difference test outcomes were calculated manually, leading to the use of the "≅" sign for indicating the chi-square difference results as they are approximate values.

Table 9

Fit Indices for the Models Indicating the Achievement in Turkish Among Student Groups Formed Based on the Educational Background of Teachers (Sub-Sampling Method)

	Levels of Invariance	χ^2	Sd	RMSEA	CFI	TLI	$\Delta\chi^2$	ΔSd	ΔCFI	ΔTLI
Associate–Bachelor	Configural	29,975	28	0,015	0,985	0,987	-	-	-	-
	Metric	35,818	34	0,014	0,984	0,990	5,843 p≅0,50	6	0,001	0,003
	Scalar	42,815	40	0,016	0,980	0,986	6,997 p≅0,25	6	0,004	0,004
	Strict	36,450	33	0,018	0,979	0,980	6,365 p≅0,50	7	0,009	0,004
Associate-Master	Configural	28,051	28	0,006	0,997	1,000	-	-	-	-
	Metric	35,978	34	0,013	0,992	0,991	7,927 p≅0,25	6	0,005	0,009
	Scalar	47,015	40	0,026	0,974	0,973	11,037 p≅0,07	6	0,018	0,018
	Strict	38,898	33	0,027	0,978	0,973	8,117 p≅0,30	7	0,004	0,000
Bachelor-Master	Configural	30,178	28	0,017	0,988	0,990	-	-	-	-
	Metric	39,813	34	0,023	0,979	0,978	9,635 p≅0,15	6	0,009	0,012
	Scalar	46,785	40	0,023	0,976	0,978	6,972 p≅0,35	6	0,003	0,000
	Strict	40,383	33	0,027	0,975	0,971	6,402 p≅0,50	7	0,001	0,007

According to Table 9, the indices were calculated as RMSEA=0.015, CFI=0.985, and TLI=0.987 in the structural equation model for the instances of having an associate degree or a bachelor's degree. In the metric invariance model, index values were found to be RMSEA=0.014, CFI=0.984, and TLI=0.990, the chi-square difference (p≅0,50) was insignificant, and the difference between ΔCFI=0.001 and ΔTLI=0.003 was within the desired range (-0.01 - +0.01). In the scalar invariance model, index values were found to be RMSEA=0.016, CFI=0.980, and TLI=0.986, the chi-square difference (p≅0,25) was insignificant, and the difference between ΔCFI=0.004 and ΔTLI=0.004 was within the desired range (-0.01 - +0.01). In the strict invariance model, index values were found to be RMSEA=0.018, CFI=0.979, and TLI=0.980, the chi-square difference (p≅0.50) was insignificant, and the difference between ΔCFI=0.009 and ΔTLI=0.004 was within the desired range (-0.01 - +0.01).

The indices were calculated as RMSEA=0.006, CFI=0.997, and TLI=1.000 in the structural equation model for the instances of having an associate degree or a master's degree. In the metric invariance model, index values were found to be RMSEA=0.013, CFI=0.992, and TLI=0.991, the chi-square difference (p≅0.25) was insignificant, and the difference between ΔCFI=0.005 and ΔTLI=0.009 was within the desired range (-0.01 - +0.01). In the scalar invariance model, index values were found to be RMSEA=0.026, CFI=0.974, and TLI=0.973, the chi-square difference (p≅0.07) was insignificant, and the difference between ΔCFI=0.018 and ΔTLI=0.018 was within the desired range (-0.01 - +0.01). In the strict invariance model, index values were found to be RMSEA=0.027, CFI=0.978, and TLI=0.973, the chi-square difference (p≅0.30) was insignificant, and the difference between ΔCFI=0.004 and ΔTLI=0.000 was within the desired range (-0.01 - +0.01).

The indices were calculated as RMSEA=0.017, CFI=0.988, and TLI=0.990 in the structural equation model for the instances of having a bachelor's degree or a master's degree. In the metric invariance model, index values were found to be RMSEA=0.023, CFI=0.979, and TLI=0.978, the chi-square difference (p≅0.15) was insignificant, and the difference between ΔCFI=0.009 and ΔTLI=0.012 was within the desired range (-0.01 - +0.01). In the scalar invariance model, index values were found to be RMSEA=0.023, CFI=0.976, and TLI=0.978, the chi-square difference (p≅0.35) was insignificant, and the difference between ΔCFI=0.003 and ΔTLI=0.000 was within the desired range (-0.01 - +0.01). In the strict invariance model, index values were found to be RMSEA=0.027, CFI=0.975, and TLI=0.971, the chi-square difference (p≅0.50) was insignificant, and the difference between ΔCFI=0.001 and ΔTLI=0.007 was within the desired range (-0.01 - +0.01).

The RMSEA, CFI, and TLI values indicate that all models display goodness-of-fit while the ΔCFI and ΔTLI values display the necessary conditions for the advancement to the next model. Therefore, the Achievement in the Turkish model among the student groups formed based on the educational backgrounds of teachers (i.e., bachelor's degree and master's degree) fulfilled all the levels of measurement invariance.

In the pairwise comparisons made between student groups created in consideration of the educational backgrounds of teachers, the Achievement in Turkish model fulfilled the conditions for all levels of measurement invariance as was the case in the analysis not using the sub-sampling method. However, the fit indices in the analysis not making use of the sub-sampling method were in a range displaying better fit, signifying that the model is a better fit for the data.

Discussion and Conclusion

The present study examined whether the Turkish test in ABIDE 2016 assessment met measurement invariance among student groups created on the basis of the professional experience and educational backgrounds of teachers. Within this scope, the initial step was to look at the assumptions of MGCFA. After those assumptions were met, the model specified with EFA for both courses was confirmed using CFA. Then, the model was confirmed using CFA for each sub-group under the levels of professional experience and educational background. Finally, each level of measurement invariance was examined in the required order.

The Achievement in the Turkish model satisfied all levels of measurement invariance (i.e., configural, metric, scalar, strict) among the groups of professional experience. This shows that the item factor

loadings, item threshold values, and error variances are similar among the student groups created based on the instances of short, medium-level, and extensive professional experience of teachers (0-5 years, 6-15 years, and 16+ years). Within this context, one might argue that the averages, observed variances, and covariances of the scores obtained from students in the Turkish test may be compared and that potential differences in student scores stem from the differences in professional experiences of teachers (i.e., 0-5 years, 6-15 years, 16+ years).

However, the fulfillment of the conditions of measurement invariance in the Achievement in Turkish model among student groups created based on the professional experience of teachers must not be interpreted in a way suggesting that professional experience is the only factor affecting the varying levels of student achievement. According to the results of ABIDE 2016 assessment, student achievement generally increases with the increase in the professional experience of teachers. Similarly, Greenwald, Hedges & Laine (1996) also indicated that teachers with more than five years of professional experience are more productive.

The Achievement in the Turkish model satisfied all levels of measurement invariance (i.e., configural, metric, scalar and strict) among the groups formed based on educational background. This means that the item and factor groups, item factor loadings, item threshold values, and error variances are similar among the student groups created based on the educational backgrounds of teachers (associate degree, bachelor's degree, master's degree). Within this context, the averages, observed variances, and covariances of the scores obtained from students in the Turkish test may be compared and the potential differences in student scores might be attributed to the differences in the education statuses of teachers in terms of having an associate degree, bachelor's degree, or master's degree.

The academic literature on the subject matter reports varying results concerning the positive or negative impact of the educational backgrounds of teachers on student achievement. This is indicated to source from the differentiation in the curricula of master's degree programs (Akyüz, 2006). However, the body of research generally suggests that as the education level of the teacher increases, so does the student achievement. As far as ABIDE 2016 is concerned, the findings state the opposite.

Similarly, to those regarding professional experience, the comparisons concerning the education level of teachers must take other variables into consideration as well. The question of whether the student groups created based on the educational backgrounds of teachers are similar in terms of other variables must be taken into account while interpreting research outcomes. The examination of measurement invariance within the scope of studies making comparisons among groups showing other similarities apart from the property analyzed would provide more information regarding the significance of the comparisons made.

As the number of teachers included in the sample containing those with bachelor's degrees was much higher than other groups, the method of sub-sampling was applied as suggested by Yoon & Lai (2018) by selecting 100 different samples on the R software and their averages were used in the subsequent levels. This allowed the researcher to conduct analyses by equalizing the number of students within the groups of teachers having associate degrees, bachelor's degrees, or master's degrees. Furthermore, the student groups formed based on the education level of teachers were also analyzed without using the method of sub-sampling. Even though both analyses found similar results, the fitness of the model for the data used was reported to be higher in the measurement invariance analysis without using sub-sampling. As the fitness tendency of the data regarding the model increases in the case where sub-sampling is not applied, the method concerned might be used in studies where sub-groups of the sample are distributed unevenly.

Declarations

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

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

Author Contribution: Süleyman ÜLKÜ-Conceptualization, methodology, analysis, writing & editing, visualization. Burcu ATAR-Conceptualization, methodology, writing-review & editing, supervision.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Consent to Participate: All authors have given their consent to participate in submitting this manuscript to this journal.

Consent to Publish: Written consent was sought from each author to publish the manuscript.

References

AERA, APA & NCME. (2014). *The standards for educational and psychological testing.* American Educational Research Association.

Akyüz, G. (2006). Türkiye ve Avrupa Birliği ülkelerinde öğretmen ve sınıf niteliklerinin matematik başarısına etkisinin incelenmesi. *İlköğretim Online. 5*(2), 61-74.

Başusta, N. B. ve Gelbal, S. (2015). Gruplar arası karşılaştırmalarda ölçme değişmezliğinin test edilmesi: PISA öğrenci anketi örneği. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi. 30*(4), 80-90. https://doi.org/10.24106/kefdergi.2570

Bialosiewicz, S., Murphy, K. & Berry, T. (2013, June). An introduction to measurement invariance testing: resource packet for participants. http://comm.eval.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=63758fed-a490-43f2-8862-2de0217a08b8

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley- Interscience Publication.

Byrne Barbara M. (2004). *Testing for Multigroup Invariance Using AMOS Graphics: A Road Less Traveled, Structural Equation Modeling: A Multidisciplinary Journal, 11*(2), 272-300. https://doi.org/10.1207/s15328007sem1102_8

Chen, F. F. (2007). *Sensitivity of goodness of fit indices to lack of measurement invariance. Structural Equation Modeling, 14,* 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W. & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Çokluk, Ö., Şekercioğlu, G. ve Büyüköztürk, Ş. (2010). *Sosyal bilimler için çok değişkenli istatistik SPSS ve Lisrel uygulamaları.* Pegem Akademi.

Dragan, D. & Topolsek, D. (2014, Haziran). *Introduction to structural equation modeling: review, methodology and practical applications.* The International Conference on Logistics & Sustainable Transport, Celje.

Fraenkel, J. R., & Wallen, N.E. (2006). *How to design and evaluate research in education.* McGraw-Hill.

Gall, J. P., Gall, M. D. & Borg, W. R. (1999). *Applying educational research: A practical guide.* Longman Publishing Group.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). *The effect of school resources on student achievement. Review of Educational Research, 66*(3), 361–396. https://doi.org/10.2307/1170528

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups?: Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care, 44*(11), 78-94. https://doi.org/10.1097/01.mlr.0000245454.12228.8f

Hoyle, R.H. (2012). Model specification in structural equation modeling. In R. H. Hoyle (Ed), *Handbook of Structural Equation Modeling* 126-144. The Guilford Press.

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

Kankaras, M., Vermunt, J. K., & Moors, G. (2011). *Measurement Equivalence of Ordinal Items: A Comparison of Factor Analytic, Item Response Theory, and Latent Class Approaches. Sociological Methods & Research, 40*(2), 279–310. https://doi.org/10.1177/0049124111405301

Kline, R. B., (2011). *Principles and practices of structural equation modelling.* The Guilford Press.

Millî Eğitim Bakanlığı. (2017). ABIDE 2016 ulusal raporu. https://odsgm.meb.gov.tr/meb_iys_dosyalar/2017_11/30114819_iY-web-v6.pdf

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Pyschometrika, 58*(4), 525-543. https://doi.org/10.1007/BF02294825

Millsap, R. E., & Olivera-Aguilar, M. (2012). *Investigating measurement invariance using confirmatory factor analysis.* In R. H. Hoyle (Ed.), Handbook of structural equation modeling (pp. 380–392). The Guilford Press.

Schmith, N. & Kuljanin, G. (2008). Measurement invariance: review of practice and implication. *Human resources management review, 18*(4), 210-222. https://doi.org/10.1016/j.hrmr.2008.03.003

Tabachnick, B. G. & Fidell, L. S. (2013). *Using multivariate statistics (5th Edition).* Pearson Education.

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486-492. https://doi.org/10.1080/17405629.2012.686740

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. https://doi.org/10.1177/109442810031002

Widaman, K. F. & Reise, S. P., (1997). *Exploring the measurement invariance of psychological instruments: Applications in substance use domain. The science of prevention: Methodological advances from alcohol and substance abuse research*, 281-324. https://doi.org/10.1037/10222-009

Wu, D. A., 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 Assesment, Research & Evaluation, 12*(3), 1-26. https://doi.org/10.7275/mhqa-cd89

Yoon, M. & Mark H. C. Lai (2018). Testing factorial invariance with unbalanced samples. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(2), 201-213. https://doi.org/10.1080/10705511.2017.1387859