

I

J

A

T

E

Volume 8

Issue 2

2021

*International Journal of  
Assessment Tools in Education*

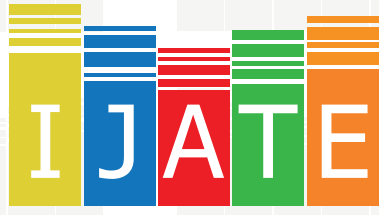
<https://dergipark.org.tr/en/pub/ijate>

<http://www.ijate.net>

e-ISSN: 2148-7456

© IJATE 2021





e-ISSN 2148-7456

<https://dergipark.org.tr/en/pub/ijate>  
<http://www.ijate.net>

**Volume 8**

**Issue 2**

**2021**

Editor : Dr. Eren Can AYBEK  
Address : Pamukkale University, Education Faculty,  
Kinikli Campus, 20070 Denizli, Turkiye  
Phone : +90 258 296 1050  
Fax : +90 258 296 1200  
E-mail : [ijate.editor@gmail.com](mailto:ijate.editor@gmail.com)

Publisher Info : Dr. Izzet KARA  
Address : Pamukkale University, Education Faculty,  
Kinikli Campus, 20070 Denizli, Turkiye  
Phone : +90 258 296 1036  
Fax : +90 258 296 1200  
E-mail : [ikara@pau.edu.tr](mailto:ikara@pau.edu.tr)

Frequency : 4 issues per year (March, June, September, December)  
Online ISSN : 2148-7456  
Website : <http://www.ijate.net/>  
<http://dergipark.org.tr/en/pub/ijate>

Journal Contact : Anıl KANDEMİR  
Address : Department of Educational Sciences, METU, Faculty of Education,  
Üniversiteler Mahallesi, No:1, Ankara, 06800, Turkiye  
E-mail : [kandemiranilk@gmail.com](mailto:kandemiranilk@gmail.com)  
Phone : +90 312 210 4040

*International Journal of Assessment Tools in Education (IJATE)* is a peer-reviewed and academic online journal.

The scientific and legal responsibility for manuscripts published in our journal belongs to the authors(s).



## International Journal of Assessment Tools in Education

*International Journal of Assessment Tools in Education* (IJATE) is an international, peer-reviewed online journal. IJATE is aimed to receive manuscripts focusing on evaluation and assessment in education. It is expected that submitted manuscripts could direct national and international argumentations in the area. Both qualitative and quantitative studies can be accepted, however, it should be considered that all manuscripts need to focus on assessment and evaluation in education.

IJATE as an online journal is sponsored and hosted by **TUBITAK-ULAKBIM** (The Scientific and Technological Research Council of Turkey).

In IJATE, there is no charged under any procedure for submitting or publishing an article.

Starting from this issue, the abbreviation for *International Journal of Assessment Tools in Education* is "*Int. J. Assess. Tools Educ.*" has been changed.

### Indexes and Platforms:

- Emerging Sources Citation Index (ESCI)
- Education Resources Information Center (ERIC)
- TR Index (ULAKBIM),
- ERIH PLUS,
- EBSCO,
- SOBIAD,
- JournalTOCs,
- MIAR (Information Matrix for Analysis of the Journals),
- idealonline,
- CrossRef,
- ResearchBib

## **Editor**

[Dr. Ozen YILDIRIM](#), Pamukkale University, Turkey

## **Editorial Board**

[Dr. Eren Can AYBEK](#), Pamukkale University, Turkey

[Dr. Beyza AKSU DUNYA](#), Bartın University, Turkey

[Dr. Selahattin GELBAL](#), Hacettepe University, Turkey

[Dr. Stanislav AVSEC](#), University of Ljubljana, Slovenia

[Dr. Murat BALKIS](#), Pamukkale University, Turkey

[Dr. Gulsah BASOL](#), Gaziosmanpaşa University, Turkey

[Dr. Bengu BORKAN](#), Boğaziçi University, Turkey

[Dr. Kelly D. BRADLEY](#), University of Kentucky, United States

[Dr. Okan BULUT](#), University of Alberta, Canada

[Dr. Javier Fombona CADAVIECO](#), University of Oviedo, Spain

[Dr. William W. COBERN](#), Western Michigan University, United States

[Dr. R. Nukhet CIKRIKCI](#), İstanbul Aydın University, Turkey

[Dr. Safiye Bilican DEMİR](#), Kocaeli University, Turkey

[Dr. Nuri DOGAN](#), Hacettepe University, Turkey

[Dr. R. Sahin ARSLAN](#), Pamukkale University, Turkey

[Dr. Anne Corinne HUGGINS-MANLEY](#), University of Florida, United States

[Dr. Francisco Andres JIMENEZ](#), Shadow Health, Inc., United States

[Dr. Nicole KAMINSKI-OZTURK](#), The University of Illinois at Chicago, United States

[Dr. Orhan KARAMUSTAFAOGLU](#), Amasya University, Turkey

[Dr. Yasemin KAYA](#), Atatürk University, Turkey

[Dr. Hulya KELECIOGLU](#), Hacettepe University, Turkey

[Dr. Hakan KOGAR](#), Akdeniz University, Turkey

[Dr. Omer KUTLU](#), Ankara University, Turkey

[Dr. Seongyong LEE](#), BNU-HKBU United International College, China

[Dr. Sunbok LEE](#), University of Houston, United States

[Dr. Froilan D. MOBO](#), Ama University, Philippines

[Dr. Hamzeh MORADI](#), Sun Yat-sen University, China

[Dr. Nesrin OZTURK](#), Izmir Democracy University, Turkey

[Dr. Turan PAKER](#), Pamukkale University, Turkey

[Dr. Murat Dogan SAHIN](#), Anadolu University, Turkey

[Dr. Ragıp TERZI](#), Harran University, Turkey

[Dr. Hakan TURKMEN](#), Ege University, Turkey

[Dr. Hossein SALARIAN](#), University of Tehran, Iran

## **English Language Editors**

[Dr. Hatice ALTUN](#) - Pamukkale University, Turkey

[Dr. Arzu KANAT MUTLUOGLU](#) - Ted University, Turkey

## **Editorial Assistant**

[Anil KANDEMİR](#) - Middle East Technical University, Turkey



## TABLE OF CONTENTS

### **Research Articles**

[Remote Assessment in Higher Education during COVID-19 Pandemic](#)

Pages: 181-199, [PDF](#)

Selma ŞENEL, Hüseyin Can ŞENEL

[Evaluation of teacher candidates' life skills in terms of departments and grade levels](#)

Pages: 200-221, [PDF](#)

Dilek ERDURAN AVCI, Fikret KORUR, Sümeyye TURGUT

[Automated Essay Scoring Effect on Test Equating Errors in Mixed-format Test](#)

Pages: 222-238, [PDF](#)

İbrahim UYSAL, Nuri DOĞAN

[An Investigation of Item Position Effects by Means of IRT-Based Differential Item Functioning Methods](#)

Pages: 239-256, [PDF](#)

Sümeyra SOYSAL, Esin YILMAZ KOĞAR

[Validity and Reliability Evidence of Professional Obsolescence Scale According to Different Test Theories](#)

Pages: 257-278, [PDF](#)

Sedagül Akbaba ALTUN, Şener BÜYÜKÖZTÜRK, Merve YILDIRIM SEHERYELİ

[Point and Interval Estimators of an Indirect Effect for a Binary Outcome](#)

Pages: 279-295, [PDF](#)

Hyung Rock LEE, Jaeyun SUNG, Sunbok LEE

[Examining the Dimensionality and Monotonicity of an Attitude Dataset based on the Item Response Theory Models](#)

Pages: 296-309, [PDF](#)

Seval KARTAL, Ezgi MOR DİRLİK

[Examination of Wording Effect of the TIMSS 2015 Mathematical Self-Esteem Scale Through the Bifactor Models](#)

Pages: 326-341, [PDF](#)

Esra OYAR, Hakan Yavuz ATAR

[Teachers' Knowledge and Perception about Dyslexia: Developing and Validating a Scale](#)

Pages: 342-356, [PDF](#)

Duygu TOSUN, Serkan ARIKAN, Nalan BABÜR

[Developing a Two-Tier Proportional Reasoning Skill Test: Validity and Reliability Studies](#)

Pages: 357-375, [PDF](#)

Kübra AÇIKGÜL

[Detecting Differential Item Functioning: Item Response Theory Methods Versus the Mantel-Haenszel Procedure](#)

Pages: 376-393, [PDF](#)

Emily DÍAZ, Gordon BROOKS, George JOHANSON

[Views of Teachers on the Potential Negative Effects of High Stake Tests](#)

Pages: 394-408, [PDF](#)

Mustafa İLHAN, Neşe GÜLER, Gülşen TAŞDELEN TEKER

[The Views of Pre-Service Elementary Teachers About Online and Traditional Peer Assessment](#)

Pages: 409-422, [PDF](#)

Ahmet Oğuz AKÇAY, Ufuk GÜVEN, Engin KARAHAN

[A Guide for More Accurate and Precise Estimations in Simulative Unidimensional IRT Models](#)

Pages: 423-453, [PDF](#)

Fulya BARİS PEKMEZCİ, Asiye ŞENGÜL AVŞAR

[Gathering evidence on e-rubrics: Perspectives and many facet Rasch analysis of rating behavior](#)

Pages: 454-474, [PDF](#)

Inan Deniz ERGUVAN, Beyza AKSU DÜNYA

---

***Review Articles***

---

[Principles for Minimizing Errors in Examination Papers and Other Educational Assessment Instruments](#)

Pages: 310-325, [PDF](#)

Irenka SUTO, Jo IRELAND

## Remote Assessment in Higher Education during COVID-19 Pandemic

Selma Senel <sup>1,\*</sup>, Huseyin Can Senel <sup>2</sup>

<sup>1</sup>Balikesir University, Faculty of Education, Department of Educational Sciences, Balikesir, Turkey

<sup>2</sup>National Defense University, Army NCO Vocational College, Department of Computer Technology, Balikesir, Turkey

### ARTICLE HISTORY

Received: Nov. 02, 2020

Revised: Dec. 31, 2020

Accepted: Jan. 30, 2021

### Keywords:

Remote assessment,  
Measurement and  
evaluation,  
Distance education,  
Online test,  
E-assessment.

**Abstract:** Universities have made a compulsory shift to distance education due to the Covid-19 pandemic. All of the higher education institutions in Turkey have completed 2019-2020 Spring semester using online tools. However, most of these institutions were not fully-prepared to have all of their courses online. Technical inadequacies, lack of qualified online tools, inexperience of instructors and students in distance education have emerged as major issues that institutions have to face. In addition to all, a new question arised; which approaches will be used for assessment. This study aimed to seek the common assessment approaches used through pandemic, how students perceived the quality of the assessment and the pros and cons of using these practices. Additionally, we examined whether participants' perceptions about quality of the assessment differ according to interaction with faculty members and use of online tests. Researchers employed survey design to reply four research questions and used a three-part instrument to collect qualitative and quantitative data. 486 students from 61 universities voluntarily participated in the study. Results indicated assignments are the mostly used tools and students are generally satisfied about the quality of the assessment practices. Another result is that students who interact with faculty members are more satisfied with the quality of the assessment practices. This emphasizes the importance of formative assessment and feedback in remote assessment. Further, students who took online tests are more satisfied with the quality of assessment. Suggestions were made for future research.

## 1. INTRODUCTION

Throughout history, pandemics are known to affect human life in many ways (Martini et al., 2019). The COVID-19 pandemic, which we still largely feel, has also caused critical changes and it also has engendered significant transformation in education activities all over the world (Daniel, 2020). Countries where the COVID-19 pandemic threat has increased, conventional education have been suspended temporarily and the distance education tools were adopted (Bozkurt & Sharma, 2020). In Turkey, as in primary and secondary education institutions affiliated to the Ministry of National Education, higher education institutions have completed 2019-2020 spring semester using distance education. A similar decision was taken for 2020-2021 fall semester.

CONTACT: Selma ŞENEL ✉ [selmahocuk@gmail.com](mailto:selmahocuk@gmail.com) 📍 Balikesir University, Faculty of Education, Department of Educational Sciences, Balikesir, Turkey

ISSN-e: 2148-7456 /© IJATE 2021

It is impossible for 21st century educational institutions to use a method away from technology. Today, innovative tools are commonly used both in-class and out of-class activities (Akçayır & Akçayır, 2018). Distance education is more common thanks to these tools and the number of distance education institutions is increasing. However, a compulsory transition to distance education without adequate preparation may cause problems in different aspects of distance education. Providing the necessary technical infrastructure for distance education, utilizing technological tools and having experienced teaching staff in sufficient numbers in distance education is among the basic needs of distance education (Veletsianos & Houlden, 2019). Absence of basic needs can be predicted to negatively affect the quality of distance education and the extent of this effect is worth researching.

Valid and reliable assessment results are crucial to be able to control whether the educational goals have been achieved or not. Assessment can be carried out during the training in order to identify and then eliminate learning deficiencies as formative assessment. The instructor may explain the assessment results and give feedback. In this respect, assessment practices have important effect in the achievement of educational goals (Chen et al., 2020). In addition, summative assessment have guiding impact by forming a basis for decisions such as being successful in a course, moving to a higher education institution, receiving a diploma or certificate (Biesta, 2009). With these in mind, both formative and summative assessment practices are considered as the cornerstones of instruction.

Formative assessment can be expected to be more prominent in distance education since the students are 'remote' and the possibility of interaction is low. Since there is no conventional classroom environment, the student needs feedback in order to see their deficiencies and mistakes. This requires effective interaction between student and instructor. Instructors should be able to provide students with the opportunity to organize their learning by providing instant feedback, through tests or performance-based techniques (Hatzipanagos & Warburton, 2009). To summarize, "monitoring" and "feedback", which is a part of formative assessment in distance education is gaining more importance. Feedback can be considered as the primary means of student-faculty communication and interaction.

All of the universities in Turkey have completed 2019-2020 spring semester with online tools. Assessment practices were conducted using various techniques like online tests, assignments, and projects. There was no face to face exams. In this period, a new issue has arisen about the quality of the assessment carried out with online tools. Assessment results form students' grade point averages and graduation besides the general achievement goals. In other words, the critical decisions that may affect the lives of individuals were made based on the assessment results and it was the first time that all assessment practices were made upon distance tools.

Learning management systems are widely used in distance education. These tools provide integrated functions like communication, interaction and storage. Canvas, Blackboard, Edmodo, Moodle, Google Classroom and Microsoft Teams are some of these tools. Similarly, video conference tools like Zoom, Skype and Adobe Connect (Koh & Kan, 2020; Nyachwaya, 2020) is latest tools that are common to have online lessons. In addition, these tools can provide a number of advantages for assessment (Araka et al., 2020). The advantages of using these tools in assessment are listed as follows:

**Instant feedback:** It is known that using instant feedback increases the performance of the students in summative assessment (Joint Information Systems Committee, 2010; Shrago & Smith, 2006). Therefore, feedback on assessment results has a critical role in increasing the quality of the learning. Among the tools used in distance education, tests using items that require selection (multiple choice, true-false, matching), are very appropriate for producing instant feedback. Using instant feedback, students may find opportunity to organize self-learning by noticing deficiencies and mistakes.

**Ease of editing based on feedback:** An assignment submitted electronically is easy to examine and edit. Students can comfortably edit and re-organize assignments in line with instructor's feedback. Instructor may plan the re-submission of assignments and students may re-submit the latest version of their work.

**Ease of submitting/responding:** Most of the learning management systems have testing or delivery tools which response and product delivery can be systematically and easily carried out. These tools are widely used for remote-assessment (Moore et al., 2011). In addition, common technological tools such as e-mail or direct messages also offers delivery preferences. Uploading or submitting an assignment to a web-based tool is easier and faster for students to maintain and submit the physically formed product.

**Control and storage:** Online storage, access, and control of tests and assignments are easier with distance education tools. Informative data such as the list of the submitted/missing assignments, submission date and time are automatically kept in most of the distance education tools. The faculty member may save all test documents to internal storage devices (computer, portable disk, etc.) or reach them independently from time and place.

**Providing statistical data:** Besides providing test statistics, distance tools present data about students' participation rates. Although there is important debate about the relation between access rates to learning management systems and completion of course outcomes, instructors may use access or participation data such as access rate, participation time, message rate and message length to gain insight (Murray et al., 2012).

**Potential to enrich assessment tools and products:** The ability to use media such as images, graphics, drawings, audios, videos and animations provided by latest technology can provide richness in assessment by changing assignment framework (Williams et al., 2005). The instructor may submit an animation and ask students to prepare a video as a reflection assignment and share this video on social media to raise the awareness of the society on related subject.

**Providing student participation and motivation:** Computer-based assessment practices, which are able to use interactive techniques and include multimedia such as audios, images, animations and videos may help to increase students' motivation (Cheng & Basu, 2006). In addition, it is well-known that use of instant feedback increases student participation and motivation in distance education (Chaiyo & Nokham, 2017).

**Re-use:** It is simple to copy or re-use an online test or assignment prepared with online tools. As reported above, storage and access to data are limitless and instructors may safely share assessment tools with each other.

In addition to the advantages of use of online tools in assessment, there are also some limitations. The most controversial topic in remote assessment is *test security* (Rovai, 2000). Test security is a critical issue to be able to rely on the test results. Test security is exceptionally important when results are used for critical decisions such as student selection, placement and graduation due to the fact that these decisions have high impact and accountability (Frey, 2018). Preventing cheating, copying and plagiarism in assessment in distance education is challenging. This may overshadow the fairness and reliability of the results obtained by assessment. To prevent this problematic situation, different technologies such as voice and retinal scans have been developed (Jain et al., 2006). However, these high-tech solutions have not yet become widespread. On the other hand, in order to prevent cheating and to increase test security, some other handy techniques may be used such as adding time limitations in online tests, presenting test items or choices randomly (in different order), creating an item pool and presenting random questions to each student, making exams using an open camera (proctoring) and hindering new

web pages/tab (Arnold, 2016; Peterson, 2019). However, *complete test security* is not yet possible even all of these measures are provided.

In addition to new technologies and techniques stated above, preferring appropriate assessment techniques and tools may be another option for higher test security (Nguyen et al., 2020). Some of these techniques may be aligned as assignments, take-home exams, performance tasks, e-portfolios and peer/self assessment forms. However, these tools must be activating higher level skills. In other words, these tools must include items or tasks triggering student's thinking, criticizing, evaluating, creating an idea or product, while preparing students for related tasks or questions. Items and tasks must be unique and must create possibilities to reply with autonomous effort. Otherwise, students may copy from web or from other sources (Rowe, 2004). Rubrics, rating scales and control lists may be used for scoring these tools. Using take-home and open-book exams (Atilgan et al., 2009) is another alternative tool. Open-book exams which allow utilizing books, notebooks and other materials may help to decrease cheating. Take-home exams may be considered as a good example for open book exams.

Besides the security limitations, ICT literacy is another competence for assessment in distance education. The ability of faculty members and students to use technology and related tools or the limitations of these devices (computers, mobile tools, internet) may adversely affect the qualified utilization of assessment tools. Participants should have all the technical infrastructure like software, hardware, and internet connection. Problems in connection speed, disconnection or other technical problems can cause hard-to-compensate results, especially in online tests. Performance-based approaches, which are time independent, can reduce the negative effects of technical deficiencies.

It cannot be denied that computer technologies have created informative, facilitating, and accelerating advantages for developing or using online assessment techniques. However, it should be kept in mind that it is up to faculty members to develop valid and reliable measurement and evaluation. Developing a valid and reliable test is incomparably important to which technology is used. Test designers must consider validity and reliability of the test rather than the type of the online tool.

During the pandemic, faculty members necessarily carried out distance education for all courses and all of the assessment practices were conducted online. However, they had been experienced in face-to-face instruction and they are not fully experienced in neither distance education nor remote assessment. Inexperience, technical problems, or lack of expert personnel might have adversely effect distance education period. Some other limitations may have negative effects on distance education and particularly on assessment. For example, the limitations of the learning management systems or decisions of the administration might have hindered preferences of faculty members. For these reasons, reliability and validity of the assessment results might have been in differentiated. Providing a shot about assessment practices carried out in this very first phase of pandemic will be an important indicator for results of remote assessment and will shed light for the future applications.

The purpose of this research is to examine the assessment practices of universities during the Covid-19 pandemic. For this purpose, answers will be sought for four research questions:

1. How are the higher education students' perceptions about the quality of assessment practices carried out during the Covid-19?
2. What are the assessment approaches that higher education institutions prefer during the Covid-19?
3. What are the views of higher education students about the assessment practices carried out during the Covid-19?



4. Do participants' perceptions about the quality of assessment practices differ according to the interaction with faculty members and use of online tests?

## 2. METHOD

The main aim of the research described in this paper was to present the assessment practices used by universities during COVID-19 pandemic and how students experienced this unique period. This includes gaining an understanding of what practices (distance tools) universities used for assessment, how they used these tools and then to determine the views of students about assessment practices. A survey model was employed in this research using quantitative and qualitative data together. The data obtained for this study consists of the responses from 486 participants from 61 different universities who took distance education for a semester and were evaluated using distance tools.

### 2.1. Study Goup

The study group for this research was determined through convenient sampling. Undergraduate students who are studying at different universities were reached through the social circle of researchers and social networks. They were informed about the research and volunteering students were identified. The study group, consists of 486 students from 61 universities and 69 departments. Since there were too many universities and reporting the names of all 61 universities would not be a necessary and useful data for the research, universities were grouped considering the University Ranking by Academic Performance (University Ranking by Academic Performance [URAP], 2020). Therefore, "university rankings", which rank universities according to various criteria, were used in reporting the universities participating in the study. The universities participating in this study were analyzed according to 11 different "university rankings list" (URAP Turkey, 2020). Being listed in "university rankings list" can provide information about the quality of universities. Accordingly, it was seen that some of the universities participated in this study were not included in any of the "rankings", while some were included in all of the 11 "rankings". Table 1 summarizes the rankings of the universities that participated in this study.

**Table 1.** *Distribution of the universities and faculties according to "university rankings".*

Faculty	0-2	3	4-8	9-11	Total	Percent
Education	10	64	37	77	188	38.68
Arts and Science	10	3	9	3	25	5.14
Fine Arts	3	0	1	1	5	1.03
Law	1	1	2	1	5	1.03
Economics and Administrative Sciences	10	8	4	9	31	6.38
Engineering	8	74	27	8	117	24.07
Medicine	4	7	4	14	29	5.97
Tourism	0	73	1	12	86	17.70
Total	46	230	85	125	486	100.00
Percent	9.47	47.33	17.49	25.72	100.00	

As can be seen in Table 1, study group consists of the students from 8 different faculties. 25% of the participants study at universities which are ranked 9-11 in the university ranking lists. More than half of the participants study at universities which are ranked 0-3 of the university ranking lists. This can indicate that a study group studying at universities with different qualifications. Gender and grades of the participants were summarized in Table 2.

**Table 2.** Gender and grade distribution of study group.

Grade	Female	Male	Total	Percentage
1	74	38	112	23.05
2	64	42	106	21.81
3	111	48	159	32.72
4	61	36	97	19.96
5	5	1	6	1.23
6	3	3	6	1.23
Total	318	168	486	100.00
Percentage	65.43	34.57	100.00	

**Table 2** reveals that that the study group is predominantly composed of female (65.43%). In addition, the frequency of 5th and 6th grades are low. This stems from that undergraduate programs are mainly 4 years in Turkey.

## 2.2. Data Collection

The data collection tool used in this study is developed as a single form. However, data collection tool includes three main parts. The aims and properties of each part of the tool is explained below.

**Part I.** In the first part of the data collection tool, an 11-item instrument was used to determine the students' perception about the quality of the assessment practices carried out during pandemic. As the first step of the scale development procedure, literature about the assessment in distance education were reviewed. Then, first version of the items was written considering the basic principles to be followed in the measurement process of a course. A total of 11 items were written. Instrument is a 5-point Likert type ranging from (1) *totally disagree*, (2) *disagree*, (3) *partially agree*, (4) *agree*, (5) *strongly agree* response categories.

Since the data collection tool is applied as a single form; expert views and pre-trial applications were carried out together for all parts of the tool. The views of three experts from *measurement and evaluation in education* department and two experts who have studies in distance education were consulted and improvements were made in the form. Data about universities, faculties, departments and gender, grade, grade point averages were added to the form to be able to describe participants. The form was uploaded to web for the pre-trial application, and it was applied with seven undergraduate students to see if it has a clear and understandable form. Minor revisions were made in line with the feedbacks.

**Part II.** The second part of the tool is primarily related with the assessment approaches used in the courses during the Covid-19 pandemic. Questionnaire consists of 7 items using 4-point likert type ranging from (1) *never used*, (2) *used in some courses*, (3) *used in most of the courses*, (4) *used in all courses*. The aim of this part is to observe what kind of approaches or techniques were preferred. In this part, the participants are also asked about whether they took online tests. Additionally, students who took the online tests were asked to mark which of the following security measures were taken in the exam.

- There was a time limit.
- The items were presented randomly to each student (order of items was unique for each student).
- Answer choices were presented randomly (order of choices was unique for each student).
- Different items were used (there was an item pool).
- Cameras were required to be open during the exams.
- There was control not to allow opening a new web page/tab.



**Part III.** In the third and last part of the data collection tool, participants' views about the assessment practices were aimed to be determined. In this section, there are open-ended questions that investigate the participants' views about assessment tools used in distance education, about uncovering the assessment preferences of the participants, and comparing face-to-face and distance assessment practices, whether participants experienced technical problems, and revealing participations' communication level with the instructors.

Volunteerism is of great importance for two main reasons; the accuracy of the data and the potential to threat validity and reliability of the instrument. Informed consent form is included on the first page of the e-form to ensure that only volunteers are included as participants. The data were collected in approximately one and a half month with efforts of the researchers using all of their social networks. Because of low number of returns to online surveys, the total number of participants could only reach 486.

### 2.3. Data Analysis

The procedure followed in the analysis of the data are as follows according to the parts of data collection tool and research purposes. MS Excel and IBM SPSS 20 were used for analysis.

1. Since researchers aimed to measure the participants' perceptions about the quality of assessment practices, first part of the instrument was developed as a measurement tool and explanatory factor analysis (EFA) was conducted for reliability.
2. Descriptive statistics were calculated for the data obtained from Part I, Part II and Part III. Frequencies, percentages, total and average points, and standard deviation score were calculated. The findings were plotted so that the results can be easily understood by the reader.
3. Content analysis was conducted for qualitative data. Qualitative data was collected through answers given by the students to the open-ended question located in the Part III of the data collection tool. Details about the trustworthiness of the qualitative analysis were presented separately.
4. One-way ANOVA and independent samples t-test were conducted to answer the fourth research question. Kolmogorov-Smirnov test results ( $p > .05$ ) indicated that data is normally distributed and Levene test showed homogeneity of variances is achieved ( $p > .05$ ).

#### 2.3.1. Construct Validity and Reliability of Instrument

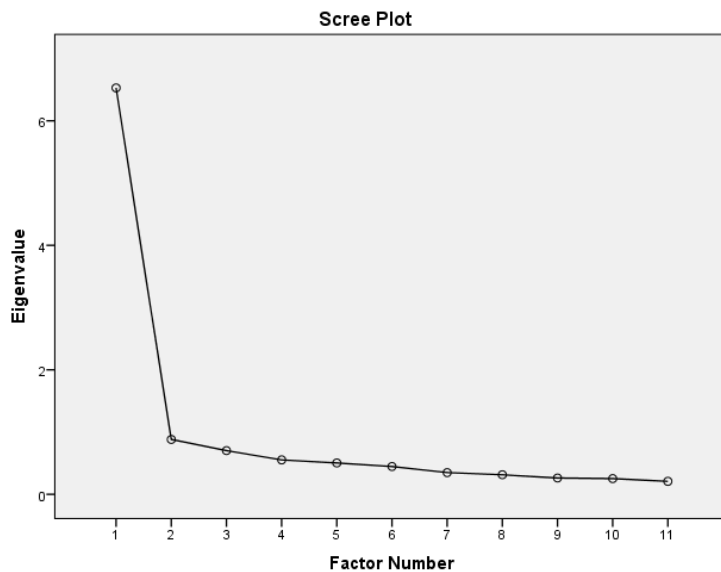
First part of the data collection tool was an *instrument* that measures *higher education students' perceptions about the quality of the assessment practices*. The instrument has 11 items and the highest score that can be obtained from the instrument is 55, and the lowest score is 11. EFA was performed using principal axis factoring method to determine the psychometric properties of instrument. First, researchers examined whether there are one-dimensional/multi-dimensional outliers in the data. It was observed that there are no extreme values. Second, sampling adequacy for EFA was examined. Kaiser-Meyer-Olkin test statistic was found as 0.932 which means perfect sample adequacy for EFA. Third, Bartlett's test of sphericity was used to investigate the multivariate normality. Results ( $X^2(66) = 3454.236; p < .01$ ) indicated that multivariate normality was achieved. EFA results showing item factor loads are presented in [Table 3](#).

Result of the EFA presented that factor loadings of each item are between 0.648-0.842 and are gathered in one dimension. Therefore, researchers decided to use all of the items. Eigenvalue Scree Plot ([Figure 1](#)) indicates that items measure only one dimension.

**Table 3.** Item factor loadings.

Item Number	Items	Factor Loadings
i1	Instructions and explanations in assessment / assignments were understandable and clear.	.759
i2	I have been informed about evaluation and scoring (rubric, evaluation criteria, etc.).	.708
i3	The techniques used in assessment (homework, portfolio, open-ended questions, tests, etc.) were appropriate for the skills desired to be acquired in the lessons.	.842
i4	Assessment was aimed to measure high level skills (creative thinking, critical thinking, problem solving, etc.).	.769
i5	The effectiveness of learning was increased by rapid assessment and giving feedback.	.814
i6	Assessment results and feedback were instant.	.761
i7	The feedback was detailed and instructive.	.800
i8	Assessment practices did not allow cheating and plagiarism.	.669
i9	The assessment results were reliable.	.735
i10	Distinctiveness of test results are high.	.656
i11	The scope of the assessment did not go beyond the provided content.	.648

**Figure 1.** EFA Eigenvalue scree plot.



As can be seen in Figure 1, Eigenvalue of the one-dimension is calculated as 6.529. The variance explained by the one-dimension is found as 55.43%. The Cronbach Alpha internal consistency coefficient of the instrument was calculated as 0.93. The Turkish form of the instrument is provided in Appendix.

### 2.3.2. Trustworthiness

While validity and reliability are used for accuracy of quantitative research, trustworthiness have the same meaning for qualitative study (Guba & Lincoln, 1994). There are some strategies that must be considered like inter-coders agreement, triangulation, peer review, debriefing and rich description (Marshall & Rossman, 2014). Researchers used inter-coder agreement and rich description to provide trustworthiness of the qualitative part of this study.

Two other coders were appointed to provide inter-coder agreement. The first coder is an assistant professor and has Ph.D. degree in measurement and evaluation. Second coder is a Ph.D. student experienced in qualitative methods. Four coders met and discussed the procedure of the study prior to coding and coded six units of data. Researcher and coders compared their findings and negotiated on differences and agreed on codes. After all coding is completed, inter-coder agreement between four coders is found as .88, as Miles and Huberman (1994) reported .80 inter-coder reliability score is satisfying. Researchers and coders compared their findings, negotiated on differences and agreed on results.

Rich description is the second strategy that researcher used for the trustworthiness of qualitative part. Researchers must indicate in-depth information about the procedure and steps of the qualitative phase of the study. The aim of detailed explanation is to provide easy understanding of phases and results (Marshall & Rossman, 2014). The researcher gave details of the qualitative phase to provide rich description so that those who wish to benefit from this research may easily understand the procedure, phases, and findings.

### 3. RESULTS/ FINDINGS

In this section, findings related to research questions will be presented. Four sub-headings were created for four research questions.

#### 3.1. Participants' Perceptions about the Quality of Assessment Practices

The descriptive statistics regarding the participants' perceptions about the quality of the assessment practices carried out during the Covid-19 pandemic process are presented in [Table 4](#).


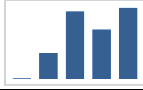

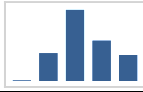
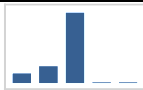


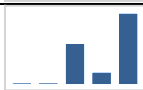


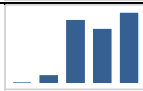
**Table 4.** Descriptives of participants' perceptions about the quality of assessment practices.

Range	Minimum	Maximum	Mean	S.D.	Variance	Skewness	Kurtosis
44.00	11.00	55.00	35.29	11.00	121.01	-.160	-.51

As can be noticed in [Table 4](#), skewness and kurtosis values prove participants' perceptions scores about the quality of assessment practices is normally distributed. However, it can be said that it is skewed to left although not at a significant level. This means big part of the observations are medium/large, with a few observations that are much smaller. As a matter of fact, the distance of the average ( $\bar{X} = 35.29$ ) is closer to maximum score than the lowest score. This presents a clue about the participants' perceptions tend to be relatively moderate to high. However, since this is not a statistically significant distortion, it can be stated that participants' perceptions of the quality of assessment practices are moderate. The distribution of the responses, the average and standard deviation values for each item are presented in [Table 5](#).

According to [Table 5](#), the average of all items except two items (i5 and i10) are found as 3.00 and above. Results show that clarity of the instructions and explanations used in assessment practices are high ( $\bar{X} = 3.56$ ,  $S = 1.23$ ). On the other hand, participants negatively valued about the use of instant assessment and feedback. In other words, instant assessment and giving feedback ( $\bar{X} = 2.94$ ,  $S = 1.34$ ) are not sufficient to increase the effectiveness of the participants' learning. In addition, participants think that the test results do not have enough power to distinguish the students ( $\bar{X} = 2.73$ ,  $S = 1.35$ ).

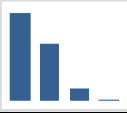
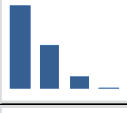

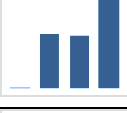



**Table 5.** Participants’ perceptions about the quality of assessment practices.

No	Item	definitely disagree	disagree	partly agree	agree	definitely agree	Mini Graph	$\bar{X}$	S
i1	Instructions and explanations in assessment / assignments were understandable and clear.	41	49	124	139	133		3.56	1.23
i2	I have been informed about evaluation and scoring (rubric, evaluation criteria, etc.).	48	78	125	105	130		3.39	1.30
i3	The techniques used in assessment (homework, portfolio, open-ended questions, tests, etc.) were appropriate for the skills desired to be acquired in the lessons.	43	75	150	108	110		3.34	1.23
i4	Assessment was aimed to measure high level skills (creative thinking, critical thinking, problem solving, etc.).	60	92	139	105	90		3.15	1.27
i5	The effectiveness of learning was increased by rapid assessment and giving feedback.	90	95	136	82	83		2.94	1.34
i6	Assessment results and feedback were instant.	56	83	133	102	112		3.27	1.30
i7	The feedback was detailed and instructive.	77	96	143	92	78		3.00	1.29
i8	Assessment practices did not allow cheating and plagiarism.	80	80	108	88	130		3.22	1.42
i9	The assessment results were reliable.	73	75	132	113	93		3.16	1.31
i10	Distinctiveness of test results are high.	122	97	122	80	65		2.73	1.35
i11	The scope of the assessment did not go beyond the provided content.	45	56	129	117	139		3.51	1.27

### 3.2. Assessment Approaches Used During the Pandemic

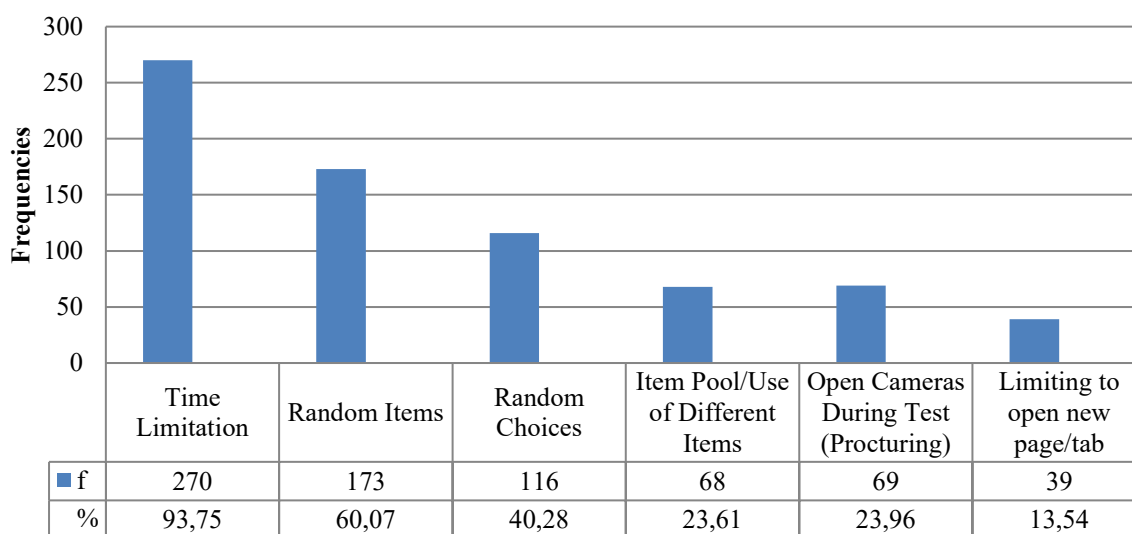
Findings regarding the usage measures of assessment approaches applied in courses during COVID-19 pandemic are presented in Table 6. As can be seen in Table 6, assignments, one of the performance-based tools, is the mostly used approach (total = 1435) overall. Assignments may be used with different techniques and in different forms. Using various approaches together such as projects, portfolios, open-ended items is the second mostly preferred approach (total = 1262). These results indicate that performance-based techniques such as open-ended questions, take home exams, product files or performance tasks were widely preferred during the pandemic. Infrequent use of online tests, peer and self assessment tools and participation indicators is an eye-catching result. As visualized in mini graphs, usage measures of these approaches are generally reported as “never used”.

**Table 6.** Usage measures of assessment approaches.

No	Item	never used	in some courses	in most courses	in all courses	Mini Graph	Total
1	Online tests using items that require selection (multiple choice, T-F, matching)	205	153	76	52		947
2	Online tests using open-ended (written) items	187	135	91	73		1022
3	Online tests using a combination of selection and open-ended items	243	122	68	53		903
4	Assignments with specified time (e.g., 1 week) (open-ended questions, take-home exams, or performance-based techniques such as portfolios, performance tasks)	43	128	124	191		1435
5	Various assessment techniques were used for evaluation (portfolio, research project, open-ended items etc.)	76	159	136	115		1262
6	Peer and / or self-assessment tools	154	175	91	66		1041
7	Discussion forums or other indicators showing participation in distance education	181	158	75	72		1010

Since test security is a problematic issue in remote assessment, participants who attended online tests were asked about the test security measures. 198 (40.74%) of the 486 students stated that they did not take an online test. Figure 2 summarizes the measures taken for the test security during the online tests.

**Figure 2.** Descriptives about the test security measures in online tests.



According to Figure 2, the most preferred test security measure in online tests is using time limitations (93.75%). The second and third mostly used measures is to present items and choices randomly (in different order) (60.07%, 40.28%) for each participant. Use of item pool/providing different questions or using open camera (procturing) are less preferred measures (23%) in online tests. Limiting to open a new web page/tab is the least preferred security measure (13.54%). Participants were asked to declare which other security measures they experienced. Replies of the participants were listed as follows.

- Recording a video narration explaining answers.
- Asking too many items in limited time (e.g. 90 items 20 minutes).
- Limiting the monitor/control of the responded items.

### 3.3. Views of Students on Assessment Practices

Third part of the data collection tool was aimed to identify views of participants about assessment practices. Both quantitative and qualitative data were collected. Four 5-point Likert type items were presented for the quantitative part. Table 7 summarizes the replies of the participants. According to Table 7, the highest average score ( $\bar{X}_4 = 3.76$ ) is related with the interaction of faculty members and students. The second highest average score ( $\bar{X}_1 = 3.24$ ) is about the test anxiety. Despite being positive about the quality of the assessment practices (first research question) and not having technical problems, participants reported that they are highly concerned about remote assessment. Similarly, participants are not likely to prefer remote assessment when face-to-face education begins. Although the low number of participants experienced technical problems is a pleasing finding indicating sufficient infrastructure of distance learning systems, the fact that even a student is experiencing a technical problem may indicate an important problem that will question the validity of the scores and prevent fair measurement.

Table 7. Participants' views on assessment practices.

No	Item	definitely disagree	disagree	partly agree	agree	definitely agree	Mini Graph	$\bar{X}$	S
1	I was more concerned about remote assessment than I feel in face-to-face assessment	90	75	100	71	150		3.24	1.49
2	I prefer remote assessment when face-to-face education begins.	180	59	107	50	90		2.61	1.52
3	I had technical problems in sending assignments/tests etc.	195	99	86	53	53		2.32	1.38
4	I could contact instructor when I had questions about assignments	30	58	100	111	187		3.76	1.25

In the last part of the data collection tool, participants were asked whether they would like to state their views about the assessment practices carried out during COVID-19 pandemic. 175 of the participants answered this part. Using content analysis, codes were grouped into the categories as negative views, positive views and demands of the participants. Codes and frequencies were given in Table 8. Additionally, it was observed that, apart from the focus of this study, participants are inclined to state views comparing face-to-face and distance education.

**Table 8.** *Frequencies of codes.*

Demands	<i>f</i>	Negative Views	<i>f</i>	Positive Views	<i>f</i>
Use of assignments	14	Distinctiveness of scores	11	Independent from time and place	6
Use of online tests	9	Items/assignments out of content	12	Aimed to measure high level skills	4
Interaction and feedback	8	Time limitations	12	Not having exam anxiety	4
Content of the tests/assignments	6	Negligence of the evaluaters	7	Interaction and feedback	2
Use of clear exam/assignment instructions	5	Use of online tests	6		
Use of varied assessment practices	5	Overrated scores	4		
Use of rubrics	4	Limited interaction and feedback	4		
Technical infrastructure	4	Limited Measurement of high-level skills	4		
Use of face-to-face exams	3	Technical problems	4		
Measuring high level skills	3	Lack of clear exam/assignment instructions	4		
Individualized assessment	3	Lack of clear instructions	2		
Test security	2				

Participants highly reported that they demand to use assignments and online tests for assessment. Another demand of the students is about the interaction and feedback. Since distance education do not offer classroom environment, student-student and student-faculty member interaction is getting more importance (Alhih & Ossiannilsson, 2017). Participants also reported negative views. The mostly declared negative view is about the distinctiveness of the scores. There is a common view among students that most there are excessively overrated scores. This may be reasoned from the heavy workload of the faculty members since all of the courses are given online and there was plenty of assignments to mark. Participants also declared negative views about “content of the tests/assignments” and “time limitations”. Participants highly criticized the exams and assignments since they think that content is extensive, faculty members demanded assignments whose subject is out of course content and there are strict time limitations, especially for assignments. Participants’ declared positive views about the time and place independence that distance education presents, aim of measuring higher level skills and exam anxiety but all of them are limited.

### **3.4. Quality of assessment according to Level of Interaction with Faculty Members and Taking Online Tests**

Literature offers strong relationship between interaction and students’ perception in distance education. In this study, researchers decided to examine if there is any significant difference in participants’ perceptions about the quality of assessment practices according to level of interaction with faculty members. Participants were grouped according to their reply one of the items (Item 4 - *I could contant to instructor when I had questions about assingments*) in the third part of the data collection tool. Descriptives are provided in [Table 9](#).



**Table 9.** Participants' replies to Item 4.

No	Item4- I could contact instructor when I had questions about assignments	f	%
1	Strongly disagree	30	6.17
2	Disagree	58	11.93
3	Partly agree	100	20.58
4	Agree	111	22.84
5	Strongly agree	187	38.48

One-way ANOVA was employed, and five groups of participants were compared. Results are presented in Table 10.

**Table 10.** ANOVA results on perception of assessment quality \* interaction level with faculty members.

	Sum of Squares	df	Mean Square	F	Sig.	Sig. Dif.
Between Groups	22149.249	4	5537.312	72.891	.000	1-2; 1-3;1-4;1-5;
Within Groups	36539.995	481	75.967			2-4; 2-5; 3-4; 3-5;
Total	58689.245	485				4-5

ANOVA results proved that participants' perceptions about the quality of assessment practices differ significantly according to participants' level of interaction with faculty members,  $F(4, 481)=72.861, p<.01$ . There is significant difference ( $p<.01$ ) between all levels of participants except *Disagree* and *Partially Agree* groups.

Online tests which are widely used in remote assessment has problems in test security. On the other hand, the qualitative phase of this study reported that students support the use of online tests. With these in mind, we decided to examine whether the use of online tests effect participants' perception about quality of assessment practices. Table 11 summarizes the results of independent samples *t*-test.

**Table 11.** Results of independent samples *t*-test.

Groups	n	$\bar{X}$	S	df	t	p
Attended Online Tests	288	40.48	11.88	484	3.26	.001
Not Attended Online Tests	198	36.48	11.56			

Results of independent samples *t*-test indicated participants' perceptions about the the quality of the assessment practices are significantly different according to participants' attendance to online tests,  $t(484)=3.26; p<.01$ . Participants who took online tests ( $\bar{X}=40.48$ ) have higher perceptions about the quality of the assessment practices than participants who did not attend online tests ( $\bar{X}=36.48$ ).

#### 4. DISCUSSION and CONCLUSION

As badly affected all the routines, pandemic changed the way we teach. While we had theoretical definitions and limited practices of distance education earlier, nowadays, distance education has a meaning for all. Today we use online tools to make remote lessons, to communicate and interact, to assign and collect homeworks and conduct assessment. In this research, we aimed to examine the very first use of remote assessment and participants' views about this unique experience asking four research questions.



First of all, participants reported positively about the quality of remote assessment. However, they reported negatively in two critical items. First, participants agreed that the use of rapid assessment and feedback was insufficient for effective learning. Instant feedback is known as an assistant to distance learners to self-evaluate their learning and increase performance in summative assessment (Koneru, 2017). Similarly, rapid assessment, is critical in distance education courses due to the asynchronous nature of these courses and additional effort was required to confirm that students were ready to receive and respond to feedback properly (Uribe & Vaughan, 2017). The second issue that participants negatively reported about the quality of the assessment practices is distinctiveness of test scores. To explain, students believe that assessment must produce fair test results. Most of the faculty members have experienced remote assessment tools for the first time and this may be a reason for students to feel that distance assessment practices did not yield distinctive test scores. Heavy workload can be pointed as another reason for unfair results. Faculty members gave all of the courses online and they must evaluate plenty of assignments, projects, and other remote assessment tools.

Results indicated that performance-based tools like assignments, performance tasks, portfolios and research projects are the mostly used assessment tools. Online tests which are easy-to-use were found to be used less. Participation to discussion forums or other indicators of participation rates to distance education are used infrequently, too. Additionally, participants reported infrequent use of peer or self assessment tools. However, literature offers to use varied tools for remote assessment (Stödberg, 2012). Limited use of online tests may have resulted from the concern about failing to meet test security. On the other hand, infrequent use of discussion forums or other indicators of participation to distance tools may be because of the inexperience of instructors about remote tools since most of the faculty members used these tools for the first time.

Even though participants did not have any technical problems and they have easy access to faculty members, they supported to use conventional exams rather than remote assessment. Moreover, participants reported that they did not experience anxiety during remote assessment. This may be explained by one of the findings of qualitative phase of the study. Students reported negative views about distinctiveness of the results, and this may be routing participants in favour of face-to-face assessment.

Further information about participants' views on remote assessment was aimed with qualitative data. Views of participants were grouped as positive views, negative views and demands. Participants declared negative views about the distinctiveness of assessment results. We know that assignments are the mostly used tool during pandemic according to the results of first research question and use of performance based tools like assignments, portfolios or projects may be laborious for faculty members (Linn et al., 1991) and this may lead to unfair assessment results. Another negative view is about the items/assignments that are out of content. With the use of online tools, a wider course content may be presented to students with the idea of having more self-studying time in distance education. Lastly, students may need more time for fulfilling performance-based tasks which requires process-oriented workload. Participants have demands, too. First of all, they demand the use of assignments and online tests and needs more interaction and feedback. Additionally, they demand well-defined exam/assignment instructions. Time and place independence, measuring higher level skills and lower level of exam anxiety are found as the positive sides of remote assessment.

Another finding is that participants who have higher levels of interaction with instructors find assessment practices more qualified. In other words, the more students can reach the faculty members and communicate, the more qualified they find the assessment. Student-student and students-instructor interaction or communication is critical in distance education since there are no conventional classrooms (Vlachopoulos & Makri, 2019). A similar result is found when

online tests and participants' perceptions about quality about assessment tools are examined. Participants who take online tests valued assessment tools more qualified. This may stem from that online test takers may feel a real assessment experience through online tests. As reported earlier, most of the students experienced distance education for the first time and they may need to involve a similar assessment tool as in conventional classrooms.

As all studies have, this study also has limitations, too. Although it is aimed a larger study group, only 486 students volunteered to participate in the study. Participants are from 61 different universities and 69 different faculties, but a larger group may yield detailed results. To overcome this limitation, a similar research with a larger group may be conducted. This study focused on students. Faculty members are the practitioners and their views about this phenomenon may help us to develop remote assessment approaches. This study was conducted considering the early stages of the pandemic. Covid-19 is still threatening the face-to-face education and institutions are now experienced in distant education. Future studies may focus on the developments and the latest techniques that institutions used for assessment.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Balikesir University Ethics Committee, 2020-8.

### Authorship contribution statement

**Selma Senel:** Investigation, Development of Data Collection Tool, Visualization, and Analysis.

**Huseyin Can Senel:** Methodology, Analysis, Supervision, and Writing the original draft.

### ORCID

Selma ŞENEL  <https://orcid.org/0000-0002-5803-0793>

Hüseyin Can ŞENEL  <https://orcid.org/0000-0002-7501-9174>

## 5. REFERENCES

- Akçayır, G., & Akçayır, M. (2018). The flipped classroom: A review of its advantages and challenges. *Computers and Education*, 126, 334-345. <https://doi.org/10.1016/j.compedu.2018.07.021>
- Alhih, M., & Ossiannilsson, E. (2017). Levels of interaction provided by online distance education models. *EURASIA Journal of Mathematics Science and Technology Education*, 13(6), 2733-2748. <https://doi.org/10.12973/eurasia.2017.01250a>
- Araka, E., Maina, E., Gitonga, R., & Oboko, R. (2020). Research trends in measurement and intervention tools for self-regulated learning for e-learning environments-systematic review (2008-2018). *Research and Practice in Technology Enhanced Learning*, 15(6), 1-21. <https://doi.org/10.1186/s41039-020-00129-5>
- Bozkurt, A. & Sharma R. C. (2020). Emergency remote teaching in a time of global crisis due to CoronaVirus pandemic. *Asian Journal of Distance Education*, 15(1), 1-6.
- Arnold, I. J. M. (2016). Cheating at online formative tests: Does it pay off? *Internet and Higher Education*, 29, 98-106. <https://doi.org/10.1016/j.iheduc.2016.02.001>
- Atılğan, H., Kan, A., & Doğan, N. (2009). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]*. Anı Yayıncılık.
- Biesta, G. (2009). Good education in an age of measurement: On the need to reconnect with the question of purpose in education. *Educational Assessment, Evaluation and Accountability*, 21(1), 33-46. <https://doi.org/10.1007/s11092-008-9064-9>
- Chaiyo, Y., & Nokham, R. (2017). The effect of Kahoot, Quizizz and Google Forms on the student's perception in the classrooms response system. *2nd Joint International*

- Conference on Digital Arts, Media and Technology 2017: Digital Economy for Sustainable Growth, ICDAMT 2017*, 178-182. <https://doi.org/10.1109/ICDAMT.2017.7904957>
- Chen, Z., Jiao, J., & Hu, K. (2020). Formative assessment as an online instruction intervention. *International Journal of Distance Education Technologies*, 19(1), 1-16. <https://doi.org/10.4018/ijdet.20210101.oa1>
- Cheng, I., & Basu, A. (2006). Improving multimedia innovative item types for computer based testing. *ISM 2006 - 8th IEEE International Symposium on Multimedia*, 557-564. <https://doi.org/10.1109/ISM.2006.92>
- Daniel, S. J. (2020). Education and the COVID-19 pandemic. *Prospects*, 49(1), 91-96. <https://doi.org/10.1007/s11125-020-09464-3>
- Frey, B. (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation*. SAGE Publications. <https://doi.org/10.4135/9781506326139>
- Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 105-117). SAGE Publications.
- Hatzipanagos, S., & Warburton, S. (2009). Feedback as dialogue: Exploring the links between formative assessment and social software in distance learning. *Learning, Media and Technology*, 34(1), 45-59. <https://doi.org/10.1080/17439880902759919>
- Jain, A. K., Bolle, R., & Pankanti, S. (2006). *Biometrics: personal identification in networked society* (Vol. 479). Springer.
- Joint Information Systems Committee. (2010). *Effective assessment in a digital age a guide to technology-enhanced assessment and feedback*. Higher Education Funding Council for England. [https://facultyinnovate.utexas.edu/sites/default/files/digiassass\\_eada.pdf](https://facultyinnovate.utexas.edu/sites/default/files/digiassass_eada.pdf)
- Koh, J. H. L., & Kan, R. Y. P. (2020). Perceptions of learning management system quality, satisfaction, and usage: Differences among students of the arts. *Australasian Journal of Educational Technology*, 36(3), 26-40. <https://doi.org/10.14742/AJET.5187>
- Koneru, I. (2017). Exploring moodle functionality for managing Open Distance Learning e-assessments. *Turkish Online Journal of Distance Education*, 18(4), 129-141. <https://doi.org/10.17718/tojde.340402>
- Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15-21.
- Marshall, C., & Rossman, G. B. (2014). *Designing qualitative research*. SAGE Publications.
- Martini, M., Gazzaniga, V., Bragazzi, N. L., & Barberis, I. (2019). The Spanish influenza pandemic: A lesson from history 100 years after 1918. *Journal of Preventive Medicine and Hygiene*, 60(1), E64-E67. <https://doi.org/10.15167/2421-4248/jpmh2019.60.1.1205>
- Moore, J. L., Dickson-Deane, C., & Galyen, K. (2011). E-learning, online learning, and distance learning environments: Are they the same? *Internet and Higher Education*, 14(2), 129-135. <https://doi.org/10.1016/j.iheduc.2010.10.001>
- Murray, M., Pérez, J., Geist, D., & Hedrick, A. (2012). Student interaction with online course content: Build it and they might come. *Journal of Information Technology Education: Research*, 11(1), 125-140. <https://doi.org/10.28945/1592>
- Nguyen, J. G., Keuseman, K. J., & Humston, J. J. (2020). Minimize online cheating for online assessments during COVID-19 Pandemic. *Journal of Chemical Education*. <https://doi.org/10.1021/acs.jchemed.0c00790>
- Nyachwaya, J. M. (2020). Teaching general chemistry (I) online during COVID-19. Process, outcomes, and lessons learned: A reflection. *Journal of Chemical Education*, 1, 17-21. <https://doi.org/10.1021/acs.jchemed.0c00891>
- Peterson, J. (2019). An analysis of academic dishonesty in online classes. *Mid-Western Educational Researcher*, 31(1), 24-36.

- Rovai, A. P. (2000). Online and traditional assessments: What is the difference? *Internet and Higher Education*, 3(3), 141-151. [https://doi.org/10.1016/S1096-7516\(01\)00028-8](https://doi.org/10.1016/S1096-7516(01)00028-8)
- Rowe, N. C. (2004). Cheating in online student assessment: Beyond plagiarism. *Online Journal of Distance Learning Administration*, 7(2). 1-8.
- Shrago, J. B., & Smith, M. K. (2006). Online assessment in the K-12 classroom: A formative assessment model for improving student performance on standardized tests. In M. Hricko & L. S. Howell (Eds.), *Online assessment and measurement: Case studies from higher education, K-12 and corporate* (pp. 181-195). Information Science Publishing. <https://doi.org/10.4018/978-1-59140-497-2.ch013>
- Stödberg, U. (2012). A research review of e-assessment. *Assessment and Evaluation in Higher Education*, 37(5), 591-604. <https://doi.org/10.1080/02602938.2011.557496>
- University Ranking by Academic Performance. (2020, November). *The position of our universities in the 11 world general rankings in 2019*. <http://tr.urapcenter.org/2019/index.php>
- Uribe, S. N., & Vaughan, M. (2017). Facilitating student learning in distance education: a case study on the development and implementation of a multifaceted feedback system. *Distance Education*, 38(3), 288-301. <https://doi.org/10.1080/01587919.2017.1369005>
- Veletsianos, G., & Houlden, S. (2019). An analysis of flexible learning and flexibility over the last 40 years of distance education. *Distance Education*, 40(4), 454-468. <https://doi.org/10.1080/01587919.2019.1681893>
- Vlachopoulos, D., & Makri, A. (2019). Online communication and interaction in distance higher education: A framework study of good practice. *International Review of Education*, 65(4), 605-632. <https://doi.org/10.1007/s11159-019-09792-3>
- Williams, D. D., Howell, S. L., & Hricko, M. (2005). *Online assessment, measurement and evaluation: Emerging practices*. Information Science Publishing. <https://doi.org/10.4018/978-1-59140-747-8>

## 6. APPENDIX

### Instrument for Student Perceptions About the Quality of the Assessment (Turkish Form)

#### [Ölçme ve Değerlendirmenin Niteliğine İlişkin Öğrenci Algısı Ölçeği]

Bu ölçekte, uzaktan eğitim sürecinde karşılaşmış olduğun ölçme ve değerlendirme işlemlerine ilişkin algının belirlenmesi amaçlanmaktadır. Maddelerin her birini okuyarak, “*Kesinlikle Katılmıyorum, Katılmıyorum, Kısmen Katılıyorum, Katılıyorum, Kesinlikle Katılıyorum*” seçeneklerinden birini işaretlemeniz beklenmektedir. Araştırmaya desteğinizden dolayı teşekkür ederiz.

No	Maddeler	Kesinlikle Katılmıyorum	Katılmıyorum	Kısmen Katılıyorum	Katılıyorum	Kesinlikle Katılıyorum
i1	Ölçme/ödevlendirme sürecindeki yönerge ve açıklamalar anlaşılır ve açıktı.					
i2	Değerlendirme ve puanlamanın nasıl yapılacağı konusunda bilgilendirildim (dereceli puanlama anahtarı, değerlendirme kriterleri vb.).					
i3	Ölçmede kullanılan teknikler (ödev, ürün dosyası, açık uçlu soru, test vb.) derslerde kazandırılmak istenen becerilere uygundu.					
i4	Ölçme ve değerlendirme üst düzey becerileri (yaratıcı düşünme, eleştirel düşünme, problem çözme vb.) yoklar nitelikteydi.					
i5	Eğitim süreci boyunca ölçme yapılarak, dönütler verilerek öğrenme sürecimin etkililiği arttırıldı.					
i6	Ölçme sonuçları ve dönütler hızlıca ulaştı.					
i7	Geribildirimler ayrıntılı ve öğreticiydi.					
i8	Ölçme ve değerlendirme kopya ve intihale (farklı kaynaklardan kaynak göstermeden alma) izin vermeyecek biçimde yapıldı.					
i9	Ölçme sonuçları güvenilirirdi (hatasızdı).					
i10	Sınav sonuçlarının başarılı ve başarısız ayırt ediciliği yüksekti.					
i11	Ölçme kapsamı, sunulan ders içeriği dışına çıkmadı.					

## Evaluation of teacher candidates' life skills in terms of departments and grade levels

Dilek Erduran Avcı <sup>1,\*</sup>, Sumeyye Turgut <sup>2</sup>, Fikret Korur <sup>1</sup>

<sup>1</sup>Burdur Mehmet Akif Ersoy University, Faculty of Education, Department of Science Education, Turkey.

<sup>2</sup>Burdur Mehmet Akif Ersoy University, Institute of Educational Sciences, Turkey.

### ARTICLE HISTORY

Received: June 08, 2020

Revised: Dec. 03, 2020

Accepted: Feb. 02, 2021

### Keywords:

Life skills scale,  
Teacher candidates,  
Department,  
Grade level,  
Confirmatory factor analysis.

**Abstract:** The purposes of this research were (i) testing the factor and model structure of the life-skills scale (LSS) on teacher candidates and (ii) inspecting the life skills of teacher candidates according to their departments and grade levels. The participants consisted of 518 teacher candidates, all of whom were students in their sophomore or senior years in the education faculty of a state university. The data were collected through the LSS, which has 83 items. The confirmatory factor analysis of LSS verified the ten-factor structure for the teacher candidates (aged between 18 and 25). There were no statistically significant differences in the mean value of teacher candidates' life skills according to the grade variable. On the contrary, there were statistically significant differences in the dependent variables according to the department. Future directions of research regarding the educational outcomes of life skills were discussed.

## 1. INTRODUCTION

The term 'life skills' was first used during the psychological consultation intervention phase of the 'project try' program, which was an initiative against poverty (Adkin, 1984). During this program, which is also referred to as "the first life skills program", the term "life skills" was used as the description of the behavioral psychological learning ability required for dealing with the predictable developmental tasks. Adkins (1984) stated that this term was spread to the general culture and gained various meanings. Following the 1960s, there was an increasing interest in life skills programs (Bailey & Deen, 2002). The objectives and the target groups of these programs varied and included, but were not limited to, reduction, adolescence problems, marriage/separation/divorce problems, protection from contagious diseases, occupational problems, occupational and industrial career development, health, death, teacher & consultant training, suicidality in young people, eating habits, and sports (Adkins, 1984; Bailey & Deen,

**CONTACT:** Dilek Erduran Avcı ✉ [derduran@mehmetakif.edu.tr](mailto:derduran@mehmetakif.edu.tr) 📧 Burdur Mehmet Akif Ersoy University, Faculty of Education, Department of Science Education, Burdur, Turkey.

ISSN-e: 2148-7456 /© IJATE 2021



2002; United Nations International Children's Emergency Fund [UNICEF], 2012, p. 10, World Health Organization [WHO], 1997, p. 13).

WHO (2004, p. 4) defined life skills as the positive behaviors that help individuals cope with daily life's difficulties and challenges efficiently. These skills were explicitly described as the psychological skills which assist people in conscious decision making, problem-solving, critical thinking, creative thinking, and efficient communication. In the related literature, there are various classifications regarding life skills. Tan (2018) summarized the definitions and the contexts of five classifications regarding life skills (Table 1) and found out that although Brooks, UNICEF (2012), WHO (1997), The Collaborative for Academic, Social, and Emotional Learning [CASEL], and Fitzpatrick et al. (2014) suggested different classifications for life skills, their definitions were similar within the frameworks of cognitive skills, personal skills, and interpersonal skills.

**Table 1.** Summary of various categories of life skills frameworks (Tan, 2018, p. 21).

Brooks (Ginter, 1999)	WHO (1997)	CASEL	UNICEF (2012)	Fitzpatrick et al. (2014)
Interpersonal communication/ Human relations	Communication/Interpersonal relationships	Self-awareness	Cognitive	Thinking
Problem-solving/ Decision making	Problem-solving/Decision making	Self-management	Personal	Learning
Physical fitness/ Health maintenance	Creative thinking/Critical Thinking	Social awareness	Interpersonal	Practical
Identity development/ Purpose in life	Self-awareness/ Empathy	Relationship skills		
	Coping with emotions/ Coping with stressors	Responsible decision making		

Today, life skills education is an integral part of the education system in many countries in the world. International organizations like UNICEF and WHO report that life skills education is crucial for young people. Since the wealth and the competitive power of the countries are directly related to the qualified workforce (Trilling & Fadel, 2009, p. 7), there is an increasing demand for individuals who possess today's life skills (Erduran Avci & Kamer, 2018). Therefore, many countries put the life skills in the curriculum (The Turkish Ministry of National Education [TMNE], 2018; Indian National Council of Educational Research and Training, 2005; Ministry of Education, Singapore, 2016); modify the curriculum according to the knowledge, skills, and competencies related to the life skills (European Commission / EACEA / Eurydice, 2012); and develop and apply programs that aim to make students gain life skills aligned with their national requirements (Allen & Lohman, 2016; Chauhan, 2016; O'Rourke et al., 2016; UNICEF, 2012).

Skill mismatch can be defined as "the mismatch between the skills of an individual and the skills required for the job they have" (Güneş, 2016; p. 210) and is a common issue in upper education which also affects the graduates (The European Centre for the Development of Vocational Training [CEDEFOP], 2010). The individuals have to learn the required skills to keep pace with life and the era's rapid changes (Khatoon, 2018). Therefore, the education systems, together with the teachers as their practitioners, have a vital role in skill learning. Tenth Development Plan of the Turkish Ministry of Development emphasizes the life skills among the educational objectives as follows:

*“The main objective of the education system is raising productive and happy individuals who possess advanced thinking, perception, and problem-solving capabilities, internalize democratic values and national culture, are open to sharing and communication, has strong artistic and aesthetic emotions, has the entrepreneurial spirit and innovative approach with self-confidence and responsibility, are familiar with using and generating science and technology, and equipped with the basic information and skills required in the information society.”* (Tenth Development Plan for the Republic of Turkey, 2013; p. 32).

The general and specific objectives of the Turkish national education and instruction programs (TMNE, 2018, p. 4) include growing individuals who possess integrated knowledge, skills, and behavior in the selected qualifications, which are defined in the qualifications framework (The Turkish Qualifications Framework [TQF], 2015). A closer look reveals that many life skills are emphasized among the skills mentioned in the programs. Therefore, all teachers, regardless of their branch, are expected to contribute to the development of students’ life skills.

Teachers play a vital role in promoting life skills that prepare students for adulthood (Amutha & Ramganes, 2013; Cassidy et al., 2018; Erduran Avcı & Kamer, 2018; Kaufman, 2013; Kurtdede-Fidan & Aydoğdu, 2018). According to the research, which predicts the causal effect of the interventions during secondary and higher education on life skills development, the ‘teacher quality’ is one of the important effects among all (Schurer, 2017). Due to the differentiating requirements of individuals and new educational approaches, teachers of today have new occupational responsibilities. These new responsibilities require new teacher qualifications in various fields. One of such qualification fields is the skills field, which includes life skills like creative thinking, analytical thinking, and developing self-awareness besides the occupational skills (TMNE, 2017). It is common to perceive that the teacher candidates, who have higher qualifications regarding these skills, would be more successful in gaining life-long learning habits and developing them (Kozikoğlu & Altunova, 2018). Evin Gencil (2013) stated that determining the level of such skills for teachers and teacher candidates contributed to planning the further stages and taking the required measures. According to the studies in the literature, students of art departments had higher skills compared to the students of other departments (Doğramacioğlu, 2016; Kayahan & Çakmakoğlu-Kuru, 2017; Milli & Yağcı, 2017; Otacioğlu, 2007; Sardoğan & Ağaoğlu, 2005).

We see that some domain-specific skills are emphasized in the specific objectives of the curriculum in compulsory education in Turkey. These skills vary according to the department courses (TMNE, 2018a, 2018b, 2018c). For instance, scientific process skills, some life skills (analytical thinking, decision-making, communication, creative thinking, entrepreneurship, and teamwork), and engineering-design skills are domain-specific skills for science course instruction program (TMNE, 2018a), where balanced diet, use of resources, personal care, self-management, and time management are domain-specific skills for the life sciences course instruction program (TMNE, 2018b). These domain-specific skills are similar to the sub-skills in some of the life-skills classifications in the literature (Fox et al., 2003; Hendricks, 1998; WHO, 2004, p. 9). Besides, Cronin and Allen (2017) view these skills as behavioral, cognitive, interpersonal, or intrapersonal competencies that can be learned, developed, and refined. Due to these aspects, it is important to evaluate teacher candidates' life skills based on their departments and grade levels.

Life skills scales are instruments that are used to measure individuals’ life skills. The life skills scales in the literature are generally applied to students in adolescence (Bailey & Deen, 2002; Erawen, 2010; Erduran Avcı & Korur, 2019, June; Greene, 2008; Kadish et al., 2001; Prasad, 2018; Vranda, 2009). There are also studies on young athletes/campers (Cronin & Allen 2017; Garst et al., 2016), teacher trainees (Chauhan, 2016), teacher candidates (Bhardwaj, 2013; Bolat & Balaman, 2017). Life skills is a broad concept that includes a lot of sub-skills (WHO, 1997).



WHO (1997) categorized the core life skills into ten categories from a broad perspective. Therefore, we examined the scales that (i) included the life skills stated by WHO and (ii) were in Turkish literature for cultural similarity. Erduran Avci and Korur's (2019, June) life skills scale (11-18 years) included ten sub-factors and each factor had many items with high representation power. The researchers provided strong evidence about the theoretical structural compatibility, validity, and reliability of this scale. In this study, we were allowed to test the structural compatibility of Erduran Avci and Korur's (2019, June) LSS on teacher candidates, who were between 17 and 25, and use it.

The purpose of this study was to examine the difference among teacher candidates' life skills according to their departments and grade levels. The term "teacher candidates" was used throughout the study with the meaning of "students trained from higher education institutions to become professional teachers" (IGI Global, n.d.). By evaluating the life skills of teacher candidates, this research may contribute to (i) developing solutions and strategies for 'skill mismatch' problem in teacher training, (ii) developing teacher training policies according to the skill needs, and (iii) planning the life skills training of the generations that will have the life skills we need. To accomplish this purpose, the research questions were as follows: (1) Is the LSS instrument valid and reliable for the students at the university level based on the results of the confirmatory factor analyses? (2) Are there any statistically significant differences between the students' average scores of life skills dimensions according to six different departments and two different grade levels?

## 2. METHOD

The descriptive survey model was used to examine the teacher candidates' life skills in terms of different variables. This model explains the information about a topic according to different independent variables. The participants' opinions or features such as interests, skills, or behavior are identified with this model. The main purpose of survey research is to describe the current situation of the research topic (Fraenkel et al., 2011, p. 393).

### 2.1. Participants

With the convenience sampling method, 640 teacher candidates in a state university's education faculty volunteered for and participated in this study. Fraenkel et al. (2011) stated that researchers in social sciences tend to use the convenience sampling method more frequently because it is not possible for researchers to use the time, money, or other resources required for random sample selection. The distribution of the remaining 518 participants by department and grade level are presented in Table 2.

**Table 2.** *Distribution of teacher candidates by department and grade level.*

Department		Grade level		Total
		1st grade	4th grade	
Math-science Education	Science	17	53	70
	Mathematics	29	44	73
Primary education	Primary school	33	32	65
	Pre-school	36	18	54
Turkish-social science education	Turkish Language	32	17	49
	Social science	15	13	28
Fine arts	Music	9	15	24
	Art	8	9	17
Educational science	Guidance and Psychological Counselling [GPC]	44	25	69
Foreign language	English Language	26	43	69
Total		249	269	518

Among these participants, the data of 122 participants whose data were found to be inconsistent (such as giving the same answers to most of the questions one after the other) and/or they left the question items in the scale blank were not included in the further analysis.

## 2.2. Variables

The variables that were used in the statistical analysis of this research are presented in Table 3. The details of two independent variables (grade level and department) and ten dependent variables, namely the scores for the dimensions, are provided in the table.

**Table 3.** *Description of the variables.*

Variable Name	Variable (wrt types)	Variable (wrt values)	Derived/Taken Items from the Scale	Variable Label / Source	Min.-Max.
Grade Level	Independent	Categorical	Demographic#1	1, 4	-
Department	Independent	Categorical	Demographic#2	1, 2, 3, 4, 5, 6	-
Critical thinking	Dependent	Continuous	1-6	Total mean	1-5
Creative thinking	Dependent	Continuous	7-16	scores within	
Decision making and problem-solving	Dependent	Continuous	17-28	each category	
Coping with stress and emotions	Dependent	Continuous	29-39		
Interpersonal relationship and communication	Dependent	Continuous	40-46		
Empathy	Dependent	Continuous	47-53		
Self-awareness	Dependent	Continuous	54-65		
Self-respect	Dependent	Continuous	66-73		
Teamwork	Dependent	Continuous	74-78		
Social responsibility	Dependent	Continuous	79-83		

## 2.3. The Instrument (LSS) and Data Collection Process

The LSS, which was developed by Erduran Avcı and Korur (2019, June) for evaluating the life skills of students at puberty, was used in this study. The scale was created by Erduran Avcı and Korur (2019, June) following the five-stage approach proposed by Hinkin (1998). The stages are as follows: item generation (creating the initial item pool), scale management (including expert views), initial item reduction (including exploratory factor analysis [EFA], confirmatory factor analysis [CFA], and convergent/discriminant validity (reporting the validity issues). The execution of the stages was performed on two different groups of students aged between 11 and 18. Six hundred seventy-nine students (EFA) were in the first study group and 585 students (EFA) were in the second study group. The factor analysis fit of the data, which was obtained by applying the scale to the first group, was evaluated using the Kaiser–Meyer Olkin (KMO) coefficient, and the sample size sufficiency was evaluated with Bartlett Sphericity Test. The fit of both values was confirmed (KMO value, .957; Bartlett Sphericity,  $\chi^2= 27350.787$ ,  $p<.001$ ). According to the explanatory factor analysis results, which was performed by varimax rotation of principal component analysis, 83 items of the LSS with load factors greater than the threshold were grouped under 10 factors with eigenvalues greater than one. These factors represented the dimensions of the scale. The dimensions and the numbers of items were as follows: Critical thinking (1-6), creative thinking (7-16), decision making and problem-solving (17-28), coping with stress and emotions (29-39), interpersonal relations and communication (40-46), empathy

(47-53), self-awareness (54-65), self-respect (66-73), teamwork (74-78), and social responsibility (79-83). The items of LSS were five-point Likert type (1: *strongly disagree*, 5: *strongly agree*) and the average scores for dimensions were 1 and 5 for minimum and maximum, respectively. Higher scores resembled students' higher perception of life skills. The total variance of these dimensions explained 51.07% of the variance. The factor load values varied between .32 and .81. Cronbach's alpha internal consistency coefficient was .964 for the whole model, where it varied between .717 and .916 for the dimensions. The average scores varied between 3.15 (teamwork) and 4.14 (empathy). After the application of LSS to the second workgroup, DFA model fit indices were calculated as  $\chi^2(3268)= 5953.19$   $p < .001$ ;  $\chi^2/sd= 1,822$ , RMSEA= .0038, SRMR= .049, CFI= .900, and IFI= .901. Cronbach's alpha internal consistency coefficient for the whole scale was .973 and .750 to .940 for the dimensions. The average scores of the second phase's dimensions varied between 3.40 (teamwork) and 4.20 (empathy). These findings were found to be coherent to the hypothetic structure of the LSS suggested by Erduran Avci and Korur (2019, June); the composite reliability, convergent validity, and discriminant validity values were in the acceptable range; and this scale was a proper instrument which could be used in assessing life skills for the future studies. We have cooperated with two domain experts to qualify LSS as a proper instrument for the university students out of the specified age range in the original study. After evaluating the appearance and content of LSS, the experts suggested that LSS could be applied without any changes. LSS was originally in Turkish and sample items in the original language are presented in Figure 1.

**Figure 1.** Sample items from the LSS (in Turkish).

	1	2	3	4	5
1. Kanıtlar yanıldığını gösterdiğinde, düşüncelerimi değiştiririm.					
2. Bir olayı çeşitli açılardan değerlendirebilirim.					
3. Bir olay sonucunda doğabilecek riskleri değerlendirebilirim.					
4. Fikirlerimi, gerçekler ve deneyimler ile oluştururum.					
5. Kendimi geliştirmek için yaptığım her hareketi eleştiririm.					
6. Nedenleri ve kanıtları temel alarak bir durumu anlamaya çalışırım.					
7. Başkalarından fikir ve öneri alırım, ancak onlara inanmadan önce kendim analiz ederim.					
8. Bir işi farklı tarzda/yenilikçi yapmaktan hoşlanırım.					
9. İşlerimi dikkatli yapmaya özen gösteririm.					

At the start of the data collection process, we obtained the required permissions to apply the LSS to the teacher candidates. We made the volunteer teacher candidates fill the LSS forms at their convenience. The first two authors conducted the data collection. It took approximately 20 minutes for a teacher candidate to fill out the LSS.

#### 2.4. Data Analysis Procedure

To analyze the answer to the first research question, we ran the default model, which was constrained by the factor loadings, in AMOS and tested the model fit to the ten-factor structure of the original LSS. CFA process is a statistical technique and it starts with a hypothesis that suggests that there is a relation between the observed variables and the hidden variables beneath them (Child, 1990). According to Mahalanobis distance  $p < .001$  (Tabachnick & Fidell, 2007, p. 99), the outliers were confirmed and 23 students' data were excluded and CFA was processed with data of 495 students. It was stated that the minimum sample size to perform the CFA can be taken as  $N \geq 100$  to 200 or can be calculated as at least 5 to 10 participants per parameter released (Bentler & Chou, 1987; Brown, 2006). Determining the sample size with general acceptances may reveal poor generalizability. For obtaining sufficient statistical power and suitable precision of parameter estimates in CFA, the sample size might be deducted from the

complexity of the model, amount of missing data, and other variables (such as number of observed variables, number of latent variables, and probability level; Brown, 2006). These features will vary widely depending on the data sets in the studies (Brown, 2006). In this context, by entering anticipated effect size as .5 (medium effect size is generally accepted in science education research), desired statistical power level as .95, number of latent variables as 45, number of observed variables as 83, and probability level as .05 values, the recommended minimum sample size was found to be 441 for CFA through an online calculator (DanielSoper, n.d.). Even though the number of participants in the sample group was appropriate according to our model, it should be considered carefully in terms of the study's generalizability. Data were examined for normal homoscedasticity. The common fit indices are given in Table 4 with their critical value ranges.

In addition to the values in Table 4, Hu and Bentler (1999) determined phased criteria, which will keep Type I and Type II errors at a minimum while maintaining an acceptable fit between the data and the model, as a) SRMR value close to or lower than .08, b) RMSEA value close to or lower than .06, and c) CFI value close to or greater than .95. In this study, to determine the model fits from the standardized scores, we used Hu and Bentler's (1999) above-mentioned model fit criteria.

**Table 4.** Fit indices and critical value ranges.

Fit indices	Good fit	Acceptable fit
$\chi^2/sd$	$0 \leq \chi^2/df \leq 2$	$2 < \chi^2/df \leq 3$
RMSEA	$0 \leq RMSEA \leq .05$	$.05 < RMSEA \leq .08$
SRMR	$0 \leq SRMR \leq .05$	$.05 < SRMR \leq .10$
IFI	$.95 \leq NFI \leq 1.00$	$.90 \leq NFI < .95$
CFI	$.95 \leq CFI \leq 1.00$	$.90 \leq CFI < .95$

Note: Adopted from Schermelleh-Engel et al. (2003).  $\chi^2$  = chi-square,  $df$ =degree of freedom, RMSEA = Root Mean Square Error of Approximation, SRMR = Standardised Root Mean Residual, IFI = Incremental Fit Index and CFI = Comparative Fit Index.

To find the answer to the second research question, we examined the interaction of six different departments and two different grades by using MANOVA. The analysis proved that there was a statistically significant interaction (grade\*department) effect on the average scores of the students [Pillai's Trace = .183,  $F(50, 2505) = 1.905, p < .05$ , partial  $\eta^2 = .037$ ]. In other words, the data suggested that the effect of studying in different departments on LSS dimension scores was not the same for 1st-grade and 4th-grade students. Since this analysis was performed on interaction with  $2*6=12$  different variables, we thought that it might be caused by the number of participants in each group (specifically the number of students in different departments). To eliminate this possibility, we assigned a new independent variable for each group and performed MANOVA again. We found that there were no statistically significant differences in further analysis. Therefore, we examined single main effects instead of department\*grade interaction. In this study, we analyzed the statistically significant differences between the students' average scores for 10 dimensions according to two different grade levels and six different departments by conducting separate MANOVAs. We confirmed that the observations were independent, and the sample size was sufficiently large for MANOVA groups. We also conducted preliminary analyses to test the assumptions of MANOVA.

The outliers in the data were analyzed in terms of Mahalanobis distances ( $p < .001$ ), for the assumption of absence of multiple variable outliers and MANOVA was carried out with 518 students' data (Tabachnick & Fidell, 2007, p. 99). For the assumption of the normal distribution of the dependent variables for each independent variable, skewness and kurtosis values for 10

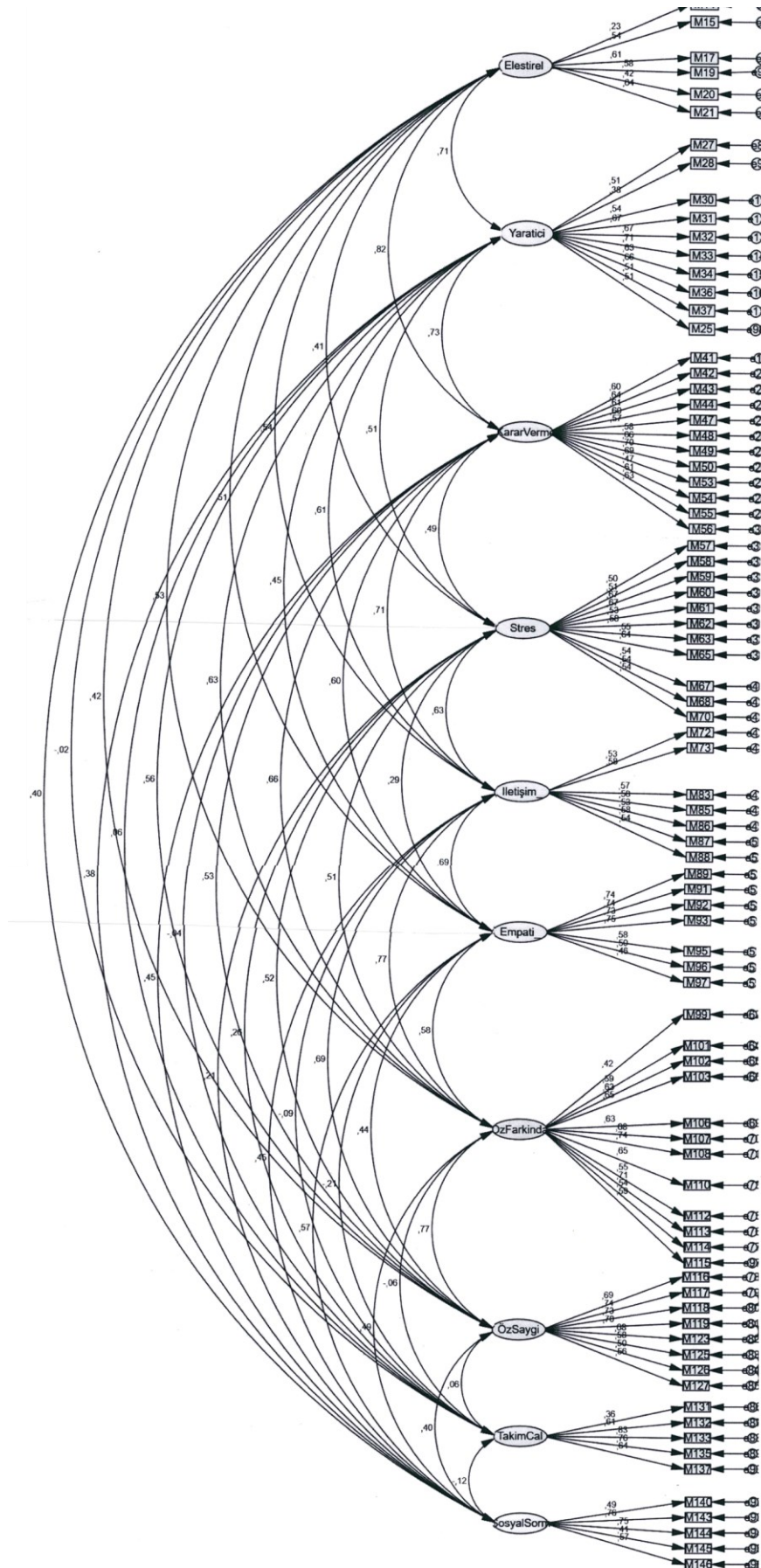
dimensions were inspected for -1.5 to +1.5 points range. At the end of this process, we assumed that the data fit normal distribution [Byrne, 2010; extremum points for the skewness between -.072 (stress) and -1.005 (social responsibility); extremum points for the kurtosis -.118 (teamwork) and .765 (social responsibility)]. To meet the absence of multicollinearity assumptions, we inspected the scatter-plot matrix graphs to confirm the linear relations among the dependent variables. Besides, we observed that there was a low to moderate correlation among the dependent variables ( $<.80$ ); and there was no multicollinearity (Tabachnick & Fidell, 2007). For the assumption of homogeneity of variable matrices, significant differentiation was found among the groups according to Box's M test performed based on grade levels and departments (according to grade levels: Box's  $M = 98.426$ ,  $F(55, 849453.550) = 1.753$ ,  $p < .05$ ; according to departments: Box's  $M = 377.470$ ,  $F(275, 164505.213) = 1.293$ ,  $p > .001$ ). If group sizes are above 30, the MANOVA is robust against violations of homogeneity of variance matrices assumption (Allen & Bennett, 2008; Hair et al., 2006; Tabachnick & Fidell, 2007). Furthermore, Tabachnick and Fidell (2007) recommended to test the Box's M at the  $p=.001$  level for unequal sample sizes; if M is not significant at the .001 level, it may be concluded that significance tests in MANOVA may be robust. The MANOVA results were evaluated with Pillai's Trace test data, which is widely accepted as a stronger test than Wilk's Lambda value (Field, 2009). According to grade levels, the findings of Levene's test showed that the assumption of homogeneity of variances was satisfied for all of the LSS dimensions ( $p > .05$ ). The findings of Levene's test according to departments showed that the assumption of homogeneity of variances was satisfied except for five dimensions: critical thinking score [ $F(5,512)=3.775$ ;  $p=.002$ ]; creative thinking score [ $F(5,512)=.481$ ;  $p=.790$ ]; decision making & problem-solving score [ $F(5,512)=.903$ ;  $p=.479$ ]; coping with stress and emotions score [ $F(5,512)=2.699$ ;  $p=.020$ ]; interpersonal relations and communication score [ $F(5,512)=1.041$ ;  $p=.393$ ]; empathy score [ $F(5,512)=4.801$ ;  $p=.000$ ]; self-awareness score [ $F(5,512)=2.191$ ;  $p=.054$ ]; self-respect score [ $F(5,512)=1.897$ ;  $p=.093$ ]; teamwork score [ $F(5,512)=4.731$ ;  $p=.000$ ], social responsibility score [ $F(5,512)=3.650$ ;  $p=.003$ ]. Further analyses provided for MANOVA (such as Tukey's HSD) are sensitive to unequal variances but multiple comparison procedures by SPSS (e.g. Tamhane's T2, Dunnett's T3, or Dunnett's C) are provided for such cases, where unequal group sizes or high variances ratios (Field, 2009; p. 374). In this study, we examined the dimensions, which did not satisfy the assumption of homogeneity of variances, with Tamhane's T2 index instead of Tukey's HSD. According to these results, the related assumptions of the MANOVA were met.

### 3. RESULT / FINDINGS

In this phase, we conducted a CFA to confirm that the structure, which was obtained by applying LSS to the teacher candidates, was compliant to the structure, which was obtained by the application of LSS to the students aged between 11 and 18. In the beginning, we run CFA for the 10-factor structure of LSS to discover the findings for the first research question. Figure 2 presents the 10-factor structure with 83 items and their corresponding loads. The inspection of model fit indices and detailed model parameter analyses revealed that the fit indices of the 10-factor structure were close to the corresponding acceptable threshold values in Table 3 [ $\chi^2(3249, 495) = 5224.521$ ,  $p < .001$ ;  $\chi^2/df = 1.608$ , RMSEA = .035, SRMR = .0527; CFI = .877, IFI = .878, RMR = .046, and AGFI = .785]. Also, the scale's fit threshold values, which are the combinations of SRMR, RMSEA, IFI, and CFI values, satisfied the phase criteria of Hu & Bentler (1999). The findings of the application of LSS to the teacher candidates were in an acceptable harmony with the Erduran Avci and Korur (2019, June)'s a hypothetical structure with 10 dimensions.



Figure 2. The path diagram of the ten-factor structure of the LSS.



The separate MANOVAs, which were conducted to answer the second research question, indicated that there were no statistically significant differences between the students' average life skills score in 10 dimensions according to two different grade levels [Pillai's Trace = .028,  $F(10, 507) = 1.445$ ,  $p = .157$ , partial  $\eta^2 = .028$ ]. There were low to medium significant differences in the student scores in the dimensions of LSS with regards to the students' departments [Pillai's Trace = .242,  $F(50, 2535) = 2.573$ ,  $p < .05$ , partial  $\eta^2 = .048$ ]. Further analyses were conducted to find out the dimensions with such interaction. It was found that there were statistically significant low to medium mean differences for the dimensions: critical thinking, low [ $F(5, 512) = 6.135$ ,  $p = .000$ , partial  $\eta^2 = .057$ ]; creative thinking, medium [ $F(5, 512) = 6.902$ ,  $p = .000$ , partial  $\eta^2 = .063$ ]; decision making & problem-solving, medium [ $F(5, 512) = 7.239$ ,  $p = .000$ , partial  $\eta^2 = .066$ ]; coping with stress and emotions, low [ $F(5, 512) = 3.581$ ,  $p = .000$ , partial  $\eta^2 = .034$ ]; interpersonal relationship and communication, low [ $F(5, 512) = 3.122$ ,  $p = .009$ , partial  $\eta^2 = .030$ ]; empathy, low [ $F(5, 512) = 5.394$ ,  $p = .000$ , partial  $\eta^2 = .050$ ]; self-awareness, medium [ $F(5, 512) = 7.340$ ,  $p = .000$ , partial  $\eta^2 = .067$ ]; self-esteem, low [ $F(5, 512) = 5.055$ ,  $p = .000$ , partial  $\eta^2 = .047$ ]; teamwork, low [ $F(5, 512) = 5.007$ ,  $p = .000$ , partial  $\eta^2 = .047$ ]; social responsibility, low [ $F(5, 512) = 3.981$ ,  $p = .001$ , partial  $\eta^2 = .037$ ] (Cohen, 1988). **Table 5** presents the results of post hoc analyses regarding this significant difference according to the departments.

After inspecting the significant differences among the departments in **Table 5**, it can be stated that the average scores of the students in the Fine Arts department in critical thinking, creative thinking, decision making, stress, self-awareness, self-respect, teamwork, and social responsibility were higher. There is at least one dimension, in which the students in the Fine Arts department was significantly higher than the students of the other five departments. On the other hand, it was found that the average scores for critical thinking, creative thinking, decision making, communication, empathy, self-awareness, and social responsibility were higher in the Primary Education department students (except the fine arts department students). Just for self-awareness, the average scores of the students in the Foreign Languages department were significantly higher than the ones of GPC students ( $p = .004$ ,  $\bar{X}_{\text{difference}} = 3.8406$ ). There were no cases where the remaining department students' average dimension scores were significantly higher than the other departments. The average scores of GPC students were lower than the corresponding average score of at least one department, except stress and communication dimensions.

**Table 5.** Post hoc Analysis for MANOVA.

Dependent Variable				Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Critical thinking	Tamhane T2	Primary education	GPC	1.4274*	.42907	.017	.1493	2.7054
			Fine arts	Math-Science	2.3114*	.44878	.000	.9572
			GPC	2.7239*	.50614	.000	1.2043	4.2435
			Foreign language	1.9703*	.49198	.002	.4915	3.4491
Creative thinking	Tukey HSD	Primary education	GPC	2.8011*	.82296	.009	.4470	5.1552
			Fine arts	Primary education	3.0119*	.98490	.028	.1945
			Math-Science	4.3981*	.96349	.000	1.6420	7.1542
			Turkish-social science	4.1334*	1.05148	.001	1.1256	7.1412
			GPC	5.8130*	1.07245	.000	2.7452	8.8808
			Foreign language	3.2913*	1.07245	.027	.2235	6.3591
Decision making and problem-solving	Tukey HSD	Primary education	Math-Science	2.9140*	.77992	.003	.6830	5.1450
			GPC	3.3280*	.95109	.007	.6073	6.0486
		Fine arts	Math-Science	5.0607*	1.11350	.000	1.8755	8.2459
			GPC	5.4747*	1.23943	.000	1.9293	9.0202
			Foreign language	4.4168*	1.23943	.005	.8713	7.9622
Coping with stress and emotions	Tamhane T2	Fine arts	Math-Science	5.4612*	1.51073	.010	.8411	10.0813
Interpersonal relationship and communication	Tukey HSD	Primary education	Math-Science	1.5041*	.51117	.040	.0418	2.9663



Table 5. *Continues.*

Empathy	Tamhane T2	Primary education	Math-Science	1.7292*	.38576	.000	.5892	2.8691	
			Turkish-social science	2.1022*	.56977	.005	.3984	3.8059	
			GPC	1.8083*	.52587	.012	.2346	3.3820	
Self-awareness	Tukey HSD	Primary education	GPC	3.7318*	.93283	.001	1.0634	6.4002	
			Fine arts	Math-Science	4.5705*	1.09212	.000	1.4465	7.6946
				Turkish-social science	3.9721*	1.19186	.012	.5628	7.3815
		GPC		6.3510*	1.21563	.000	2.8736	9.8284	
		Foreign language	GPC	3.8406*	1.04957	.004	.8382	6.8429	
			Self-esteem	Tukey HSD	Fine arts	Math-Science	3.5478*	.86535	.001
Turkish-social science	3.0184*					.94437	.018	.3169	5.7198
Foreign language	GPC	4.5345*			.96321	.000	1.7792	7.2898	
	Foreign language	2.9548*			.96321	.027	.1995	5.7101	
Teamwork	Tamhane T2	Fine arts	Primary education	3.7866*	1.01728	.007	.6772	6.8960	
			Math-Science	3.7339*	.99683	.007	.6767	6.7912	
			GPC	3.8657*	1.02796	.006	.7272	7.0042	
Social responsibility	Tamhane T2	Primary education	Math-Science	1.0358*	.33415	.032	.0484	2.0232	
			GPC	1.6172*	.47608	.014	.1913	3.0431	
		Fine arts	Math-Science	1.3012*	.43006	.048	.0056	2.5968	
			GPC	1.8826*	.54767	.012	.2427	3.5226	

#### **4. DISCUSSION and CONCLUSION**

This study was conducted for two purposes: i) to test the 10-factor theoretical structure of LSS for teacher candidates aged between 18 and 25, ii) to find out whether the life scale dimension scores of the teacher candidates varied according to the departments and grade levels. The findings of the study are discussed below based on these two purposes.

LSS, which was developed by Erduran Avcı and Korur (2019, June) was applied to the teacher candidates in the research group of this study. The results of CFA indicated that the structure model of the scale, which included 10 dimensions and 83 items, was confirmed. We can say that LSS did not perform perfectly according to the fit indices and some correlation incompatibilities. However, the 10-factor structure was very close to the acceptable ranges according to the model fit indices and the values obtained by detailed parameter analyses for the model. Reasons for this fact might include (i) Erduran Avcı and Korur (2019, June) followed a well-planned and systematic process to develop the scale-LSS, which was used in this study, and (ii) the structural validity of the scale, together with the items in its dimensions, was high. Also, possible similar expectations and perceptions of the students of puberty, to whom the original scale was applied, and the teacher candidates, to whom the scale was applied in this study, might be another reason. It is common knowledge that puberty can continue until the twenties (Çardak, 2013, pp. 62-64). At ages 18 to 25, one usually attends university and this period covers late puberty and early adulthood. During this period, young individuals build new social relations and keep improving themselves for the rest of their life. According to the “life-span, life-space” theory (Super, 1990), the period between ages of 15 to 24 is the exploration phase. The individuals in the exploration phase explore their interests, skills, values, and more (Eryılmaz & Mutlu, 2017). Therefore, although the age groups of the samples in this study and Erduran Avcı and Korur (2019, June) were different, it can be stated that these two age groups have some intersections, common skills, and perceptions. A few studies also examine the life skills of teacher candidates in the literature (Bhardwaj, 2013; Bolat & Balaman, 2017; Chauhan, 2016).

The analyses of ten sub-factors of LSS showed that there were no statistically significant differences in the life skills of teacher candidates according to their grade levels but there were significant differences according to department variable. Teacher candidates' scores for all of the LSS sub-factors (critical thinking, creative thinking, decision-making & problem-solving, coping with stress and emotions, interpersonal relations & communication, empathy, self-awareness, self-respect, teamwork, and social responsibility) varied significantly according to their departments. There were significant differences in favor of fine arts, primary education, and foreign language departments compared to many other departments. Among those, the most significant differences were observed in the fine arts department. The scores of the students in the fine arts department were different compared to many other departments in eight dimensions (critical thinking, creative thinking, decision making & problem-solving, coping with stress and emotions, self-awareness, self-respect, teamwork, and social responsibility). Specifically, there was a significant difference in favor of the fine arts department in creative thinking sub-dimension when compared to the other departments. In Turkey, the fine arts departments accept students by a special talent exam, which is unique to each fine arts department, where all other departments accept students by a central exam named higher education institutions exam [HEIE]. Therefore, the researchers think that this result, which is in favor of the fine arts students, is natural because the students of the fine arts department were accepted to the university with a completely different assessment process. Similarly, Sardoğan and Ağaoğlu (2005) stated that the students in visual arts, music, and physical training departments had a higher level of emphatic skills than the students who were accepted to the university HEIE. Kayahan and Çakmakoğlu Kuru (2017) states that the departments like visual communication design, which accept students by a talent exam, were more successful than the other

departments when evaluated according to criteria like interest in the domain lessons, the success in the application courses, hand-eye-brain coordination, symbolic thinking skill, creativity, class harmony in the application courses, and participation in the social activities. Similar results were observed for the students of the fine arts high schools (Doğramacıoğlu, 2016). Milli and Yağcı (2017) indicated that the music department teacher candidates' communication skill was better than the students of the other departments. Similarly, Otacıoğlu (2007) found that the music department teacher candidates demonstrated a higher level of problem-solving skills than the GPC department teacher candidates. In contrast to these studies, a study in India on teacher candidates found a significant difference between science teacher candidates' life skills and art teacher candidates in favor of science teacher candidates (Pal & Chandra, 2019). Bhardwaj (2013) found that student teachers from the science stream had better composite life skills than the ones from the arts stream. The research results of Balaman et al. (2018), who compared the life skill levels of university students and pedagogical formation students, revealed that the life skill levels of the pedagogical formation students were significantly higher than the ones of the undergraduate students. Göksün and Kurt (2017) stated that the usage of 21st-century learning skills and the 21st-century teaching skills of the teacher candidates varied according to their universities and departments; and this might be caused by the department's HEIE admission threshold score & HEIE score type, the learning life of the teacher candidates in the universities, and other factors like different professors and course contents. Studying in different departments create differences in the life skills of the teacher candidates. This result indicates a need for longitudinal studies on the factors that may affect life skills, considering the attributes of both the departments and the teacher candidates who study there.

Since life skills have an impact on the prediction of many variables like success (Chien et al., 2012; Cronin et al., 2019; Erduran Avcı & Korur, 2019, June), metacognitive awareness (Zorlu et al., 2019), and self-efficacy (Koyuncu, 2018; Kozikoğlu & Altunova, 2018), it is vital to make students gain them from the early ages. One of the dominant factors in student's learning during the formal learning process is teachers. Therefore, it can be predicted that teacher candidates with highly developed life skills will contribute to the teaching-instruction process and the success of our students. Amutha and Ramganes (2013) emphasize that teachers should gain and develop the life skills to use them in their personal and professional life. Simona (2015) emphasizes the need for vocational teachers and trainers for practical training and support activities in embedding the life skills in their specialties. In this context, courses, activities, and applications regarding life skills can be inserted into the teacher training programs (Amutha & Ramganes, 2013; Pal & Chandra, 2019) and learning environments, that allow the candidates to integrate these skills into cognitive, affective, and psychomotor acquisitions, can be designed. This way, teacher candidates can attune the professional skills to daily life skills (Güneş & Uygun, 2016) and they can be supported in adopting these skills to the learning environments.

As with every research, there are several limitations for this study. The first limitation is related to the type of instrument used for the evaluation of life skills. We tried to limit the impact of this limitation by applying the steps in the development phase of the scale, providing the participants with adequate time and accompanying them during the data acquisition phase, and reminding the participants to read all items of the questions before making their markings. The second limitation is the fact that the instruments with closed-end questions rely on the honesty of the provided answers. Therefore, different measurement instruments may be merged in future studies that aim to evaluate young people's life skills (Jacobs Foundation, 2011). The third limitation is the varied distribution of the teacher candidates to the departments. Future studies can be conducted with relatively similar sample sizes according to the variables. This study's structure is not appropriate to reveal the cause-and-effect relations, which can be stated as the

last limitation. The longitudinal studies with different research designs may help determine the causality relations among the factors that impact life skills.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

### Authorship contribution statement

**Dilek Erduran Avcı:** Investigation, Literature review, Research design, Data collection, Data analysis, and Writing the manuscript. **Sümeyye Turgut:** Investigation, Literature review, Methodology, Data collection, and Data analysis. **Fikret Korur:** Investigation, Methodology, Data analysis, Visualization, Software, and Writing the manuscript.

### ORCID

Dilek Erduran Avcı  <https://orcid.org/0000-0001-6695-7348>

Sümeyye Turgut  <https://orcid.org/0000-0001-8620-1070>

Fikret Korur  <https://orcid.org/0000-0003-2676-6234>

## 5. REFERENCES

- Adkins, W.R. (1984). Life skills education: A video-based counselling/learning delivery system. In D. Larson (Ed.), *Teaching Psychological Skills: Models for giving psychology away* (pp.44-68). Brooks/Cole.
- Allen, P., & Bennett, K. (2008). *SPSS for the Health and Behavioural Sciences*. Thomson Learning.
- Allen, B.S., & Lohman, B.J. (2016). Positive youth development life skills gained at the Iowa 4-H Youth Conference. *Journal of Youth Development*, 11(1), 62-76. <https://www.semanticscholar.org/paper/Positive-Youth-Development-Life-Skills-Gained-at-Allen-Lohman/15a78d98314f59ad9966bfe44c2008b67db17204>
- Amutha, S., & Ramganes, E. (2013). Efficiency of teacher educators through life skill building. *International Journal of Management and Development Studies*, 2(3), 13-18. <http://www.indianjournals.com/ijor.aspx?target=ijor:ijmdsl&volume=2&issue=3&article=002>
- Bailey, S.J., & Deen, M.Y. (2002). Development of a web-based evaluation system: A tool for measuring life skills in youth and family programs. *Future Relations*, 51(2), 138-147. <https://www.jstor.org/stable/3700199?seq=1>
- Balaman, F., Bolat, Y., & Baş, M. (2018). A comparison of life skill levels between students of education faculty and students of pedagogical formation: The case of Mustafa Kemal University, Turkey. *European Journal of Education Studies*, 4(5), 75-91. <https://oapub.org/edu/index.php/ejes/article/view/1582/4214>
- Bentler, P. M., & Chou, C. P. (1987). Practical issues in structural modeling. *Sociological Methods & Research*, 16, 78–117. <https://doi.org/10.1177/0049124187016001004>
- Bhardwaj, S. (2013). *Teaching skills, study skills and life skills among student teachers in relation to their sex and stream of study at graduation level* [Unpublished Doctoral Dissertation]. Himachal Pradesh University.
- Bolat, Y., & Balaman, F. (2017). Life skills scale: validity and reliability study. *Journal of the Human and Social Science Researches*, 6(4), 22-39. <http://www.itobiad.com/tr/download/article-file/336258>
- Brown, T. A., (2012). *Confirmatory factor analysis for applied research*. The Guilford Press.

- Byrne, B. M. (2010). *Multivariate applications series*. Structural equation modeling with AMOS: Basic concepts, applications, and programming (2nd ed.). Routledge/Taylor & Francis Group.
- Can, S., & Can, Ş. (2011). Kamu personeli seçme sınavı öncesinde öğretmen adaylarının stress düzeyleri [Stress levels of students before KPSS (Public Personnel Choosing Exam)]. *Kastamonu Education Journal*, 19(3), 765-778. <https://dergipark.org.tr/tr/download/article-file/817384>
- Cassidy, K., Franco, Y., & Meo, E. (2018). Preparation for adulthood: A teacher inquiry study for facilitating life skills in secondary education in the United States. *Journal of Educational Issues*, 4(1), 33-46. <https://files.eric.ed.gov/fulltext/EJ1172806.pdf>
- Chauhan, S. (2016). Effectiveness of a life skills programme on teacher trainees. *International Multidisciplinary e-Journal*, 5(4), 90-98. <http://www.shreeprakashan.com/Documents/20160430142939588.13.Sarika%20Chauhan.pdf>
- Chien, N., Harbin, V., Goldhagen, S., Lippman, L., & Walker, K.E. (2012). Encouraging the development of key life skills in elementary school-age children: a literature review and recommendations to the tauck family foundation. *Child Trends*, 28, 1-11. <https://www.childtrends.org/wp-content/uploads/2013/06/2012-28KeyLifeSkills.pdf>
- Child, D. (1990). *The essentials of factor analysis*, (2nd ed.). Cassel Educational Limited.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Earlbaum Associates.
- Cronin, L.D., & Allen, J. (2017). Development and initial validation of the Life Skills Scale for Sport. *Psychology of Sport and Exercise*, 28, 105-119. <http://dx.doi.org/10.1016/j.psychsport.2016.11.001>
- Cronin, L., Allen, J., Ellison, P., Marchant, D., Levy, A., & Harwood, C. (2019). Development and initial validation of the life skills ability scale for higher education students. *Studies in Higher Education*. <https://doi.org/10.1080/03075079.2019.1672641>
- Çardak, M. (2013). Fiziksel gelişim [Physical development]. In Ş. Işık-Terzi (Ed.), *Eğitim psikolojisi* (pp. 48-71). Pegem Akademi.
- DanielSoper (n.d.) *Statistics calculators*. <https://www.danielsoper.com/statcalc/calculator.aspx?id=89>
- Doğramacıoğlu, B. (2016). *Comparison of creativity levels of students in fine arts high school and other high school types* [Unpublished Master Thesis]. Marmara University.
- Erawen, P. (2010). Developing life skills scale for high school students through mixed methods research. *European Journal of Scientific Research*, 47(2), 169-186. <https://pdfs.semantic scholar.org/f25a/954616d717592158165bea1142f36af8f416.pdf>
- Erduran Avci, D., & Kamer, D. (2018). Views of teachers regarding the life skills provided in science curriculum. *Eurasian Journal of Educational Research*, 77, 1-18. <https://files.eric.ed.gov/fulltext/EJ1192961.pdf>
- Erduran Avci, D., & Korur, F. (2019, June). *Bir ölçek geliştirme çalışması: Yaşam becerileri ölçeği* [A scale development study: Life skills scale]. Paper presented at VI<sup>th</sup> International Eurasian Educational Research Congress (EJER), Ankara University.
- Eryılmaz, A., & Mutlu, T. (2017). Career development and health from the perspective of life-span development approach. *Current Approaches in Psychiatry*, 9(2), 227-249. <https://dergipark.org.tr/en/download/article-file/268979>
- European Commission/EACEA/Eurydice (2012). *Developing key competences at school in Europe: Challenges and opportunities for policy*. Publications Office of the European Union. [https://eacea.ec.europa.eu/national-policies/eurydice/content/developing-key-competences-school-europe-challenges-and-opportunities-policy\\_en](https://eacea.ec.europa.eu/national-policies/eurydice/content/developing-key-competences-school-europe-challenges-and-opportunities-policy_en)



- Evin Gencil, İ. (2013). Prospective teachers' perceptions towards lifelong learning competencies. *Education and Science*, 38(170), 237-252. <http://egitimvebilim.ted.org.tr/index.php/EB/article/viewFile/1847/558>
- Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Sage Publications.
- Fitzpatrick, S., Twohig, M., & Morgan, M. (2014). Priorities for primary education? From subjects to life- skills and children's social and emotional development. *Irish Educational Studies*, 33(3), 269-286. <https://doi.org/10.1080/03323315.2014.923183>
- Fox, J., Schroeder, D., & Lodl, K. (2003). Life skill development through 4-H clubs: The perspective of 4-H alumni. *Journal of Extension*, 41(6). [https://www.joe.org/joe/2003december/rb2.php#:~:text=The%20mission%20of%204%2DH,people%20\(Cox%2C%201996\).&text=His%20research%20showed%20that%204,service%20ethic%2C%20and%20social%20skills.](https://www.joe.org/joe/2003december/rb2.php#:~:text=The%20mission%20of%204%2DH,people%20(Cox%2C%201996).&text=His%20research%20showed%20that%204,service%20ethic%2C%20and%20social%20skills.)
- Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2011). *How to design and evaluate research in education*. McGraw-Hill Humanities/Social Sciences/Languages.
- Garst, B.A., Gagnon, R., & Whittington, A. (2016). A closer look at the camp experience: Examining relationships between life skills, elements of positive youth development, and antecedents of change among camp alumni. *Journal of Outdoor Recreation, Education, and Leadership*, 8(2), 180-199. <https://doi.org/10.18666/JOREL-2016-V8-I2-7694>
- Ginter, E. J. (1999). David K. Brooks' contribution to the developmentally based life-skills approach. *Journal of Mental Health Counseling*, 21(3), 191-202. <https://search.proquest.com/docview/198716822?pq-origsite=gscholar&fromopenview=true>
- Göksun, D.O., & Kurt, A.A. (2017). The relationship between pre-service teachers' use of 21st century learner skills and 21st century teacher skills. *Education and Science*, 42(190), 107-130. <http://dx.doi.org/10.15390/EB.2017.7089>
- Green, H.A. (2008). *Learn from yesterday, live for today, hope for tomorrow: The development of a life skills scale* [Unpublished Master thesis]. Miami University.
- Güneş, F. (2016). Problems of divergence of skills in Turkish teaching and recommendations for solution. *Bartın University Journal of Faculty of Education*, 5(2), 205-222. <https://dergipark.org.tr/tr/download/article-file/224463>
- Güneş, F., & Uygun, T. (2016). Skill discrepancy in teacher education. *Ahi Evran Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 2(3), 1-14. <https://dergipark.org.tr/tr/download/article-file/263737>
- Hair, Jr., J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (6th ed.). Pearson Prentice Hall.
- Hu, L. T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1-55. <https://doi.org/10.1080/10705519909540118>
- IGI Global (n.d.). *What is pre-service teachers*. <https://www.igi-global.com/dictionary/investigating-the-factors-influencing-pre-service-teachers-acceptance-to-use-mobile-devices-for-learning/23201>
- Indian National Council of Educational Research and Training. (2005). *India National Curriculum*. <http://www.ncert.nic.in/rightside/links/pdf/framework/english/nf2005.pdf>
- Jacobs Foundation. (2011). *Monitoring and evaluating life skills for youth development* (Volume 2: The Toolkit). <https://jacobsfoundation.org/app/uploads/2017/08/Monitoring-and-Evaluating-The-Toolkit.pdf>
- Kadish, T.E., Glaser, B.A., Calhoun, G.B., & Ginter, E.J. (2001). Identifying the developmental strengths of juvenile offenders: Assessing four life-skills dimensions. *Journal of Addictions & Offender Counseling*, 21, 85-95. <https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2161-1874.2001.tb00154.x>



- Kaufman, K.J. (2013). 21 Ways to 21st century skills: Why students need them and ideas for practical implementation. *Kappa Delta Pi Record*, 49(2), 78-83. <https://doi.org/10.1080/00228958.2013.786594>
- Kayahan, Z., & Çakmakoğlu Kuru, A. (2017). Türkiye'deki görsel iletişim tasarımı bölümlerine öğrenci alımında uygulanan sınavlar hakkında öğretim elemanı görüşleri [Opinions of academicians regarding the examinations applied in student admission to departments of visual communication design in Turkey]. *Sanat ve Tasarım Dergisi*, 19, 75-87. <https://dergipark.org.tr/tr/download/article-file/317170>
- Khatoon, S. (2018). Developing life skills approach in the teaching-learning process based on Johari Window Model: Dealing with change. *The Research Journal of Social Sciences*, 9(6), 65-75. [www.aensi.in/](http://www.aensi.in/)
- Koyuncu, B. (2018). The effect of pre-service teachers' life skills on teacher self-efficacy. *Journal of Education and Learning*, 7(5), 188-200. <https://doi.org/10.5539/jel.v7n5p188>
- Kozikoğlu, İ., & Altunova, N. (2018). The predictive power of prospective teachers' self-efficacy perceptions of 21st century skills for their lifelong learning tendencies. *The Journal of Higher Education and Science*, 8(3), 522-531. [https://pdfs.semanticscholar.org/b414/4ab36c3e160fe6da12373e31672f50cbbb22.pdf?\\_ga=2.126820109.1836896345.1612253960-2067958555.1612253960](https://pdfs.semanticscholar.org/b414/4ab36c3e160fe6da12373e31672f50cbbb22.pdf?_ga=2.126820109.1836896345.1612253960-2067958555.1612253960)
- Kurtdede-Fidan, N., & Aydogdu, B. (2018). Life skills from the perspectives of classroom and science teachers. *International Journal of Progressive Education*, 14(1), 32-55. <https://files.eric.ed.gov/fulltext/EJ1169814.pdf>
- Milli, M. S., & Yağcı, U. (2017). Research on communication skills of pre-service teachers. *Bolu Abant İzzet Baysal University Journal of Education Faculty*, 17(1), 286-298. <https://dergipark.org.tr/tr/download/article-file/292008>
- Ministry of Education, Singapore. (2016). *Primary School Education Booklet*. <https://www.moe.gov.sg/education/primary/primary-schooleducation-booklet>
- O'Rourke, M., Hammond, S., O'Sullivan, D., Staunton, C., & O'Brien, S. (2016). The LifeMatters programme for developing life-skills in children: An evaluation. *International Journal of Mentoring and Coaching in Education*, 5(2), 144-157. <https://doi.org/10.1108/IJMCE-10-2015-0031>
- Otacıoğlu, A. S. G. (2007). Comparison of problem solving skill levels of students who are trained in different branches of education faculties. *Eurasian Journal of Educational Research*, 29, 73-83. <https://app.trdizin.gov.tr/makale/TnpBeE9UWTI/egitim-fakultelerinin-farkli-branslarinda-egitim-alan-ogrencilerin-problem-cozme-beceri-duzeylerinin-karsilastirilmesi>
- Pal, S., & Chandra, S. (2019). A study of life skills of pupil – teachers. *International Journal of Science and Research*, 8(8), 2201-2205. [https://www.ijsr.net/search\\_index\\_results\\_paperid.php?id=ART2020796](https://www.ijsr.net/search_index_results_paperid.php?id=ART2020796)
- Prasad, J. (2018). *Awareness of life skills among senior secondary school students of east and south districts of Sikkim* [Unpublished Master thesis]. Sikkim University.
- Sardoğan, M.E., & Ağaoğlu, S.A. (2005). The comparison between university students who enters the university with special ability examination and general ability examination in terms of the empathic abilities. *Gazi Journal of Physical Education and Sport Sciences*, 4, 33-40. <https://dergipark.org.tr/tr/download/article-file/292439>
- Schurer, S. (2017). Does education strengthen the life skills of adolescents? *IZA World of Labor*, 366, 1-11. <https://doi.org/10.15185/izawol.366>
- Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research*, 8(2), 23–74. <https://psycnet.apa.org/record/2003-08119-003>

- Simona, G. (2015). Teacher training for embedding life skills into vocational teaching. *Procedia-Social and Behavioral Sciences*, 180, 814-819. <https://doi.org/10.1016/j.sbspro.2015.02.215>
- Super, D.E. (1990). A life span, life-space approach to career development. In D Brown & L Brooks (Eds.), *Career choice and development* (pp.197-261). Jossey-Bass.
- Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics* (5th ed.). Pearson Education.
- Tan, S. (2018). *Life skills education: Teachers' perceptions in primary school classrooms in Finland and Singapore* [Unpublished Master thesis]. University of Jyväskylä.
- T.C. Kalkınma Bakanlığı. (2013). *Onuncu kalkınma planı* (2014-2018) [The tenth development plan]. Kalkınma Bakanlığı. <http://www.sbb.gov.tr/wp-content/uploads/2018/11/Onuncu-Kalk%C4%B1nma-Plan%C4%B1-2014-2018.pdf>
- The European Centre for the Development of Vocational Training (CEDEFOP) (2018). *Insights into skill shortages and skill mismatch*. <http://www.cedefop.europa.eu>
- The Turkish Ministry of National Education (TMNE) (2017). *Öğretmenlik mesleğinin genel yeterlikleri* [General competencies of the teaching profession]. Öğretmen Yetiştirme ve Geliştirme Genel Müdürlüğü. [http://oygm.meb.gov.tr/meb\\_iys\\_dosyalar/2017\\_12/1111\\_5355\\_YYRETMENLYK\\_MESLEYY\\_GENEL\\_YETERLYKLERY.pdf](http://oygm.meb.gov.tr/meb_iys_dosyalar/2017_12/1111_5355_YYRETMENLYK_MESLEYY_GENEL_YETERLYKLERY.pdf)
- The Turkish Ministry of National Education (TMNE) (2018). *Fen bilimleri dersi öğretim programı* [Science course curriculum]. T.C. Milli Eğitim Bakanlığı. <https://mufredat.meb.gov.tr/ProgramDetay.aspx?PID=325>
- The Turkish Ministry of National Education (TMNE) (2018b). *Hayat bilgisi dersi öğretim programı* [Life sciences course curriculum]. T.C. Milli Eğitim Bakanlığı. <https://mufredat.meb.gov.tr/ProgramDetay.aspx?PID=326>
- The Turkish Ministry of National Education (TMNE) (2018c). *Sosyal bilgiler dersi öğretim programı* [Social sciences course curriculum]. T.C. Milli Eğitim Bakanlığı. <https://mufredat.meb.gov.tr/ProgramDetay.aspx?PID=354>
- The Turkish Qualifications Framework (TQF) (2015). *Türkiye yeterlilikler çerçevesi* [Turkish Qualifications framework]. Ankara: Mesleki Yeterlilik Kurumu. <https://www.myk.gov.tr/index.php/tr/turkiye-yeterlilikler-ercevesi>
- Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. John Wiley & Sons.
- United Nations International Children's Emergency Fund (UNICEF) (2012). *Global evaluation of life skills education programmes: Final report*. UNICEF. [https://www.eccnetwork.net/sites/default/files/media/file/GLSEE\\_Booklet\\_Web.pdf](https://www.eccnetwork.net/sites/default/files/media/file/GLSEE_Booklet_Web.pdf)
- World Health Organization (WHO) (1997). *Life skills education for children and adolescents in schools*. Programme on mental health world health organization. [https://apps.who.int/iris/bitstream/handle/10665/63552/WHO\\_MNH\\_PSF\\_93.7A\\_Rev.2.pdf?sequence=1&isAllowed=y](https://apps.who.int/iris/bitstream/handle/10665/63552/WHO_MNH_PSF_93.7A_Rev.2.pdf?sequence=1&isAllowed=y)
- World Health Organization (WHO) (2004). *Skills for health: Skills-based health education including life skills*. Information series on school health, Document 9. The World Health Organization's information series on school health. <https://apps.who.int/iris/bitstream/handle/10665/42818/924159103X.pdf?sequence=1&isAllowed=y>
- Vranda, M.N. (2009). Development and standardization of life skills scale. *Indian Journal of Social Psychiatry*, 25(1-2), 17 - 28. [http://iasp.org.in/IJSP/IJSP\\_2009\\_\(1-2\).pdf#page=19](http://iasp.org.in/IJSP/IJSP_2009_(1-2).pdf#page=19)
- Zorlu, D.Y., Zorlu, F., & Dinç, S. (2019). Investigation of relationships between the preservice science teachers' life skills and metacognitive awareness. *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education*, 13(1), 302-327. <https://doi.org/10.17522/balikesirnef.511546>

## 6. APPENDIX

## Life Skills Scale

## Yaşam Becerileri Ölçeği

Sayın Katılımcı,

Bu ölçek yaşam becerilerini belirlemeye yönelik maddelerden oluşmaktadır. Sizden beklenen her maddeyi okuyup 1 ile 5 arası derecelerden birini işaretlemenizdir. Maddeleri içtenlikle işaretlemeniz araştırma sonuçları açısından oldukça önemlidir. Lütfen tüm maddeleri işaretleyiniz. Katkılarınızdan dolayı teşekkür ederiz.

**1: En az katılıyorum.....5: En çok katılıyorum**

	1	2	3	4	5
<b>Eleştirel Düşünme</b>					
1. Kanıtlar yanlışlığı gösterdiğinde, düşüncelerimi değiştiririm.					
2. Bir olayı çeşitli açılardan değerlendirebilirim.					
3. Bir olay sonucunda doğabilecek riskleri değerlendirebilirim.					
4. Fikirlerimi, gerçekler ve deneyimler ile oluştururum.					
5. Kendimi geliştirmek için yaptığım her hareketi eleştiririm.					
6. Nedenleri ve kanıtları temel alarak bir durumu anlamaya çalışırım.					
<b>Yaratıcı Düşünme</b>					
7. Başkalarından fikir ve öneri alırım, ancak onlara inanmadan önce kendim analiz ederim.					
8. Bir işi farklı tarzda/yenilikçi yapmaktan hoşlanırım.					
9. İşlerimi dikkatli yapmaya özen gösteririm.					
10. Yeni şeyler yapmayı tercih ederim.					
11. Yeni fikirler üretirim.					
12. Başkalarından farklı düşünceler üretebilirim.					
13. Sorunlar karşısında kendi yenilikçi çözümlerimi oluştururum.					
14. Herhangi bir işi yapmanın birçok yolunu bulabilirim.					
15. Kendi özgün fikirlerimin peşinden giderim.					
16. Problemlerimi çözerken genellikle hayal gücüme başvururum.					
<b>Karar verme ve problem çözme</b>					
17. Kararlarımın sonuçları hakkında sorumluluk alırım.					
18. Sorunun tüm çözümlerini değerlendirip en iyisini seçerim.					
19. Karar almadan önce sorunun tüm yönlerini analiz ederim.					
20. Verdiğim kararların sonuçlarını tahmin edebilirim.					
21. Ne pahasına olursa olsun bir sorunun çözümünü bulmaya çalışırım.					
22. Bir karara varmadan önce tüm bakış açıları dikkate alırım.					
23. Sorunlarımı çözerken ve önemli kararlar alırken deneyimlerimden yararlanırım.					
24. Kararlarım ya da çözümlerim işe yaramazsa tekrar gözden geçiririm.					
25. Karar almadan önce sonuçlardan nasıl etkileneceğimi düşünürüm.					
26. Karar almadan önce, başkalarını nasıl etkileyeceğini düşünürüm.					
27. Karar alırken önceliklerimi düzenleyebilirim.					
28. Bir problemi akıl yürüterek çözerim.					

<b>Stresle ve Duygularla Başa Çıkma</b>				
29. Stresle başa çıkmak için farklı yollar denerim.				
30. Olumsuz duygularımı çevremdeki insanlara yansıtmam.				
31. Olumsuz duygularla başa çıkabilirim.				
32. Stresi engelleyebilmek için bir plan dahilinde çalışabilirim.				
33. Stresi arttıracak mükemmeliyetçilik duygusundan vazgeçebilirim.				
34. Fikir çatışmalarımı başa çıkabilirim.				
35. Öfke ile baş edebilirim.				
36. Hayatımdaki herşey için olumlu düşünürüm.				
37. Durumlar karşısında kontrolsüz tepkiler vermem.				
38. Duygularımı uygun şekilde ifade ederim.				
39. Genellikle kaygı düzeyim düşüktür.				
<b>Kişiler arası ilişki ve iletişim</b>				
40. Amacıma uygun iletişim yöntemlerini seçmeye dikkat ederim.				
41. İletişim becerilerimi geliştirmek için çaba gösteririm.				
42. İnsanlarla kolayca iletişim kurabilirim.				
43. Konuşurken niyetimi çok açık bir şekilde ifade ederim.				
44. İnsanlarla konuşurken göz teması kurarım.				
45. Birisi konuşurken çok dikkatli dinlerim.				
46. İnsanlar benimle konuşurken rahat hisseder.				
<b>Empati</b>				
47. Başkalarının görüşlerini özgürce ifade etmelerine fırsat veririm.				
48. Kendimi karşımdaki bireyin yerine koyabilirim.				
49. Başkalarına yardım etmek için kendi sorumluluğumun farkındayım.				
50. Başkalarının hislerini anlayabilirim.				
51. Başkalarına yardım ettiğimde mutlu hissederim.				
52. Acı çeken birilerini gördüğümde kendimi kötü hissederim.				
53. Kimseyi incitmemeye çalışırım.				
<b>Öz Farkındalık</b>				
54. Sevdiğim şeyleri biliyorum.				
55. Duygularımın farkındayım.				
56. Kendi ihtiyaçlarımın farkındayım.				
57. Neleri başarabileceğimin farkındayım.				
58. Duygularımı uygun bir şekilde ifade edebilirim.				
59. Becerilerimi etkili bir şekilde kullanırım.				
60. Güçlü yönlerimi biliyorum.				
61. Sahip olduğum yetenekleri biliyorum.				
62. Yaptığım işleri/eylemleri değerlendiririm.				
63. İhtiyaçlarımı biliyorum.				
64. Hayatımın amaçları hakkında net bir fikrim var.				
65. Hak ve sorumluluklarımı biliyorum.				
<b>Öz Saygı</b>				
66. Birçok iyi özelliğe sahip olduğumu düşünüyorum.				
67. Kendi özelliklerimi seviyorum.				
68. Kendimi bütünüyle değerli hissediyorum.				

69. Birçok şeyi diğer insanlar kadar iyi yapabiliyorum.					
70. Birçok şeyi yapabileceğime inanıyorum.					
71. Hayatı değerli olarak görüyorum.					
72. Sahip olduklarımdan memnunum.					
73. Yaptığım işlerde kendime güveniyorum.					
<b>Takım Çalışması</b>					
74. Kendimden başka birinin yaptığı işe güvenmem.					
75. Takım çalışmalarında sorumluluk almaktan çekinirim.					
76. Takım çalışmalarında benden farklı düşünenlere tahammül edemem.					
77. Takım çalışmalarında “Her koyun kendi bacağından asılır.” düşüncesini taşıyım.					
78. Takımla çalışma ortamında kendi isteklerimi yaparım.					
<b>Sosyal Sorumluluk</b>					
79. Çevremi kirlettiğimde kendimi suçlu hissederim.					
80. Topluma faydalı işlerde gönüllü olmak isterim.					
81. Bencil davrandığımda kendimi suçlu hissederim.					
82. Birlikte çalıştığım grup başarısız olduğunda suçlu hissederim.					
83. Davranışlarımdan ötürü başkaları sorun yaşarsa kendimi kötü hissederim.					

## Automated Essay Scoring Effect on Test Equating Errors in Mixed-format Test

Ibrahim Uysal <sup>1,\*</sup>, Nuri Dogan <sup>2</sup>

<sup>1</sup>Department of Educational Sciences, Faculty of Education, Bolu Abant İzzet Baysal University, Bolu, Turkey

<sup>2</sup>Department of Educational Sciences, Faculty of Education, Hacettepe University, Ankara, Turkey

### ARTICLE HISTORY

Received: Oct. 24, 2020

Revised: Dec. 31, 2020

Accepted: Feb. 07, 2021

### Keywords:

Test equating,  
Automated scoring,  
Classical test theory,  
Item response theory,  
Mixed-format tests.

**Abstract:** Scoring constructed-response items can be highly difficult, time-consuming, and costly in practice. Improvements in computer technology have enabled automated scoring of constructed-response items. However, the application of automated scoring without an investigation of test equating can lead to serious problems. The goal of this study was to score the constructed-response items in mixed-format tests automatically with different test/training data rates and to investigate the indirect effect of these scores on test equating compared with human raters. Bidirectional long-short term memory (BLSTM) was selected as the automated scoring method for the best performance. During the test equating process, methods based on classical test theory and item response theory were utilized. In most of the equating methods, errors of the equating resulting from automated scoring were close to the errors occurring in equating processes conducted by human raters. It was concluded that automated scoring can be applied because it is convenient in terms of equating.

## 1. INTRODUCTION

Test developers often have a dilemma in choosing the item format to be included on the tests. Reasons for this include suitability for the measurement of cognitive features, cost of application and scoring, the effect of item types used in tests on teaching, and psychometric properties. With practicality in mind, tests can be designed to include only multiple-choice items, only constructed-response items, or both multiple-choice and constructed-response items (Martinez, 1999; Rodriguez, 2002). Martinez (1999) states that a single-format test is not suitable for all purposes and situations, while Messick (1993) states that using different test item formats together will benefit from the strengths of each format and compensate for weaknesses. Therefore, it is essential to use both multiple-choice and constructed-response items, especially in large-scale tests. Because with constructed-response items, students have opportunity to organize and apply what they learn in a deeper way (Tankersley, 2007). However, it is difficult, time-consuming, and costly to score constructed-response items in large-scale testing applications. Due to the scoring difficulties of constructed-response items, test developers searched for and introduced the concept of automated scoring (Page, 1966).

**CONTACT:** İbrahim UYSAL ✉ [ibrahimuysal06@gmail.com](mailto:ibrahimuysal06@gmail.com) 📍 Department of Educational Sciences, Faculty of Education, Bolu Abant İzzet Baysal University, Bolu, Turkey

ISSN-e: 2148-7456 /© IJATE 2021



Using automated essay scoring systems in tests will ensure efficient use of funds, reduce scoring time, and efforts (Attali & Burstein, 2006; Chen et al., 2014). The use of this system will eliminate the necessity to use many raters. Besides, scoring bias can be prevented. Reliability problems arising from differently trained raters will be overcome, as will generalizability (Adesiji et al., 2016). However, the effectiveness of automated scoring systems in applications such as test equating, which is important in ensuring justice between individuals taking different test forms or participating in the test at different times, has not been adequately investigated in the literature. Applying automated scoring without such research can cause serious problems (such as making wrong decisions about individuals). When automated scoring conditions change, equating error is also likely to change. In this respect, it is necessary to determine the acceptable automated scoring limits for test equating. The current study was designed based on these problem situations.

This study is important in determining whether automated scoring and training/test data rates in automated scoring increase test equating errors and whether the equating errors that occur because of automated scoring are different from the equating errors that occur with human raters. Thus, test equating after automated scoring can be performed under relevant conditions. When the literature was examined, a test equating study that Almond (2014) conducted on constructed-response items by automatically scoring common items in a sample of 500 people was found. In this study, the linear logistic equating method, a variant of Tucker linear equating, was used. Also, there was only one test equating study using automated scoring in mixed-format tests. This study, conducted by Olgar (2015), contains 30 multiple-choice items and one open-ended item in tests. The studies carried out by Almond (2014) and Olgar (2015) used the linear logistics equating method. The current study focused on equating tests with a large number of constructed-response items with automated scoring.

Moreover, this study was not based on a single test equating method but was carried out using both classical test theory (CTT) and item response theory (IRT) based test equating methods. It was seen that test equating methods based on IRT were not used in test equating studies carried out with automated scoring. So, to investigate which method works better in equating with automated scoring, both CTT and IRT were used in the study.

In the literature, similar studies compared the equating methods based on CTT and IRT in mixed-format tests and between nonequivalent groups using a common item pattern (Hagge & Kolen, 2011; Hagge et al., 2011; He, 2011; Lee et al., 2012; Liu & Kolen, 2011; Wolf, 2013). In the current study, CTT-based equating methods (Tucker linear, chained linear, chained equipercentile, frequency equipercentile), and IRT-based true score equating methods (mean-mean, mean-sigma, Stocking-Lord and Haebara) were used. Most of the literature studies (Hagge & Kolen, 2011; Hagge et al., 2011; He, 2011; Liu & Kolen, 2011; Wolf, 2013) compared CTT-based chained equipercentile and frequency estimation methods and IRT-based true and observed score equating methods. Among these studies, Hagge and Kolen (2011) and Hagge et al. (2011) used the Haebara method, Wolf (2013) used simultaneous scaling and He (2011) and Liu and Kolen (2011) used the Stocking-Lord method in IRT-based true score equating. In their research, Lee et al. (2012) compared Tucker, Levine observed score, Levine true score, chained equipercentile, frequency estimation, Stocking-Lord, and IRT observed score equating methods.

In the current study, in cases where equipercentile equating, based on CTT, was used, pre-smoothing with the bivariate log-linear function was applied. Similar to this study, Hagge et al. (2011), Lee et al. (2012), and Wolf (2013) pre-smoothed with the log-linear function. On the other hand, Liu and Kolen (2011) used pre-smoothing while obtaining the results for the population to make a comparison in the equating process. In addition, they changed synthetic population ratios of equating methods other than chained equating methods. Similarly, Hagge

and Kolen (2011), Hagge et al. (2011), and Wolf (2013) changed the synthetic population ratio to 1 in their study. However, these studies did not evaluate the effect of the synthetic population ratio but showed the results based on the new group that took the test. While Hagge and Kolen (2011) and Liu and Kolen (2011) conducted their research on real data, Wolf (2013) worked on simulated data. Of these researchers, Liu and Kolen (2011) included only multiple-choice items in tests as common items, while Hagge and Kolen (2011) and Wolf (2013) used mixed-format tests as common items in tests.

More constructed-response items should be included in large-scale tests to measuring more complex skills such as higher-order, critical thinking and reasoning, better evaluating items involving multiple steps in the solution process. But these items should also be easily and accurately scored. Therefore, the current study is important. In addition, test equating studies on restricted constructed-response items with automated scores are not enough. This study has two purposes: i) to evaluate the effect of constructed-response items scored by automated scoring systems in the test equating process on equating errors, ii) to examine the change of equating errors in the change of the conditions in the automated scoring systems.

## **2. METHOD**

### **2.1. Design**

The study was correlational, as it aims to determine the effect of automated scoring of constructed-response items on test equating in mixed-format tests by comparing it with test equating performed by human raters. Creswell (2012) stated that it is possible to see how a difference in one variable affects the other variable in correlational studies.

### **2.2. Sample**

The data for this study were obtained from the eighth-grade Turkish test that is part of the Academic Skills Monitoring and Evaluation (ABİDE) project implemented by the Ministry of National Education (MoNE) in 2016. Data for 1000 students who answered the A<sub>1</sub> and B<sub>1</sub> booklets on the Turkish test were selected randomly. After selecting and cleaning data, 607 students from the A<sub>1</sub> booklet and 584 students from B<sub>1</sub> booklet were studied. Details were given in the data analysis section. Spence (1996) stated that at least 500 individuals must answer each test form for test equating studies. The number of students answering the A<sub>1</sub> and B<sub>1</sub> booklets in this research met this criterion.

### **2.3. Data Collection Tools**

Multiple-choice and constructed-response items are included in ABİDE tests, which aim to examine students' higher-order thinking skills using different types of items. Two human rater groups scored Constructed-response items, and a third rater group was consulted in case of a dispute between the first two raters' groups. The focus of the research was the data obtained from two Turkish test forms (A<sub>1</sub> and B<sub>1</sub>) with 18 items. 9 items in the A<sub>1</sub> test and 10 items in the B<sub>1</sub> test were constructed-response items. Constructed-response items were scored as either 0-1 or 0-1-2. Nine items were common in A<sub>1</sub> and B<sub>1</sub> tests (MoNE, 2017).

Since the tests used in the study contain common items, they were equated using the common-item nonequivalent group (CINEG) design. However, some criteria must be met to equate the tests using a CINEG design. Angoff (1984) stated that even if the test length increases, the proportion of common items in the test should not be less than 20%. In this application, the proportion of common items was 50%. Considering the data characteristics, it is necessary to use dichotomously and polytomously scored item types together in common items in tests. As a matter of fact, Tate (2000) proposed the use of both types of items as common items in mixed-format tests. The reason for this is that the common items should represent the entire test. In the A<sub>1</sub> and B<sub>1</sub> booklets, five of the nine common items were constructed-response and four were multiple-choice.

Cramer's  $V$  coefficient calculated the consistency between raters for each constructed-response item included in the tests in the ABIDE study. Cramer's  $V$  ranged from .83 to .98 for items included in the Turkish test in A<sub>1</sub> booklet, and from .87 to .99 for items included in the Turkish test in B<sub>1</sub> booklet. Internal consistency coefficients for test scores were stated as .73 for booklet A and .76 for booklet B (MoNE, 2017).

## 2.4. Data Analysis

The data were entered based on the balanced distribution of the categories regarding the scores obtained from the constructed-response items. This was done to avoid the problem of prevalence regarding constructed-response items in the data. Indeed, this is important in automated scoring. Taking into account 9 items for A<sub>1</sub> booklet and 10 items for B<sub>1</sub> booklet 697 data entries from A<sub>1</sub> booklet and 701 data entries from booklet B<sub>1</sub> were made. Then, within the researchers' criteria, students responding to half or more of the constructed-response items and multiple-choice items in the test were selected. After this process, the missing data rates were calculated for each constructed-response and multiple-choice item. The data were cleaned so that the missing data rate remained below 5%. It was anticipated that a large number of blank answers will show higher interrater reliability coefficients in automated scoring. As there were few data in some categories, individuals scoring in these categories were retained in the response data as much as possible. Then, the scores given by the two groups of human raters (group 1 and group 2) were examined. Due to the missing data, a group of students were also excluded from the study. In the last case, 90 students using the A<sub>1</sub> booklet and 117 students using the B<sub>1</sub> booklet were excluded. Thus, the data preparation process was completed, and the automated scoring process was started with 607 data from the A<sub>1</sub> booklet and 584 from the B<sub>1</sub> booklet.

In the study, an automated scoring system was created using the Python program on the Linux operating system. Automated scoring was done using supervised machine learning algorithms by mapping the computer's scoring features through human raters. Five methods were used in automated scoring: SVM (support vector machine), LR (logistic regression), MNB (multinomial naive Bayes), LSTM (long-short term memory), and BLSTM (bidirectional long-short term memory). Two libraries were used in the software prepared through Python. 90% of the data was used to train the system and 10% to test the system. Random sampling method was applied with cross validity. Ten-fold cross-validation was used. Turkish test constructed-response items belonging to "Monitoring, Research and Development Project for Measurement and Evaluation Applications" implemented by MoNE were used while developing the software. This test is different from the ABIDE tests used in this research. It is given to fifth-grade students (10–11 years old) and includes five constructed-response items. Five constructed-response items were used while preparing the software. Three of the five constructed-response items are scored as 0-1, while two are scored as 0-1-2. Two human rater groups scored each student's answer, and a third rater group was applied in case of dispute. Rubrics were used in scoring processes. Table 1 shows the sample results of 0-1 scored item 16 and 0-1-2 scored item 20. While 0-1 scored item 16 was tested with 303 data, 0-1-2 scored item 20 was tested with 637 data. Since item 20 was scored in three categories, it was found appropriate to experiment on more data.

**Table 1.** *Agreement percentages between automated and human scoring.*

	Number of data	Number of category	SVM (%)	LR (%)	MNB (%)	LSTM (%)	BLSTM (%)
Item 16	303	2	98.0	98.3	96.1	99.0	99.0
Item 20	637	3	85.5	82.4	75.1	87.3	88.7

Note: Agreement percentages above 80% indicate an acceptable fit (Hartmann, 1977).

Table 1 shows that the percentages of agreement obtained for item 16 were relatively high. The methods that showed the highest agreement percentage for this item were LSTM and BLSTM. Therefore, the agreement percentages obtained for item 20 are sufficient. The method that showed the best agreement in item 20 was the BLSTM method. The fact that the percentages of agreement obtained for all methods were at the expected level showed that the system created would be sufficient to score the current study's constructed-response items.

The entry of the student answer sheets in JPEG format for constructed-response items was done manually. This is because students' handwriting was difficult to read and because optical character recognition (OCR) systems cannot be used on account of the use of adjacent handwriting. In addition, it was to eliminate errors that may arise from OCR programs. In order to completely match the manually entered data with student answers, the data were checked by a team of six people and errors were corrected. Student answers were directly conveyed and were not subject to any correction.

The automated scoring system was trained in the automated scoring phase using the human raters' final scores. In this way, it was taught how to score by human raters and the scoring features were mapped to the system. Test data, which were not used in the training of the system, were scored automatically. The amount of data used to test the system was a factor studied in the research. The data rates used to test the system were determined as 10%, 20% and 33%. Therefore, the amount of data used in training the system was 90%, 80% and 67% respectively. These values indicated that 61, 121 and 200 of the 607 data for the A<sub>1</sub> booklet were used to test the system, respectively, while 546, 486 and 407 data, respectively, were used to train the system. From the B<sub>1</sub> booklet, 584 data, 58, 117 and 193 are used to test the system, respectively; 526, 467 and 391, respectively, were used to train the system. The amount of data to be used for training the system was reduced as much as possible, and the effect of this on automated scoring and indirect effect on test equating examined. While calculating the results, 10-fold cross-validation was used for the 10% test data rate, 5-fold cross validity was used for the 20% test data rate, and 3-fold cross validity was used for the 33% test data rate. In this way, training and test data were differentiated and all data from both booklets were converted into test data. As a result, the system obtained 607 data scored for the A<sub>1</sub> booklet and 584 scored for the B<sub>1</sub> booklet.

Automated scoring was performed for 10%, 20% and 33% test data rates using the BLSTM method, which shows the best fit, and equating was started. In order to make comparisons, the test forms were equated by using the final scores of the human raters for each test form. In the equating process, methods based on CTT and IRT were used. The test data's statistics and reliability values to this research were examined before the equating process. The statistics and reliability coefficients of the A<sub>1</sub> and B<sub>1</sub> booklet for human raters and automated scoring (BLSTM 10%, BLSTM 20% and BLSTM 33%) are given in Table 2. The reliability coefficient was examined in two ways. In the first case, reliability was determined by Cronbach's alpha coefficient (Cronbach, 1951) and in the second case by McDonald's omega coefficient (McDonald, 1999) based on factor analysis. While the alpha coefficient was used because it gave the lower bound estimate of reliability, the omega coefficient was chosen because it had less and more realistic assumptions (Bendermacher, 2010; Dunn et al., 2014).

Table 2 shows that the average score generated by human rating was slightly lower than the average score calculated after automated scoring. When using human raters, the standard deviation was slightly higher than automated scoring. Omega and Cronbach's alpha reliability coefficients were found to be close to each other under both human rating and automated scoring. However, when using human raters, both Cronbach's alpha and omega coefficients were slightly higher.

**Table 2.** Test statistics on  $A_1$  and  $B_1$  booklets.

	Human Raters		BLSTM %10		BLSTM %20		BLSTM %33	
	$A_1$	$B_1$	$A_1$	$B_1$	$A_1$	$B_1$	$A_1$	$B_1$
Number of Item	18	18	18	18	18	18	18	18
Sample Size	607	584	607	584	607	584	607	584
Mean	13.152	14.101	13.259	14.300	13.283	14.361	13.273	14.346
Standart Deviation	4.530	4.964	4.331	4.777	4.333	4.765	4.313	4.760
Median	14	15	13	15	14	15	14	15
Minimum	1	0	2	0	2	0	2	1
Maximum	23	23	23	23	23	23	23	23
Skewness	-.249	-.466	-.208	-.520	-.218	-.538	-.209	-.518
Reliability (Alfa)	.766	.797	.746	.784	.746	.783	.747	.786
Reliability (Omega)	.868	.893	.857	.885	.856	.882	.858	.884

Chained linear (LC), Tucker linear (LT), chained equipercentile (EC), and frequency estimation (EF) equating methods based on CTT were chosen. Synthetic population value was changed to  $w_1 = 1$  ( $WS = 1$ ) and the effect of this situation was investigated. When the synthetic population was determined as  $w_1 = 1$ , the group that takes the new test form in the common item design in nonequivalent groups was determined as the synthetic universe (Kolen, & Brennan, 2014). When the synthetic population value was not changed, the synthetic population was determined according to the number of samples in the groups (to be  $w_1 + w_2 = 1$ ). However, since chained equating did not support the synthetic population, synthetic population ratios had not been changed in methods using chained equating (Kolen, & Brennan, 2014). In addition, presmoothing (PSM) was performed for equipercentile equating methods. For the EF method, PSM is performed and the synthetic population ratio was changed. With these changes, the effects of synthetic population parameters and/or PSM on the equating results were also evaluated. "equate" (Albano, 2016) package in R (R Development Core Team, 2018) was used while equating test forms according to CTT methods. PSM was carried out using PROC IML (Moses & von Davier, 2006) code in SAS 9.4 (SAS Institute, 2015). The reason for performing this procedure outside the R program was that the total scores obtained from the  $A_1$  booklet or the  $B_1$  booklet and the total scores obtained from the common tests should be subtracted because some of the frequencies associated with the score combinations were zero (Moses et al., 2004). However, the "equate" package in the R software did not allow this.

PSM was performed using polynomial bivariate loglinear function distribution due to the use of nonequivalent group design. The best model was chosen for each form by comparing 11 different models in the polynomial bivariate loglinear function distribution. The equating was carried out by using 10000 replications with the bootstrap technique.

The mean-mean (MM), mean-sigma (MS), Haebara (HB) and Stocking-Lord (SL), which are true score equating methods based on separate calibration in IRT, were used. Before equating, IRT assumptions were examined. The first assumption was unidimensionality. Factor analysis for mixed tests for each test form was carried out for both human scorers and automated scoring conditions using the MPLUS (Muthén & Muthén, 2012) program. Due to the use of mixed-



format tests, polycoric and tetracoric correlations were utilized. The weighted least square mean and variance adjusted (WLSMV) were used as the estimation method in the factor analysis. WLSMV estimation method is known as one of the most suitable methods when using polycoric and tetracoric correlations (Barendse et al., 2015). In addition, parallel analysis (Timmerman & Lorenzo-Seva, 2011) was carried out through the Factor 10.5 program (Lorenzo-Seva & Fernando, 2006) in order to decide the number of dimensions. Parallel analysis results showed that each test form has a single factor structure for both automated scoring (with 10%, 20% and 33% test data rates) and human raters.

Five models were compared to determine which IRT model fit the data for each test form. Since there were re-constructed-response items rated binary and there was no possibility to respond to these items by chance, all binary items were examined based on one parameter model (1PLM) and two-parameter model (2PLM). Models reviewed include 1) 1PLM and partial credit model (PCM), 2) 1PLM and generalized partial credit model (GPCM), 3) 1PLM and graded response model (GRM), 4) 2PLM and GPCM, 5) 2PLM and GRM. When comparing models, the differences between  $-2\log$  likelihood values and degrees of freedom were calculated, and these values were compared with the chi-square table. If the value obtained was greater than the value determined for the 5% error in the chi-square table, a higher model had been adopted. When comparing models with the same degrees of freedom, standard error averages related to theta estimation were used. EAP method was used to estimate ability parameters. Accordingly, models with lower standard errors were used to estimate the ability and item parameters. Model comparisons were made for all of the human raters' final scores and the rating done by the automated scoring systems and it was concluded that the 2PLM and GPCM methods were more appropriate overall. Ability and item parameters were estimated using XCalibre 4.1 (Yoes, 1996). The XCalibre program estimates the discrimination and difficulty parameters with a lower error (RMSE) than BILOG (Mislevy & Bock, 1997; Weiss & Minden, 2012). Test equating was performed by transferring the ability parameters and item parameters estimated in the XCalibre program to the IRTEQ program.

Standard error of equating (SEE), bias (BIAS), and root mean squared error (RMSE) were calculated to be used in comparisons after test equating with methods based on CTT and IRT. The random error (SEE) was designed based on the standard deviation of the equated scores and results from the sample. Bias, that is, systematic error, was based on the difference between the estimated equation and the criterion (real) equation relationship. Bias results from reasons such as the common items do not represent the test form in terms of content and statistical properties in nonequivalent groups, the serious differences between the groups and the difference of common items from one application to another. Bias was not a coefficient directly affected by the sample. RMSE is a combination of bias and standard error (Kolen & Brennan, 2014; LaFlair et al., 2017). The bias value was not directly used in comparing the performance of the methods due to the high level of negative and positive values can neutralize each other (Zu & Liu, 2010). Absolute BIAS values have not been studied since the negative BIAS value indicates that the skills are predicted to be lower than they are and the positive indicates that the skills are predicted higher than they are (Pang et al., 2010). The methods were compared over SEE and RMSE, which is a combination of SEE and BIAS. While choosing the best method, RMSE values were used due to the combination of systematic and random error.

SEE, BIAS, and RMSE values were calculated through the “equate” package (Albano, 2016) after the equating process in CTT and the MSEXCEL module after the IRT equating process. By choosing the same error coefficients, CTT and IRT equating methods were compared. To make it easier to compare with the CTT, theta was used to calculate the IRT errors. Below are the equations used to calculate BIAS (equation 1), RMSE (equation 2) and SEE (equation 3) in the CTT (Gonzalez & Wiberg, 2017).  $L$  is the number of bootstraps performed,  $l$  are the



samples,  $\hat{\varphi}(x_i)$  is the estimated equated scores,  $\varphi(x_i)$  is the real equated scores, and  $\bar{\varphi}(x_i)$  is the estimated equated mean scores:

$$BIAS(x_i) = \frac{1}{L} \sum_{l=1}^L [(\hat{\varphi}_1(x_i) - \varphi_1(x_i))] \tag{1}$$

$$RMSE(x_i) = \sqrt{\frac{1}{L} \sum_{l=1}^L [(\hat{\varphi}_1(x_i) - \varphi_1(x_i))^2]} \tag{2}$$

$$SEE(x_i) = \sqrt{RMSE(x_i)^2 - BIAS(x_i)^2} \tag{3}$$

The following equations can be used when calculating SEE (equality 4), BIAS (equality 5) and RMSE (equality 6) values based on IRT. The resources of Deng and Monfils (2017) and Keller and Keller (2011) were used for equations.  $\theta_i$  is the ability of the individual  $i$ ,  $\hat{\theta}_i$  is the ability of the individual  $i$  estimated by the equating method used, and  $N$  is the sample size:

$$SEE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{\theta}_i - \theta_i - BIAS)^2} \tag{4}$$

$$BIAS = \frac{1}{N} \sum_{i=1}^N (\hat{\theta}_i - \theta_i) \tag{5}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{\theta}_i - \theta_i)^2} \tag{6}$$

After the equating errors were obtained for three automated scoring conditions, they were compared with the human raters. It was then decided to perform a difference test to determine the status of showing significant difference in the errors (RMSE) of the rater type in the equating process. Accordingly, the average of three conditions related to automated scoring was calculated. Normality was then tested for each group. A Shapiro-Wilks test was used while testing normality. The results showed that the RMSE values of the equating process performed through human raters did not distributed normally ( $W(sd = 13) = .860, p < .05$ ), and the RMSE values of the equating process performed through automated scoring system were normally distributed ( $W(sd = 13) = .914, p = .210 > .05$ ). As a result, since one of the groups did not provide the assumption of normality, the difference test was carried out with the Mann-Whitney U test, a nonparametric technique. To determine the effect of the scoring type on the RMSE, the effect size was calculated through Cliff’s Delta coefficient (Cliff, 1996). The Cliff’s Delta coefficient used to compare two groups ranges from -1 to +1. If the coefficient is closer to -1 or +1 the effect size is increased and if closer to 0 effect size is decreased (Cliff, 1993). For this purpose, R “effsize” package (Torchiano, 2020) was used.

After calculating the effect size, the correlation between the errors of the human raters’ equating and the errors of the automated scoring equating were examined. According to the normality tests, the relationship was examined using Spearman’s rho correlation since one of the variables did not meet the normality assumption.

### 3. RESULT / FINDINGS

Table 3 shows the errors related to the test equating process. Equating was made with human scores for both forms and equating errors displayed in the “human” column. Equating using machine scores was performed for both forms and equating errors are shown in the “BLSTM” column. Table 3 shows the equating errors using the scores obtained with 10%, 20%, 33% test data rates via the BLSTM method. In Table 3, the lowest error methods are shown in bold and the highest error are shown in italics for each rater and type of error condition.

When the human raters were re taken into consideration in [Table 3](#), that the lowest random error (SEE) was .050 obtained in the MS method based on IRT. MM method followed this with .061. When using methods based on IRT, the highest SEE (.083) showed in SL and HB methods. When using human raters, the method that showed the lowest SEE (.197) in CTT based equating methods was the Tucker linear in which the synthetic population ratio was determined as 1 (LT[WS = 1]). This value was followed by (.198) the LT equating method in which the random universe ratio was not changed and the random universe ratio was determined based on the sample numbers. The method with the highest SEE (.357) was the PSMEC equating method, which was pre-smoothed with a bivariate logarithmic linear function. In the case where human raters were used, the highest SEEs were obtained in equipercentile equating methods. In this condition, methods based on IRT generally showed lower SEEs than methods based on CTT.

When test equating results made after automated scoring performed with a 10% test data rate and the BLSTM method were evaluated in terms of random error, the lowest random error (.047) was found in MS method. This value (.047) was lower than that of human raters (.050). This value (.047), which was obtained at the 10% test data rate, was followed by the MM method with .079. When using methods based on IRT, HB method showed the highest SEE (.110). When automated scoring was performed at a rate of 10% test data, LT[WS = 1] was the method that shows the lowest SEE (.200) in test equating methods based on CTT. This value was followed by the LT equating method with .201. The method with the highest SEE (.407) is the EC. In the equating performed after automated scoring with the 10% test data rate and BLSTM method, the highest SEEs were obtained in equipercentile equating methods. In this condition, methods based on IRT generally showed less SEEs than methods based on CTT. The SEEs calculated for all methods were close to the SEEs of equating with human raters. In two conditions, automated scoring (using BLSTM method with 10% test data rate) led to test equating with fewer errors.

When test equating results made after automated scoring performed with a 20% test data rate were evaluated in terms of random error, the lowest random error (.006) was found in the MS method. The value obtained was quite close to 0 (.006) and was much lower than the SEE (.050) obtained when human raters are used. This value (.006), which was obtained at the 20% test data rate, was followed by the MM method with .098. When using methods based on IRT, HB method showed the highest SEE (.127). When automated scoring was performed at a rate of 20% test data, LT[WS = 1] was the method that shows the lowest SEE (.196) in equating methods based on CTT. This value is followed by the LT equating method with .197. The method with the highest SEE (.405) was the PSMEC equating method. In the equating performed after automated scoring with the 20% test data rate and BLSTM method, the highest SEEs were obtained in equipercentile equating methods in general. In this condition, methods based on IRT generally showed lower SEEs than methods based on CTT. The SEEs calculated for all methods are close to the SEEs of equating with human raters. In four conditions, automated scoring (using BLSTM method with 20% test data rate) led to test equating with fewer errors.

When test equating results made after automated scoring performed with a 33% test data rate were evaluated in terms of random error, the lowest random error (.012) was found in the MS method. This value obtained is quite close to 0 (.012) and is much lower than the SEE (.050) obtained when human raters were used. This value (.012), which was obtained at the 33% test data rate, was followed by the MM method with .071. When using methods based on IRT, the HB method showed the highest SEE (.137). When automated scoring was performed at a rate of 33% test data, LT[WS = 1] was the method that shows the lowest SEE (.200) in test equating methods based on CTT. This value was followed by the LT equating method, with an SEE of .202. The method with the highest SEE (.398) is the EC equating method.

**Table 3.** Errors related to equating methods based on CTT and IRT.

		SEE				BIAS				RMSE			
		Human	BLSTM			Human	BLSTM			Human	BLSTM		
			%10	%20	%33		%10	%20	%33		%10	%20	%33
CTT	LC	.211	.213	.209	.215	<b>.003</b>	<b>.002</b>	<b>.002</b>	<b>.003</b>	.211	.213	.209	.215
	LT	.198	.201	.197	.202	<b>.003</b>	<b>.002</b>	<b>.002</b>	<b>.003</b>	.198	.201	.197	.202
	LT (WS=1)	.197	.200	.196	.200	<b>.003</b>	<b>.002</b>	<b>.002</b>	.004	.197	.200	.196	.200
	EC	.351	<i>.407</i>	<i>.396</i>	<i>.398</i>	.061	<i>.216</i>	<i>.159</i>	<i>.142</i>	<i>.357</i>	<i>.461</i>	<i>.427</i>	<i>.423</i>
	EF	.330	.336	.347	.336	.062	.032	.052	.071	.336	.337	.351	.344
	EF (WS=1)	.330	.362	.371	.348	.059	.048	.158	.062	.335	.365	.403	.353
	PSMEC	<i>.357</i>	.328	<i>.405</i>	.350	.044	.042	.087	.041	<i>.359</i>	.331	.414	.352
	PSMEF	.321	.341	.360	.307	.023	.021	.084	.021	.322	.342	.369	.307
	PSMEF (WS=1)	.333	.349	.371	.317	.023	.021	.078	.021	.334	.349	.379	.318
IRT	MM	.061	.079	.098	.071	-.010	.022	.039	.010	<b>.062</b>	<b>.083</b>	<b>.106</b>	<b>.072</b>
	MS	<b>.050</b>	<b>.047</b>	<b>.006</b>	<b>.012</b>	.064	.128	.127	.079	.081	.136	.127	.080
	HB	.083	.110	.127	.137	<i>-.079</i>	-1.08	-0.087	-1.127	.114	.154	.154	.187
	SL	.083	.100	.118	.119	<i>-.079</i>	-0.098	-0.078	-1.118	.114	.140	.141	.167

Note: In terms of SEE, BIAS and RMSE, the lowest coefficient is shown in bold and the highest coefficient in italics in each condition.

In the equating performed after automated scoring with the 33% test data rate and BLSTM method, the highest SEEs were obtained in equipercentile equating methods. In this condition, methods based on IRT generally showed lower SEEs than methods based on CTT. The SEEs calculated for all methods were close to the SEEs of equating with human raters. In four conditions, automated scoring (using BLSTM method with 33% test data rate) made test equating with fewer errors.

When the random errors obtained in all equating processes were evaluated, the errors were very close to each other. In the equating performed by automated scoring, in some cases, lower SEE values were obtained than in the equating performed by human raters. IRT based methods had lower SEE values than methods based on CTT, even if human raters were used or automated scoring was performed. Considering all the equating processes, the lowest SEE value (.006) was obtained using the MS method with BLSTM in automated scoring based on a 20% test data rate. The highest SEE value (.407) was obtained by the EC equating method in all test equating processes performed using BLSTM in automated scoring based on a 10% test data rate.

Systematic error (BIAS) sizes obtained in the equating process with human raters vary between .003 and .079. BIAS values obtained after equating with scores obtained through the BLSTM method based on a 10% test data rate vary between .002 and .216. BIAS values obtained after equating with scores obtained through the BLSTM method based on a 20% test data rate vary between .002 and .159. BIAS values obtained after equating with scores obtained through the BLSTM method based on a 33% test data rate vary between .003 and .142.

When the human raters were taken into consideration, as shown in [Table 3](#), the lowest RMSE was .062 obtained by the MM method based on IRT. This value was followed by .081 with the MS method. When using IRT methods, the highest RMSE (.114) was found in the SL and HB methods. These results mean that moment methods (MM and MS) show lower RMSEs than characteristic curve methods (SL and HB) based on IRT. When using human raters, the method that shows the lowest RMSE (.197) in CTT based equating methods is the LT[WS = 1]. This value is followed by .198 with the LT equating method. The method with the highest RMSE (.359) was the PSMEC equating method. In the case where human raters are used, the highest RMSEs were obtained in equipercentile equating methods. In this condition, methods based on IRT generally showed less RMSEs than methods based on CTT.

When test equating results made after automated scoring performed with a 10% test data rate were evaluated in terms of RMSE, the lowest RMSE (.083) was found in the MM method. This value (.083) was close to the lowest RMSE value (.062) obtained when human raters are used. This value (.083), which was obtained at the 10% test data rate was followed by MS method with .136. When using methods based on IRT, HB method showed the highest RMSE (.154). When automated scoring was performed at a rate of 10% test data, LT[WS = 1] was the method that shows the lowest RMSE (.200) in test equating methods based on CTT. This value was followed by the LT equating method with .201. The method with the highest RMSE (.461) was the EC equating method. In the equating performed after automated scoring with the 10% test data rate and BLSTM method, the highest RMSEs were obtained in equipercentile equating methods in general. In this condition, methods based on IRT generally showed less RMSEs than methods based on CTT. The RMSEs calculated for all methods were close to the RMSEs calculated from equating with human raters. In one condition (PSMEC), automated scoring (using BLSTM method with 10% test data rate) led to test equating with fewer RMSE.

When test equating results conducted after automated scoring performed with a 20% test data rate and BLSTM were evaluated in terms of RMSE, the lowest RMSE (.106) was found in the MM method. This value (.106) was close to the lowest RMSE value (.062) obtained when human raters were used. This value (.106), which was obtained at the 20% test data rate, was followed by the MS method with .127. When using methods based on IRT, HB method showed

the highest RMSE (.154). When automated scoring was performed at a rate of 20% test data, LT[WS = 1] was the method that shows the lowest RMSE (.196) in equating methods based on CTT. This value was followed by the LT equating method with .197. The method with the highest RMSE (.427) was the EC equating method. In the equating performed after automated scoring with the 20% test data rate and BLSTM method, the highest RMSEs were obtained with equipercentile equating methods. In this condition, methods based on IRT generally showed lower RMSEs than methods based on CTT. The RMSEs calculated for all methods are close to the RMSEs calculated by equating with human raters. In three conditions, automated scoring (using BLSTM method with 20% test data rate) performed test equating with fewer RMSEs.

When test equating results made after automated scoring performed with a 33% test data rate are evaluated in terms of RMSE, the lowest RMSE (.072) was found in the MM method. This value (.072) was very close to the lowest RMSE value (.062) obtained by human raters. This value (.072), which was obtained at the 33% test data rate, was followed by the MS method with .080. When using methods based on IRT, the HB method showed the highest RMSE (.187). When automated scoring was performed at a rate of 33% test data, LT[WS = 1] shows the lowest RMSE (.200) in equating methods based on CTT. This value was followed by the LT equating method with .202. The method with the highest RMSE (.423) was the EC equating method. In the equating performed after automated scoring with the 33% test data rate and BLSTM method, the highest RMSEs were obtained with equipercentile equating methods. In this condition, methods based on IRT generally showed lower RMSEs than methods based on CTT. The RMSEs calculated for all methods are close to the RMSEs of equating with human raters. In four conditions, automated scoring (using BLSTM method with 20% test data rate) performed test equating with fewer RMSEs.

Figure 1 shows RMSE values of the equating performed by human raters and automated scoring based on 10%, 20% and 33% test data rates. The chart was drawn in the range of 0 to 1, since in the literature it was noted that RMSE values below 1% are not important (Pang et al., 2010).

Figure 1. RMSE values of the methods according to the rater type.

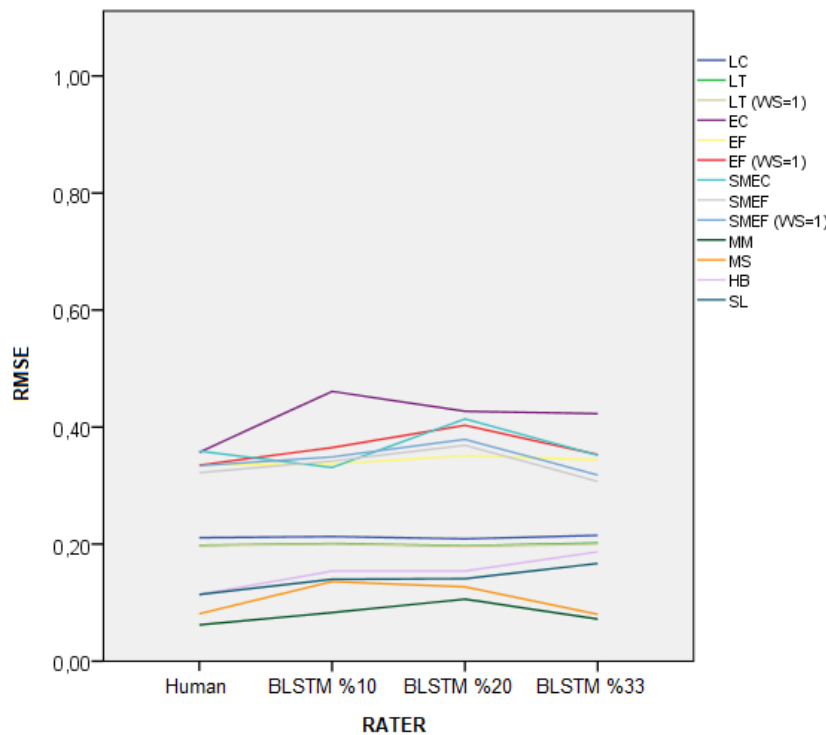




Figure 1 shows that the RMSE values obtained with all equating methods are close to each other. In the equating performed with automated scoring, in some cases, lower RMSE values were obtained than in the equating performed with human raters. IRT based methods had lower RMSE values than methods based on CTT, even if human raters were used or automated scoring was performed. Considering all the equating processes, the lowest RMSE value (.062) was obtained in MM method with the using human raters. In equating with automated scoring scores, the lowest RMSE value (.072) was obtained with the MM method. When IRT test equating methods were compared for each condition, it can be indicated that moment methods showed less error (RMSE) than characteristic curve methods. The highest RMSE value (.359) was obtained in the PSMEC equating method in all test equating process performed using human raters. In automated scoring, the highest RMSE value (.461) was obtained with the EC equating method. In general, equipercentile equating methods equate tests with more RMSE. Changing the synthetic population ratio to 1 generally reduced RMSE values in linear methods. However, in equipercentile equating methods and when pre-smoothing was applied in equipercentile equating methods RMSE values generally increased. Changing the ratio of synthetic population to 1 did not create very large decreases or increases in RMSE coefficients. The pre-smoothing process decreased RMSE values in some cases but increased it in other cases.

The average of errors resulting from test equating performed with the scores obtained by automated scoring with the test data rates of 10%, 20% and 33% were calculated. Then, the significant difference between these averages and the errors of the equating obtained through human raters was examined. Equating methods, variations in synthetic population ratios and/or pre-smoothing versions of these methods have been investigated to determine whether there is a difference between human raters and automated scoring averages. A Mann-Whitney U test was used because the normal distribution assumption was not met for each group. The results are shown in Table 4.

**Table 4.** Difference test regarding RMSE values obtained as a result of human raters and automated scoring.

	Rater	N	Mean Rank	Sum of Ranks	U	p
RMSE	Human Scoring	13	12.000	156.000	65.000	.336
	Automated Scoring	13	15.000	195.000		

Table 4 shows that the RMSE values (median = .211) of 13 equating methods obtained through human raters did not differ significantly from the mean RMSE values (median = .212) of 13 equating methods obtained through automated scoring ( $U = 65,000$ ,  $p = .336 > .05$ ). Accordingly, the use of human raters or automated scoring did not have a significant effect on the RMSE values obtained as a result of the equating process. The effect size was investigated through the Cliff's Delta coefficient and -.18 was found. This effect size is small (Cliff, 1993). The relationship between the errors of the equating (RMSE) performed by human raters and the averages of the equating errors (RMSE) performed by automated scoring was evaluated with the correlation of Spearman rank differences and at a high and significant level relationship was found ( $r = .96$ ,  $p = .00 < .05$ ).

#### 4. DISCUSSION and CONCLUSION

Three equating procedures were performed in the study according to the test data rates used in automated scoring. The equating process was carried out for human scorers as well as for automated scoring. In the equating process for human raters, the final scores of the human raters for the A<sub>1</sub> and B<sub>1</sub> booklets were used. In the equating process for automated scoring, the scores



obtained by the automated scoring of the constructed-response items in both test forms were used. Constructed-response items and objectively scored items are not subjected to equating separately. Methods based on CTT and IRT have been used as the equating method.

This study had found that the errors (RMSE) obtained in all methods and different combinations of methods in automated scoring conditions and in the condition where human raters were found similar. In some cases, lower RMSE values were found in the equating performed through automated scoring than human raters' equating processes. It was observed that pre-smoothing decreased RMSE values in some cases but increased in other cases. Hagge et al. (2011) determined that the pre-smoothing reduced the standard error of chained equipercentile equating and frequency estimation methods. This study changed the ratio of synthetic population decreased RMSE values in linear equating methods, while it increased RMSE values in equipercentile equating methods. However, it should be noted that equating errors presented here were based on automated scoring conditions. The result of the equating showed that methods based on IRT equate tests with lower errors (in terms of SEE and RMSE) compared to methods based on CTT either in automated scoring conditions or when human raters were used. Hagge and Kolen (2011) and Liu and Kolen (2011) stated that methods based on IRT showed lower errors than the methods based on CTT according to the root mean squared error in conditions like this study. Liu and Kolen (2011) also found that IRT true score equating methods had lower SEE values than frequency estimation and chained equipercentile equating methods. Although the same criterion is not considered, Lee et al. (2012) stated that IRT true score equating performed better than Tucker linear, chained equipercentile, frequency estimation, pre-smoothed chained equipercentile, and pre-smoothed frequency estimation methods in terms of primary level equality. Wolf (2013) also found that in terms of primary level equality, IRT true score equating performed better than frequency estimation and chained equipercentile equating. Hagge et al. (2011) stated that IRT based methods had lower SEE values than CTT based methods. However, these studies weren't equating based on automated scoring. When methods based on IRT were compared for each condition, moment methods equate with less error than characteristic curve methods. This situation may be related to linearity besides the number of common items and test length. The highest RMSE and SEE values are found in equipercentile equating methods.

Regarding RMSE and SEE, the highest errors were obtained in the chained equipercentile and pre-smoothed chained equipercentile equating methods. Hagge and Kolen (2011) and Hagge et al. (2011) also stated that the method with the highest SEE value was chained equipercentile equating. However, He (2011) stated that the chained equipercentile equating method performed better than frequency estimation method according to primary level equality criterion. The difference between this study and He (2011) is thought to be due to the sample size. In automated scoring, the average RMSE values of different test data rates for each equating method were calculated and the statistical differences of these values from the errors of equating performed by human raters were examined. As a result, it was determined that there was no significant difference between the errors and that the errors showed a high level of compliance. Olgar (2015) used the open-ended items as common items by scoring them automatically and stated that even though the common items were multiple-choice items or open-ended items scored automatically with multiple-choice items, the results were similar. He even found that the including automatically scored open-ended items in common items yielded better results in some cases. Almond (2014) stated that in tests consisting only of constructed-response items, linear logistic equating can be used as an alternative by automatically scoring common items with generic e-rater.

In cases where automated scoring is made, based on the results of this study, methods based on IRT in equating procedures are recommended. This study was carried out on approximately 1200 people. In subsequent studies, the effect of automated scoring on the equating process can

be examined using larger samples. This study determined the effect of changing the synthetic population ratio on equating errors under automated scoring conditions. In future studies, when there is a difference between the number of groups to be equated, the effect of the synthetic population ratio to .5 can be evaluated. This study also discussed the effect of pre-smoothing under automated scoring conditions. In further research, pre- and post-smoothing can be compared, and different pre- and post-smoothing methods can be examined under different patterns.

### Acknowledgments

This paper was produced from part of the first author's doctoral dissertation prepared under the supervision of the second author. Thanks to Behzad Naderalvojud for his support in the creation of the automated essay scoring software used in this research.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

### Authorship contribution statement

**Ibrahim UYSAL:** Investigation, Software, Methodology, Formal Analysis, Visualization, Resources, and Writing the original draft. **Nuri DOĞAN:** Investigation, Software, Methodology, Supervision, and Validation.

### ORCID

Ibrahim UYSAL  <https://orcid.org/0000-0002-6767-0362>

Nuri DOĞAN  <https://orcid.org/0000-0001-6274-2016>

## 5. REFERENCES


- Adesiji, K. M., Agbonifo, O. C., Adesuyi, A. T., & Olabode, O. (2016). Development of an automated descriptive text-based scoring system. *British Journal of Mathematics & Computer Science*, 19(4), 1-14. <https://doi.org/10.9734/BJMCS/2016/27558>
- Albano, A. D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software*, 74(8), 1-36. <https://doi.org/10.18637/jss.v074.i08>
- Almond, R. G. (2014). Using automated essay scores as an anchor when equating constructed-response writing tests. *International Journal of Testing*, 14(1), 73-91. <https://doi.org/10.1080/15305058.2013.816309>
- Angoff, W. H. (1984). *Scales, norms and equivalent scores*. Educational Testing Service.
- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v.2. *Journal of Technology, Learning, and Assessment*, 4(3), 1-30. <http://www.jtla.org>.
- Barendse, M. T., Oort, F. J., & Timmerman, M. E. (2015). Using exploratory factor analysis to determine the dimensionality of discrete responses. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(1), 87-101. <https://doi.org/10.1080/10705511.2014.934850>
- Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology*, 3(2), 77-85. <https://doi.org/10.1111/j.2044-8317.1950.tb00285.x>
- Chen, H., Xu, J., & He, B. (2014). Automated essay scoring by capturing relative writing quality. *The Computer Journal*, 57(9), 1318-1330. <https://doi.org/10.1093/comjnl/bxt117>
- Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3), 494-509. <https://doi.org/10.1037/0033-2909.114.3.494>
- Cliff, N. (1996). *Ordinal methods for behavioral data analysis*. Routledge.

- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *The Journal of Applied Psychology*, 78(1), 98-104. <https://doi.org/10.1037/0021-9010.78.1.98>
- Creswell, J. W. (2012). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (4th ed.). Pearson.
- Deng, W., & Monfils, R. (2017). *Long-term impact of valid case criterion on capturing population-level growth under Item Response Theory equating* (Research Report 17-17). Educational Testing Service. <https://doi.org/10.1002/ets2.12144>
- Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, 141(1), 2-18. <https://doi.org/10.1037/a0024338>
- Gonzalez, J., & Wiberg, M. (2017). *Applying test equating methods: Using R*. Springer.
- Hagge, S. L., & Kolen, M. J. (2011). Equating mixed-format tests with format representative and non-representative common items. In M. J. Kolen & W-C. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 1, pp. 95-135). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
- Hagge, S. L., Liu, C., He, Y., Powers, S. J., Wang, W., & Kolen, M. J. (2011). A comparison of IRT and traditional equipercentile methods in mixed-format equating. In M. J. Kolen & W-C. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 1, pp. 19-50). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
- He, Y. (2011). Evaluating equating properties for mixed-format tests [Doctoral dissertation, University of IOWA]. <https://ir.uiowa.edu/etd/981/>
- Kaiser, H. F. (1970). A second-generation little jiffy. *Psychometrika*, 35(4), 401-415. <https://doi.org/10.1007/BF02291817>
- Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark IV. *Educational and Psychological Measurement*, 34(1), 111-117. <https://doi.org/10.1177/001316447403400115>
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling and linking* (2nd ed.). Springer.
- Keller, L. A., & Keller, R. R. (2011). The long-term sustainability of different Item Response Theory scaling methods. *Educational and Psychological Measurement*, 71(2), 362-379. <https://doi.org/10.1177/0013164410375111>
- LaFlair, G. T., Isbell, D., May, L. D. N., Arvizu, M. N. G., & Jamieson, J. (2017). Equating in small-scale language testing programs. *Language Testing*, 34(1), 127-144. <https://doi.org/10.1177/0265532215620825>
- Lee, E., Lee, W-C., & Brennan, R. L. (2012). *Exploring equity properties in equating using AP® examinations* (Report No. 2012-4). CollegeBoard.
- Liu, C., & Kolen, M. J. (2011). A comparison among IRT equating methods and traditional equating methods for mixed-format tests. In M. J. Kolen & W-C. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 1, pp. 75-94). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
- Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods*, 38(1), 88-91. <https://doi.org/10.3758/BF03192753>
- Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, 34(4), 207-218. [https://doi.org/10.1207/s15326985ep3404\\_2](https://doi.org/10.1207/s15326985ep3404_2)
- Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed-response, performance testing, and portfolio assessment* (pp. 61-73). Lawrence Erlbaum Associates, Inc.

- MoNE. (2017). Monitoring and evaluation of academic skills (ABİDE) 2016 8th grade report. [https://odsgm.meb.gov.tr/meb\\_iys\\_dosyalar/2017\\_11/30114819\\_iY-web-v6.pdf](https://odsgm.meb.gov.tr/meb_iys_dosyalar/2017_11/30114819_iY-web-v6.pdf)
- Moses, T. P., & von Davier, A. A. (2006). *An SAS macro for loglinear smoothing: Applications and implications* (Report No. 06-05). Educational Testing Service.
- Moses, T., von Davier, A. A., & Casabianca, J. (2004). *Loglinear smoothing: An alternative numerical approach using SAS* (Research No. 04-27). Educational Testing Service.
- Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Muthén & Muthén.
- Olgar, S. (2015). *The integration of automated essay scoring systems into the equating process for mixed-format tests* [Doctoral dissertation, Florida State University]. <http://diginole.lib.fsu.edu/islandora/object/fsu%3A253122>
- Page, E. B. (1966). The imminence of grading essays by computers. *Phi Delta Kappan*, 47(5), 238–243. <http://www.jstor.org/stable/20371545>
- Pang, X., Madera, E., Radwan, N., & Zhang, S. (2010). *A comparison of four test equating methods* (Research Report). Eqao.
- R Development Core Team. (2018). *R: A language and environment for statistical computing* (version 3.5.2) [Computer software]. R Foundation for Statistical Computing.
- Rodriguez, M. C. (2002). Choosing an item format. In G. Tindal & T. M. Haladyna (Eds.), *Large scale assessment programs for all students: Validity, technical adequacy and implementation* (pp. 213-231). Lawrence Erlbaum Associates.
- SAS Institute. (2015). *Statistical analysis software* (version 9.4) [Computer software]. SAS Institute.
- Spence, P. D. (1996). *The effect of multidimensionality on unidimensional equating with item response theory* [Doctoral dissertation, University of Florida]. <https://www.proquest.com/docview/304315473>
- Tankersley, K. (2007). *Tests that teach: Using standardized tests to improve instruction*. Association for Supervision and Curriculum Development.
- Tate, R. (2000). Performance of a proposed method for the linking of mixed-format tests with constructed-response and multiple-choice items. *Journal of Educational Measurement*, 37(4), 329-346. <http://www.jstor.org/stable/1435244>
- Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, 16(2), 209-220. <https://doi.org/10.1037/a0023353>
- Torchiano, M. (2020). *effsize: Efficient Effect Size Computation* (Version 0.8.1) [Computer software]. <https://CRAN.R-project.org/package=effsize>
- Weiss, D. J., & Minden, S. V. (2012). *A comparison of item parameter estimates from Xcalibre 4.1 and Bilog-MG* (Technical Report). Assessment Systems Corporation.
- Wolf, R. (2013). *Assessing the impact of characteristics of the test, common-items, and examinees on the preservation of equity properties in mixed-format test equating* [Doctoral dissertation, University of Pittsburgh]. <https://core.ac.uk/download/pdf/19441049.pdf>
- Yoes, M. E. (1996). *User's manual for the XCALIBRE marginal maximum-likelihood estimation program* [Computer software]. Assessment Systems Corporation.
- Zu, J., & Liu, J. (2010). Observed score equating using discrete and passage-based anchor items. *Journal of Educational Measurement*, 47(4), 395-412. <https://doi.org/10.1111/j.1745-3984.2010.00120.x>



## An Investigation of Item Position Effects by Means of IRT-Based Differential Item Functioning Methods

Sumeyra Soysal <sup>1,\*</sup>, Esin Yilmaz Kogar <sup>2</sup>

<sup>1</sup>Necmettin Erbakan University, Faculty of Education, Department of Educational Sciences, Konya, Turkey

<sup>2</sup>Niğde Ömer Halisdemir University, Faculty of Education, Department of Educational Sciences, Niğde, Turkey

### ARTICLE HISTORY

Received: Aug. 13, 2020

Revised: Jan. 08, 2021

Accepted: Feb. 11, 2021

### Keywords:

Item position effects,  
Item Response Theory,  
Differential item function,  
Raju's unsigned area,  
Lord's chi-square.

**Abstract:** In this study, whether item position effects lead to DIF in the condition where different test booklets are used was investigated. To do this the methods of Lord's chi-square and Raju's unsigned area with the 3PL model under with and without item purification were used. When the performance of the methods was compared, it was revealed that generally, the method of Lord's chi-square identified more items with DIF than did the method of Raju's unsigned area. The differentiation of the booklets with respect to item position resulted in a higher number of items displaying DIF with item purification conditions. Based on the findings of the present study, to avoid the occurrence of DIF due to item position effects, it is recommended to position the same items across different booklets in similar locations when forming different booklets.

## 1. INTRODUCTION

With the help of measurement tools used in the field of education, various decisions such as passed/failed, successful/unsuccessful were intended to reach about individuals and it is aimed to affect individuals' lives as accurately as possible. Various methods are used in large-scale assessments in education in line with this aim. To make the results of these kinds of assessments more reliable, one of the widely used methods in different positions or locations within the tests (Bulut et al., 2017). Thus, problems such as individuals memorizing items or copying answers of other examinees during the test application can be overcome (Bulut, 2015). Thus, the effect of these factors that may affect the psychometric properties of the test can be reduced. However, although the use of different test forms or booklets has positive aspects, it may lead to psychometric issues such as position effects of items (Bulut, 2015). The consequences of the position effect on individuals' abilities are ignored in many test creation processes. If such an effect occurs, it is assumed to be the same for all persons and all items therefore it is thought to not affect the person's ability or item difficulty (Hahne, 2008). However, in practice, individuals' test scores can vary according to item position (Kleinke, 1980). In that case, item position effects that cause changes in individuals' test scores may threaten the validity of test score interpretations (Trendtel & Robitzsch, 2018). Hence, examining the positioning of the

CONTACT: Sümeýra Soysal ✉ [sumeyrasoysal@hotmail.com](mailto:sumeyrasoysal@hotmail.com) 📍 Necmettin Erbakan University, Faculty of Education, Department of Educational Sciences, 42090, Konya, Turkey

same items in various ways across different booklets should be examined and investigated to see whether or not one book type is more advantageous for some groups of test takers which is important for the test development process. The positions of items in booklets or test forms created by item position manipulations may lead to differential item functioning (DIF) (Akayleh, 2018; Balta & Omur Sunbul, 2017; Debeer & Janssen, 2013; Erdem, 2015). The present examines whether item position effects lead to DIF in test items or not.

### **1.1. Item Position Effects**

The interaction between the position of a test item in a test booklet and the performance a test taker displays on the same item is called item position effects – IP effects (Qian, 2014). Kingston and Dorans (1984) stated that, in the most classical way, IP effects may emerge in two conditions; namely, items in a measurement instrument that are positioned towards the end may be found easy by test takers owing to practice or learning effect (a positive IP effect) or they can be found difficult owing to fatigue effect (a negative IP effect).

An item displaying IP effects means that the item parameters (e.g., difficulty or discrimination) can vary according to the item's position in the booklet (Weirich et al., 2017). For example, Weirich et al. (2017) stated that considering IP effects on item difficulty, an item administered at the end of a test often is more difficult than the same item administered at the beginning of the test (p.115). Similarly, Le (2017) concluded that items tend to be more difficult when placed towards the end of the test. The test-takers in this study may have found the items positioned towards the end difficult owing to their decrease in motivation in the exam. However, whatever the underlying reason is, conditions that occur owing to IP effects negatively impact the validity of the results. Various studies have also indicated that it is important to consider position effect to test the validity of an assessment (Hahne, 2008; Hohensinn et al., 2008; Qian, 2014).

Studies in the literature investigated whether creating different test forms, arranging the location of the items in the test, and ordering the items from easy to hard or hard to easy affect the individuals' performance or item parameters. However, the results of the studies that examined this subject are not the same. While some studies have determined that the item position has a role on individuals' performance (Debeer & Janssen, 2013; Hartig & Buchholz, 2012; Ollenu & Etsey, 2015; The West African Examinations Council [WAEC], 1993), others have concluded that item position does not affect the performance of students or examinees (Doğan Gül & Çokluk Bökeoğlu, 2018; Perlini et al., 1988; Tal et al., 2008). In some studies, it was determined that the item position caused bias in item parameter estimates (Debeer & Janssen, 2013; Doğan Gül & Çokluk Bökeoğlu, 2018; Hecht et al., 2015; Meyers et al., 2009). Although there is no clear conclusion about the item position on which different studies have been conducted, different booklets are used in many exams for example the Program for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), and the Progress in International Reading Literacy Study (PIRLS). For the item security in such large-scale assessments (such as memorizing the item by those taking the exam), booklets created with items in different orders and different clusters could be used (Frey, Hartig, & Rupp, 2009). In such test administrations where there is awareness of the possibility of IP effects leading to negative outcomes (such as bias in item parameters, test score differences), booklet design is used as a measure. However, studies are reporting that IP has an impact even in administrations where booklet design is used as a measure (Hartig & Buchholz, 2012; Le, 2007; Martin et al., 2004).

Although the studies on the IP effects are mostly based on Classical Test Theory (CTT), there are also studies conducted with Item Response Theory (IRT) framework, the use of which has become widespread in many fields (Debeer & Janssen, 2013; Hahne, 2008; Hohensinn et al., 2008; Qian, 2014; Weirich et al., 2014). The fundamental assumptions of IRT are that the individual's ability measures can be obtained independently of the tests applied to test takers



and that invariant item and ability parameters can be reached (Hambleton et al., 1991). However, this assumption of item parameter invariance could be in the booklets in which the same items are positioned differently in an achievement test (Weirich et al., 2017).

Since IP effects are not the same for every test-takers, ignoring this effect limits to make a fair comparison. Recent research shows that there can be individual differences as a result of IP effects (Debeer & Janssen, 2013; Verguts & De Boeck; 2000). So, this situation may lead to biased ability parameter estimates. Moreover, IP effects can cause a different source of variation which can have an impact on test scores (Tippets & Benson, 1989). For this reason, the IP effects can cause significant validity issues.

IP effects have a crucial role in almost all moderate to extensive lengths tests using different booklets (Leary & Dorans, 1985). And IP effects is a practical concern in the professional development of test instruments in large-scale assessments (Qian, 2014). Therefore, it is highly worthwhile for test developers to focus and to attention on this issue.

## 1.2. Differential Item Functioning

Differential item functioning (DIF) developed by Holland and Thayer (1988) compares the probability of correct answers to items in test takers from different subgroups with the same level of ability. DIF occurs when different groups of the same underlying ability have different probabilities of responding to an item correctly (Holland & Wainer, 1993).

In DIF studies, it is common that there are at least two groups, i.e. focus and reference groups. The focal group generally refers to a minority group or study group, while the majority group is called the reference group (Schmitt & Crone, 1991). However, when naming the groups is not clear, it can be completely random. There are two types of DIF, namely uniform and non-uniform DIF. Uniform DIF exists when an item is constantly in favor of one group over another group across the  $\theta$  continuum (Zumbo, 1999). In other words, almost all members of a group show better performance than almost all the members of the group who are at the same ability levels. Non-uniform DIF occurs when the item provides a relative advantage, the magnitude of which changes as the  $\theta$  level changes, or when a group has a relative advantage at the low  $\theta$  level, whereas the other group has a relative advantage at the high  $\theta$  level (Penfield & Lam, 2000). If an item shows DIF, it does not mean that item is biased. Generally, DIF analysis is considered as the first step in deciding whether an item can be biased towards a particular group. If the factor causing DIF is irrelevant to the construct being measured by the test, it is a source of bias (Karami, 2012). Kamata and Vaughn (2004, p.51) stated that DIF can arise for reasons other than bias, and therefore an item with DIF should be interpreted as "possibly biased item" or simply called "DIF item".

McNamara and Roever (2006, p. 93) have discussed the DIF detecting methods in four categories: (1) Analyses based on item difficulty. These approaches compare item difficulty estimates. (2) Nonparametric approaches. These procedures use contingency tables, chi-square, and odds ratios. (3) Item-response-theory-based approaches which include 1, 2, and 3 parameter logistic models. (4) Other approaches. These include logistic regression, which also employs a model comparison method, as well as generalizability theory and multifaceted measurement, which are less commonly used in classic DIF studies. As IRT methods were employed in the present study, only these methods were focused on. Methods based on IRT essentially compare item parameters or item characteristic curves that show the focus and reference group test-takers' probability of giving correct answers to items (Camilli & Shepard, 1994). The chi-square test and Raju's area measurement, which are used in the present study, are among the most frequently used IRT-based DIF methods.

### **1.3. Differential Item Functioning Based on Position Effects**

There are numerous studies on IP effects on psychometric item characteristics in the related literature (Hambleton, 1968; Hambleton & Traub, 1974; Kelnke, 1980; Klosner & Gellman, 1973; Leary & Dorans, 1985; Lee, 2007; Newman et al., 1988; Perlina et al., 1998). However, there are fewer studies on whether using different forms or booklets in achievement exams leads to certain psychometric problems such as DIF, and in the majority of these studies, while some focus on item order effects by ordering items from easy to difficult, difficult to easy, or randomly based on item difficulty index (Balta & Omur Sunbul, 2017; Çokluk et al., 2016; Freedle & Kostin, 1991; Plake et al., 1988; Ryan & Chiu, 2001), others focus on IP effects (Avcu et al., 2018; Bulut, 2015; Erdem, 2015).

Ryan and Chiu (2001) developed two forms consisting of 40-items which included topics they had addressed, namely algebra, trigonometry, geometry, and analytic geometry. The items in form-1 were ordered from easy to difficult, while the items in form-2 were ordered from easy to difficult based on the topics. This study reported that the variance in item order did not significantly affect the occurrence of DIF. Çokluk, Gül, and Doğan-Gül (2016) administered three different forms in which the items of a 20-item achievement exam in a science and technology course were ordered from easy to difficult, from difficult to easy, and completely randomly to the seventh-grade students. They investigated whether there was DIF in different forms created by positioning items differently via CTT and IRT-based methods. They concluded that positioning items differently caused a significant difference in the probability of the test takers at the same ability level responding correctly to the items.

Another study, conducted by Bulut (2015), aimed to examine the relationship between gender-based DIF and booklet effect stemming from using test booklets in which the same items were used but positioned differently. By using large-scale verbal reasoning test data in the study, Bulut (2015) conducted uniform and nonuniform DIF analyses using CTT-based DIF detection methods. The study revealed that even though the general difficulty level of the booklets for the male and female groups was found to be similar, some items in each test booklet were observed to be marked as showing uniform and non-uniform DIF. In this study, where the number of non-uniform DIF items was found to be higher than the number of uniform DIF items in each type of booklet. It was deduced that different test booklets were problematic in terms of the exam results of male and female test-takers. In another study, conducted by Erdem (2015), whether the subtests of six different courses in the TEOG (Transition System from Elementary Education to Secondary Education) administered during the fall term of the 2014-2015 academic year displayed DIF based on booklet type was examined using CTT based DIF detection methods. The study revealed that, in terms of the test booklet, there was a high number of DIF displaying items in the subtests of Religion, Culture and Ethics, Turkish Revolution History and Kemalism, and Foreign Language (English), while the number of DIF displaying items decreased in subtests of Turkish and Science and Technology. There was no item displaying DIF in the mathematics subtest.

Findings reported by previous studies show that the location and order of items in a test can affect test results. Hence, it can be claimed that the position of test items should be taken into consideration during a test development process. Thus, the present study aimed to examine whether or not IP effects led to DIF arising from using different test booklets. In large-scale assessments in Turkey are not usually administered as a pilot test. Therefore, items cannot be placed in these booklets based on item difficulty indices.

Instead, items addressing similar learning outcomes are generally clustered together and positioned in the booklets based on these clusters. For this reason, IP effects, not item order, is the focus of the present study. Moreover, it was observed that in the studies where IP effects were examined by using data obtained from large-scale exams, mostly CTT based methods

were used to identify DIF. The current study has some strengths since IRT-based DIF methods are used on real data. In IRT-based DIF studies, generally, 1 parameter logistic (PL) or 2PL models are used without checking for model-data compatibility. However, in the present study, the model was selected by testing the model-data fit. It is believed that the results of the present study will provide test developers preparing different booklets with foresight regarding whether IP effects will lead to DIF or not.

## 2. METHOD

The study group of the present study was comprised of 9737 students who took the TEOG exam during the first term in the 8th-grade on 23rd-24th November 2016. The number of male and female participants were 5049 (51.9%) and 4688 (48.1%), respectively.

### 2.1. Instrument

TEOG is a large-scale assessment administered to 8th-grade students by the Ministry of National Education, General Directory of Measurement, Assessment, and Exam Services in Turkey between the years 2013 and 2017. The scores obtained from this exam are used to place primary school graduates in secondary education institutions (Ministry of National Education [MoNE], 2013). TEOG consists of six subtests, each of which includes 20 multiple-choice items. These subtests are (i) Turkish, (ii) Mathematics, (iii) Science and Technology, (iv) Religion, Culture and Ethics, (v) Turkish Revolution History and Kemalism, and (vi) Foreign Languages (English). In this exam, four booklets (A, B, C, D) formed by varying the positions of the same questions were used. In the present study, the data obtained from the TEOG administered during the first term of the 2016-2017 academic year were used. The study focused only on the Turkish subtest.

### 2.2. Data Analysis

In the data analysis phase of the study, first of all, the missing data in the four booklets, each of which included the responses of 2500 students, were deleted. Booklet A was regarded to be the original booklet, and the responses of the students who took Booklet B, C, or D were reorganized according to Booklet A. Finally, the data set was converted to a categorical score of either 0 or 1. The descriptive statistics of the data set by booklet type used in the study are presented in [Table 1](#).

**Table 1.** *Descriptive statistics by booklets.*

Booklet	N	Min	Max	$\bar{X}$	Std. Dev.	Skewness (Std. Error)	Kurtosis (Std. Error)	KR-20
A	2416	.00	20.00	11.082	4.497	.049 (.050)	-.982 (.100)	.816
B	2453	1.00	20.00	10.824	4.525	.084 (.049)	-.912 (.099)	.817
C	2438	1.00	20.00	10.967	4.475	.118 (.050)	-.940 (.099)	.811
D	2430	.00	20.00	11.003	4.427	.083 (.050)	-.927 (.099)	.808
Total	9737	.00	20.00	10.968	4.481	.083 (.025)	-.940 (.050)	.813

There are no clear-cut guidelines for interpreting measures of skewness and kurtosis. However, Huck (2012, p.27) stated that most researchers accept the range between -1 and +1 for approximately normal distribution. When the statistics regarding skewness and kurtosis coefficients in [Table 1](#) are examined, a normal distribution of the data for all the booklets is observed. As the KR-20 reliability coefficients ranged between .81 and .82 across the booklets, the results obtained from these booklets were considered to be reliable. Because values greater than 0.80 are considered to have high reliability (Salvucci et al., 1997).

Whether the data for each booklet are unidimensional or not was examined through a confirmatory factor analysis based on the WLSMV (weighted least squares mean and variance adjusted) estimation method. WLSMV has been recommended for estimating CFA model parameters with categorical variables (Muthén & Muthén, 2010). To run this analysis, the “lavaan” (Rosseel et al., 2019) package in the R software was utilized. The results obtained are summarized in Table 2.

**Table 2.** Dimensionality analysis by booklets.

Goodness of Fit	A	B	C	D	Criterion*
$\chi^2/df$	294.217/17 0=1.731	336.128/170= 1.977	333.534/170= 1.961	268.263/170= 1.578	$\leq 5$ Moderate fit $\leq 3$ Perfect fit
CFI	.993	.991	.990	.994	$\geq .90$ Good fit $\geq .95$ Perfect fit
NNFI	.992	.990	.989	.993	$\geq .90$ Good fit $\geq .95$ Perfect fit
RMSEA	.017	.020	.020	.015	$\leq .05$ Perfect fit $\leq .08$ Good fit
SRMR	.024	.026	.026	.023	$\leq .05$ Perfect fit $\leq .08$ Good fit

\*Hu & Bentler, 1999; Sümer, 2000; Kline, 2005; Brown, 2006; Hooper, Coughlan & Mullen, 2008.

When Table 2 is examined, the model-data compatibility for each of the four booklets is observed to be a perfect fit. Based on these findings, it was concluded that the measured construct that unidimensional. This outcome also indicates that the data sets displayed local independence (Hambleton et al., 1991). Finally, model-data compatibility analyses were run to decide which unidimensional parametric IRT model was the most appropriate for the data set used in the study. The results that the analyses yielded are summarized in Table 3.

**Table 3.** Comparison of models with the likelihood-based statistics.

Booklet	Model	Model Fit Indices			Difference		
		AIC	BIC	Log-likelihood	$\Delta\chi^2$	$\Delta df$	<i>p</i>
Booklet A (N=2416)	1PL	56918.35	57039.93	-28438.17			
	2PL	56226.22	56457.82	-28073.11	730.1	19	.00
	3PL	55939.61	56287.00	-27909.80	326.6	20	.00
Booklet B (N=2453)	1PL	58145.46	58267.36	-29051.73			
	2PL	57491.92	57724.12	-28705.96	691.5	19	.00
	3PL	57245.04	57593.34	-28562.52	286.9	20	.00
Booklet C (N=2438)	1PL	58016.17	58137.95	-28987.09			
	2PL	57401.83	57633.79	-28660.92	652.3	19	.00
	3PL	57102.99	57450.93	-28491.50	338.8	20	.00
Booklet D (N=2430)	1PL	57598.17	57719.88	-28778.08			
	2PL	57041.56	57273.39	-28480.78	594.6	19	.00
	3PL	56791.39	57139.13	-28335.69	290.2	20	.00
Total (N=9737)	1PL	230673.50	230824.30	-115315.70			
	2PL	228105.80	228393.20	-114012.90	2605.7	19	.00
	3PL	226973.90	227404.90	-113426.90	1172.0	20	.00

When the item parameters obtained from the 1-, 2- and 3PL models and the  $\Delta\chi^2$  differences summarized in Table 3 were examined, it was concluded that the 3PL model is fitted the Turkish subtest of TEOG. For this reason, the 3PL model was used for the DIF analyses run by utilizing the Lord's chi-square (Lord's  $\chi^2$ ) and Raju's unsigned area methods. These two methods were tested for both with and without item purification. Item purification is used to decrease the effect of items displaying DIF based on the results obtained from DIF methods and is, hence, used to increase the validity of the results (Candell & Drasgow, 1988). In IRT-based methods, item purification is realized by rescaling item parameters in both of the two groups generally based on the reference group scale, while in each step of the purification process, all the items identified as DIF are eliminated and the remaining items are rescaled (Magis & Facon, 2012). In the analyses where items with DIF are taken into consideration, there is a high possibility of Type I error occurrence owing to the fact that items without DIF can be identified as items with DIF (Clauser et al., 1993). However, with the item purification approach the inflation in Type I error rates can be avoided and the power to identify items with DIF can be increased (Magis & Facon, 2012). Hence, in the present study, the effect of item purification on DIF results has also been examined. DIF analyses were run with "difR" package in the R software (Magis et al., 2015) and on the maximum likelihood method. The methods used in the research are, in brief, as follows:

### 2.2.1. Lord's chi-square test

Lord's  $\chi^2$  the hypothesis whether the item parameters (depending on the IRT model used) in one group are different from those in other groups. This method looks at whether there are significant differences between the two groups with statistics (Price, 2014). Lord's  $\chi^2$  is for the item characteristic curves (ICCs) equality between reference groups and focus groups, and is calculated using the following equation:

$$\chi^2 = (v_{iR} - v_{iF})' \Sigma^{-1} (v_{iR} - v_{iF})$$

where  $(v_{iR} - v_{iF})'$  is a vector of differences in the  $i$ -th item parameter estimations (discrimination, difficulty, and pseudo-guessing) between the focus group and the reference group, while  $\Sigma^{-1}$  is the inverse of the asymptotic variance-covariance matrix for differences in item parameter estimations. Lord's  $\chi^2$  test allows for detecting uniform or non-uniform DIF among two groups by setting an appropriate item response model (Lord, 1980, pp. 217-223). When the estimated  $\chi^2$  for  $i$ -th item is significant at .05 level in the present study, this item is flagged as DIF.

### 2.2.2. Raju's area method

Raju (1988, 1990) enhanced the formulas from the area method originally proposed by Rudner, Geston, and Knight (1980) for calculating the exact area between two item response functions (IRFs) derived from two different groups, and presented two statistical tests, called signed and unsigned area methods, for assessing whether the area between two estimated IRFs is significantly different from zero for the 1-, 2- and 3PL models. According to Raju (1988), the signed area (SA) is referred to as the difference between two item characteristic curves, whereas the unsigned area (UA) is referred to as the distance. The SA is computed from the difference between item difficulty parameters, whereas the UA is calculated from the difference between both difficulty and discrimination parameters. Thus, the SA is about uniform DIF, while the UA is related to the non-uniform DIF. Raju (1988) showed that when the  $c$ -parameters (pseudo-guessing parameter) are unequal, the area between two IRFs was infinite and that infinite procedures for estimating the area between two IRFs with unequal  $c$ -parameter yield misleading results. Raju (1988, 1990) proposed to make equal or fixed  $c$ -parameters for this problem. Therefore,  $c$ -parameters in the focal group were fixed to those in the reference group of the present study. Raju's UA is calculated through the following equation:



$$\text{Raju's UA} = (1 - c) \left| \left( \frac{2(a_2 - a_1)}{Da_1a_2} \right) \ln \left[ 1 + e^{\frac{Da_1a_2(b_2 - b_1)}{a_2 - a_1}} \right] - (b_2 - b_1) \right|$$

where a, b and c are the estimation of item discrimination, difficulty, and pseudo-guessing estimates, respectively.

### 2.2.3. Identify DIF items

To identify DIF items in the present study, each booklet was analyzed using the Lord's  $\chi^2$  and Raju's UA methods with and without purification, separately. Then DIF items were flagged in each booklet. Booklet A was optionally chosen as the reference group and the remaining booklets were used as focus groups in all analyses. The results are presented in such a way that Booklet A was compared against booklets B, C, and D.

## 3. FINDINGS

With Booklet A being used as the reference group, the data obtained through the pairwise comparisons of the booklets based on the methods of Lord's  $\chi^2$  and Raju's UA are summarized in Tables 4, 5, and 6.

**Table 4.** Results of DIF analysis of the booklet A versus booklet B.

Item		Lord's $\chi^2$		Raju's UA	
Position in A	Position in B	Without purification	With purification	Without purification	With purification
1	4	11.94*	13.27*	-1.05	-.78
2	5	3.99	4.63	-1.38	-1.38
3	6	1.02	1.71	.71	.21
4	3	3.34	2.03	-1.69	-1.60
5	2	5.49	5.74	-1.88	-2.41*
6	1	10.50*	10.55*	-2.81*	-3.73*
7	12	.93	2.29	.61	-4.25*
8	13	4.12	5.95	-1.84	-3.66*
9	14	10.21*	7.68	1.69	1.40
10	15	4.54	4.06	-1.83	-2.01*
11	16	1.49	2.28	-1.22	-2.49*
12	17	3.88	3.24	.81	.41
13	11	3.86	5.35	-1.77	-2.88*
14	10	3.85	2.49	-.75	-1.32
15	9	9.44*	11.29*	-.95	-1.56
16	8	7.53	9.08*	1.43	1.05
17	7	7.07	5.87	-2.19*	-2.80*
18	19	3.38	1.93	-1.52	-1.18
19	20	8.16*	6.14	-1.46	-1.77
20	18	1.21	1.50	-1.00	-1.46

\* $p < .05$

As can be observed in Table 4, in the Lord's  $\chi^2$  method, the items displaying DIF without item purification are items 1, 6, 9, 15, and 19, while items displaying DIF with item purification are items 1, 6, 15 and 16. In the Raju's UA method, items displaying DIF without item purification are items 6 and 17, while those with item purification are identified as items 5, 6, 7, 8, 10, 11, 13, and 17.

As can be observed in Table 5, in the Lord's  $\chi^2$  method, the items with DIF for both with and without item purification conditions are items 1, 2, 13, and 16. In the Raju's UA method, items displaying DIF without item purification are items 13 and 16, while those with item purification are identified as items 10, 13, and 16.



**Table 5.** Results of DIF analysis of the booklet A versus booklet C.

Item			Lord $\chi^2$		Raju's UA	
	Position in A	Position in C	Without purification	With purification	Without purification	With purification
1		6	9.19*	9.73*	.52	.41
2		3	13.31*	14.12*	-1.32	-1.36
3		2	4.83	5.43	.24	.21
4		1	3.79	2.85	-1.62	-1.83
5		4	1.43	2.11	.35	.09
6		5	6.52	7.26	-1.12	-1.14
7		13	2.15	2.50	1.41	1.01
8		11	5.02	4.57	-1.30	-1.38
9		12	4.80	4.16	1.87	1.93
10		9	6.22	5.59	-1.91	-2.07*
11		8	1.66	2.15	-1.02	-1.02
12		7	1.09	1.53	.47	.35
13		14	11.27*	11.51*	-2.29*	-2.33*
14		15	1.14	1.41	-.17	-0.28
15		10	5.39	6.33	1.13	1.11
16		19	7.90*	8.79*	2.05*	2.03*
17		20	3.63	3.82	-0.66	-.89
18		17	2.18	2.16	-1.44	-1.51
19		18	2.58	2.89	-1.58	-1.64
20		16	.35	.74	.48	.33

\* $p < .05$ **Table 6.** Results of DIF analysis of the booklet A versus booklet D.

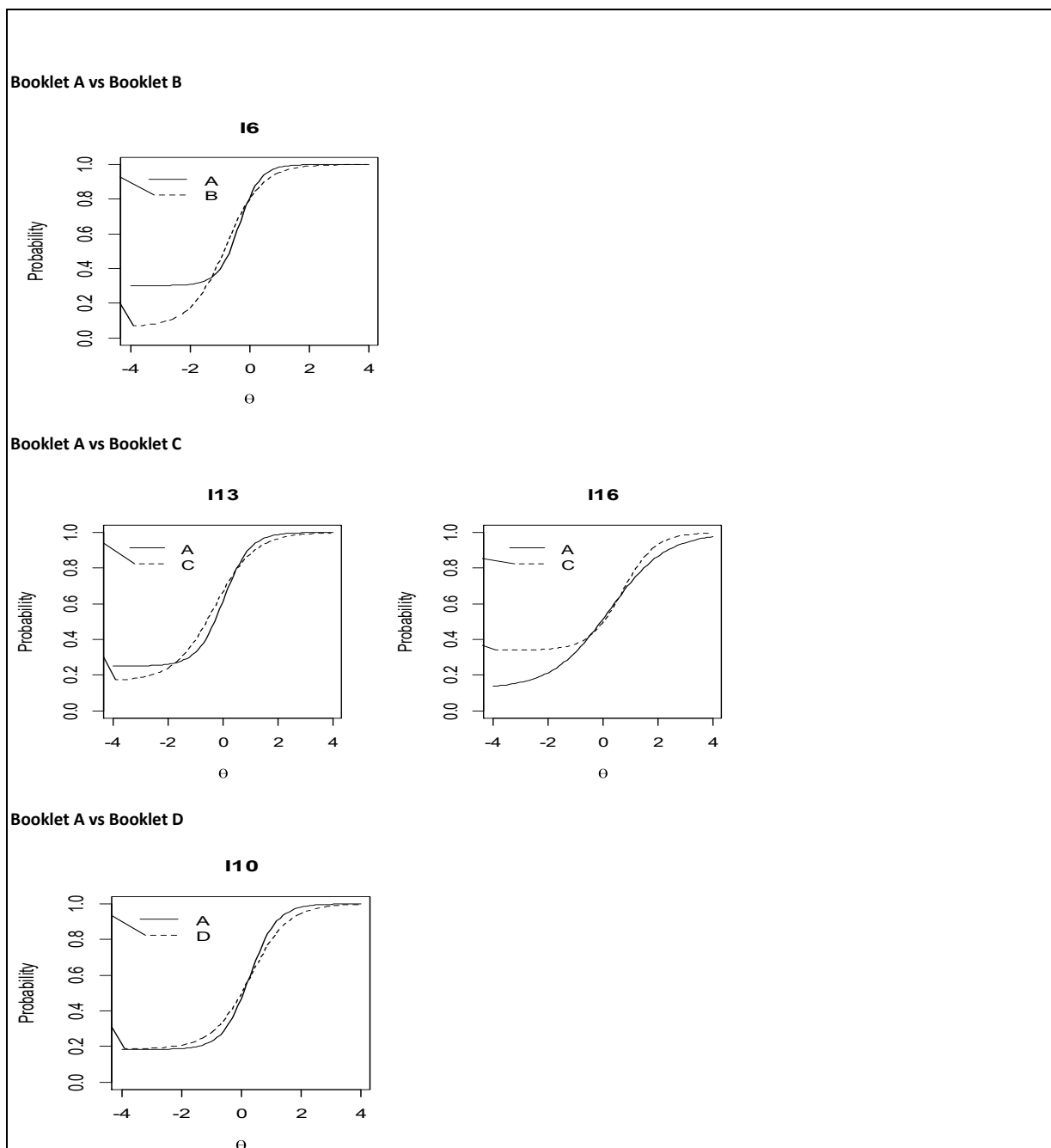
Item			Lord $\chi^2$		Raju's UA	
	Position in A	Position in D	Without purification	With purification	Without purification	With purification
1		2	.13	.30	.22	.13
2		4	5.77	21.52*	-.28	-1.02
3		5	4.42	8.98*	-.45	-.67
4		6	1.18	3.69	-.76	-.42
5		1	9.86*	15.73*	-1.25	-1.09
6		3	7.23	23.21*	-1.65	-2.28*
7		11	4.93	28.85*	-2.04*	-4.14*
8		14	9.44*	25.83*	-.98	-1.07
9		13	1.10	1.98	.93	.44
10		16	10.21*	18.06*	-2.24*	-2.31*
11		17	3.60	13.14*	-1.78	-2.50*
12		15	8.70*	7.17	-2.10*	-2.45*
13		10	2.47	12.23*	-.99	-1.87
14		9	1.83	5.97	-1.03	-1.29
15		12	.55	4.04	.50	.30
16		7	3.60	5.42	1.50	1.24
17		8	2.76	5.23	-1.48	-2.16*
18		20	4.00	6.16	-1.53	-1.01
19		18	3.34	3.76	-1.47	-1.65
20		19	2.01	3.46	-1.24	-1.36

\* $p < .05$

As can be observed in Table 6, in the Lord’s  $\chi^2$  method, the items displaying DIF without item purification are items 5, 8, 10, and 12, while items displaying DIF with item purification are items 2, 3, 5, 6, 7, 8, 10, 11, and 13. In the Raju’s UA method, items displaying DIF without item purification are items 7, 10, and 12, while those with item purification are identified as items 6, 7, 10, 11, 12, and 17.

Besides, ICCs were examined for the items flagged as DIF in all conditions (methods x purification) Item 6 was flagged as DIF in both Booklet A and Booklet B. Item 6 in Booklet A is item 1 in Booklet B. In the comparison of Booklet A and Booklet C, items 13 and 16 were flagged as DIF. Items 13 and 16 in Booklet A are items 14 and 10 in Booklet C, respectively. In the comparison of Booklet A and Booklet D, only item 10 was flagged as DIF. This item is item 16 in Booklet D. The ICCs of these four items were shown in Figure 1. It could be observed in Figure 1 that these items displayed non-uniform DIF.

Figure 1. ICCs of DIF items flagged by Lord’s  $\chi^2$  and Raju’s UA methods.



#### 4. DISCUSSION and CONCLUSION

In the present study, the effect of using different booklets formed by changing the position of the same items, which is frequently a preferred practice in large-scale tests, on test-takers' responses was investigated. To this end, four booklets of the Turkish subtest in the 2016 TEOG exam was examined. First, Lord's  $\chi^2$  identified more items with DIF than Raju's unsigned area did in the without item purification condition. Then, items flagged as DIF in the Raju's unsigned area method are generally flagged as DIF in the Lord's  $\chi^2$  method, as well.

In the condition of with item purification, as in the condition of without item purification, fewer items with DIF were observed in the Raju's UA method than in the Lord's  $\chi^2$  method. However, the results that both methods yielded were not revealed to be as consistent as they were in the condition of without item purification. In both methods, the items flagged as DIF when Booklet A was compared against booklets B were more than the items Booklet A was compared against booklets C and D. This could result from the fact that the highest level of similarity in terms of item position was between Booklets A and Booklet C. Thus, it made us think that performing item purification with Lord's  $\chi^2$  and Raju's UA methods tended to be more sensitive than performing without purification. The results of the present study showed consistency with those reported by Özdemir (2015), the study of whom yielded results that were obtained in both with and without item purification using the methods of Lord's  $\chi^2$  and Raju's signed area. Özdemir reported that both Lord's chi-square and Raju's signed area (for 1PL) methods with or without item purification affected both the number of DIF items and DIF items.

In the literature, there are not only studies reporting that item position can have an impact on individuals' performance (Leary & Dorans, 1985; Hambleton, 1968; Wu et al., 2019), but also studies reporting that item position can lead to bias in item parameter estimations (Debeer & Janssen, 2013; Meyers et al., 2009). Meyers et al. (2009), who researched the effect of item position based on IRT, stated that 56% of the variance in item difficulty between the two tests stemmed from the change in the order of the items. Similarly, Debeer and Janssen (2013) reported that in the 2006 PISA reading test, the fact that the item was positioned in a cluster further below the test led to estimations of item difficulty. Taking into consideration that the differentiation in the item parameters reflects onto the ICCs, it can be claimed that this can result in statistically significant results in differential item functioning.

In the present study, the fact that the items flagged as DIF are generally positioned at considerably different places between booklets can indicate that DIF may result from the position of the item in the test. To illustrate, among the items flagged as DIF in at least one method, items 6, 9, 15, and 17 in Booklet A are in the order of 1, 14, 9, and 7 in Booklet B, respectively. Thus, the results obtained in the present study display consistency with those reported in the related literature. However, in the present study, the same items positioned close to each other in different booklets were also revealed to flag as DIF in some conditions (with or without purification) in at least one method (e.g. such items as 2 and 13 in Booklet A are in 3rd and 14th order in Booklet C). In this case, the reason underlying DIF may not be based on item position. It may have arisen due to a type 1 error caused by sampling.

With the consideration of the effects of item position on item difficulty, an item positioned at the end of a test is generally more difficult than the same item positioned at the beginning of the test (Hambleton, 1968; Li et al, 2012; Rose et al., 2019; Weirich et al., 2017). In consistence with the literature, the analyses conducted in the present study also yielded similar results. When the items flagged as DIF were examined in at least one method, item 15 in Booklet A was found to be 9 in Booklet B, and this item was found more difficult by the test takers of Booklet A (see [Appendix-1](#)). This could be attributed to the fatigue effect, mentioned in the study by Davis and Ferdous (2005).

There are also studies reporting that ordering items in a test from easy to difficult has an impact on the probability of giving correct responses to the items (Balta & Omur Sunbul, 2017; Çokluk et al., 2016). In the present study, some items flagged as DIF were evaluated within this scope. To illustrate, the first item in Booklet A, which flagged as DIF, was item 6 and item 5 in Booklets B and C, respectively. When [Appendix-1](#), which presents a summary of the item parameters, is examined, it is observed that this item is the most difficult in the test. Hence, starting a booklet with an easy or difficult item can be an advantage or a disadvantage.

In conclusion, based on the findings of the present study, it can be claimed that the method of Lord's  $\chi^2$  has a higher tendency of flagging items as DIF when compared to the method of Raju's UA. Moreover, it should not be ignored that there may be some prediction error in the DIF results obtained from Raju's UA method since the guessing parameters of the focus group is fixed to the ones of the reference group. As can also be observed in the present study, no method can definitely identify the presence of items flagged as DIF. Even though an item flagged as DIF in any method is no evidence that this item has DIF, it may still require this item to be examined. As a criterion, items flagged as DIF in more than one condition can be examined in detail. When item parameters, the positions of the items, and/or their content are examined carefully, conditions that could be causing DIF can be understood. In the present study which focused on the impact of item position on DIF, it was deduced that an item being positioned at first or last when compared to another booklet could provide an advantage or disadvantage to the test takers.

It is believed that the findings of the present study could provide test developers who prepare different booklets with insight into whether or not IP effects may result in DIF. When forming different booklets, to avoid the occurrence of DIF resulting from IP effects, it is recommended that the same items be positioned in similar locations in the different booklets. The present study is believed to be a significant contribution to the related literature as there is a limited number of studies including DIF analysis based on the 3PL model (Choi et al., 2014; Monahan & Ankenmann, 2010; Uysal et al., 2019; Zwick et al., 1995). In fact, no recent study that tested the Raju's area method based on the 3PL model with real data was encountered in the literature. Hence, in future studies, IP effects based on Raju's area with the 3PL model can be compared with other methods under different conditions. With this kind of simulation study, the results obtained in a condition where there is a fixed c-parameter can be examined. Researchers are recommended to conduct further studies examining the effect of item position together with item order and/or item content on DIF.

### **Declaration of Conflicting Interests and Ethics**

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

### **Authorship contribution statement**

**Sumeyra Soysal:** Investigation, Resources, Software, Visualization, Methodology, Formal Analysis, and Writing - **Esin Yilmaz-Kogar:** Investigation, Resources, Methodology, Formal Analysis, and Writing.

### **ORCID**

Sümeýra SOYSAL  <https://orcid.org/0000-0002-7304-1722>

Esin YILMAZ KOĞAR  <https://orcid.org/0000-0001-6755-9018>

## 5. REFERENCES

- Akayleh, A. S. A. (2018). *Precision of the estimations for some methods of the CTT and IRT as a base to display the differential item functions on the different item ordered test formats.* <https://bit.ly/3aJeFKx>
- Avcu, A., Tunç, E. B., & Uluman, M. (2018). How the order of the items in a booklet affects item functioning: Empirical findings from course level data?. *European Journal of Education Studies*, 4(3), 227-239. <http://doi.org/10.5281/zenodo.1199695>
- Balta, E., & Omur Sunbul, S. (2017). An investigation of ordering test items differently depending on their difficulty level by differential item functioning. *Eurasian Journal of Educational Research*, 72, 23-42. <https://doi.org/doi:10.14689/ejer.2017.72.2>
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.
- Bulut, O. (2015). An empirical analysis of gender-based DIF due to test booklet effect. *European Journal of Research on Education*, 3(1), 7-16. <https://bit.ly/3cKkhqf>
- Bulut, O., Quo, Q., & Gierl, M. J. (2017). A structural equation modeling approach for examining position effects in large-scale assessments. *Large-scale Assessments in Education*, 5(1), 8. <http://doi.org/10.1186/s40536-017-0042-x>
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage Publications.
- Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, 12(3), 253-260. <https://conservancy.umn.edu/bitstream/handle/11299/107645/v12n3p253.pdf?sequence=1>
- Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education*, 6(4), 269-279. [https://doi.org/10.1207/s15324818ame0604\\_2](https://doi.org/10.1207/s15324818ame0604_2)
- Choi, Y., Alexeev, N., & Cohen, A. (2014). DIF analysis using a mixture 3PL model with a covariate on the TIMSS 2007 mathematics test. In *KAERA Research Forum*, 1(1), 4-14. [http://www.columbia.edu/~ld208/KAERA\\_2014.pdf#page=5](http://www.columbia.edu/~ld208/KAERA_2014.pdf#page=5)
- Çokluk, Ö., Gül, E., & Dogan-Gül, Ç. (2016). Examining differential item functions of different item ordered test forms according to item difficulty levels. *Educational Sciences: Theory and Practice*, 16(1), 319-330. <http://dx.doi.org/10.12738/estp.2016.1.0329>
- Davis, J., & Ferdous, A. (2005). *Using item difficulty and item position to measure test fatigue*. American Institutes for Research. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.110.847&rep=rep1&type=pdf>
- Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, 50(2), 164-185. <https://ppw.kuleuven.be/okp/pdf/DeBeer2013MIPEW.pdf>
- Doğan Gül, Ç., & Çokluk Bökeoğlu, Ö. (2018). The comparison of academic success of students with low and high anxiety levels in tests varying in item difficulty. *Inonu University Journal of the Faculty of Education*, 19(3), 252-265. <https://doi.org/10.17679/inuefd.341477>
- Erdem, B. (2015). *Ortaöğretime geçişte kullanılan ortak sınavların değişen madde fonksiyonu açısından kitapçık türlerine göre farklı yöntemlerle incelenmesi* [Investigation of Common Exams Used in Transition to High Schools in Terms of Differential Item Functioning Regarding Booklet Types with Different Methods] [Unpublished master dissertation]. Hacettepe University. Ankara.

- Freedle, R., & Kostin, I. (1991). *The prediction of SAT reading comprehension item difficulty for expository prose passages* (ETS Research Report, RR-91-29). Princeton, NJ: Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.1991.tb01396.x>
- Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, 28(3), 39-53. <https://doi.org/10.1111/j.1745-3992.2009.00154.x>
- Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science Quarterly*, 50(3), 379–390. <https://bit.ly/3aHHyGD>
- Hambleton, R. K. (1968). *The effects of item order and anxiety on test performance and stress*. Paper presented at the meeting of American Educational Research Association. <https://files.eric.ed.gov/fulltext/ED017960.pdf>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.
- Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling*, 54(4), 418-431. <https://core.ac.uk/download/pdf/25705605.pdf>
- Hecht, M., Weirich, S., Siegle, T., & Frey, A. (2015). Effects of design properties on parameter estimation in large-scale assessments. *Educational and Psychological Measurement*, 75(6), 1021-1044. <https://doi.org/10.1177/0013164415573311>
- Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the linear logistic test model. *Psychology Science Quarterly*, 50, 391-402. <https://bit.ly/39Sb9xY>
- Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E., & Khorramdel, L. (2011). Analysing item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation*, 17(6), 497-509. <https://doi.org/10.1080/13803611.2011.632668>
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp.129-143). Erlbaum.
- Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Lawrence Erlbaum Associates.
- Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modeling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods* 6(1), 53-60. <http://arrow.dit.ie/cgi/viewcontent.cgi?article=1001&context=buschmanart>
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55. <https://doi.org/10.1080/10705519909540118>
- Huck, S. W. (2012). *Reading statistics and research* (6th ed.). Pearson.
- Kamata, A., & Vaughn, B. K. (2004). An introduction to differential item functioning analysis. *Learning Disabilities: A Contemporary Journal*, 2(2), 49-69. (EJ797693). ERIC. <https://eric.ed.gov/?id=EJ797693>
- Karami, H. (2012). An introduction to differential item functioning. *The International Journal of Educational and Psychological Assessment*, 11(2), 59-76. <https://psycnet.apa.org/record/2012-28410-004>
- Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, 8(2), 147-154. <https://conservancy.umn.edu/bitstream/handle/11299/101880/1/v08n2p147.pdf>



- Kleinke, D. J. (1980). Item order, response location, and examinee sex and handedness on performance on multiple-choice tests. *Journal of Educational Research*, 73(4), 225–229. <https://doi.org/10.1080/00220671.1980.10885240>
- Kline, R. B. (2005). *Principles and practice of structural equation modeling*. The Guilford Press.
- Klosner, N. C., & Gellman, E. K. (1973). The effect of item arrangement on classroom test performance: Implications for content validity. *Educational and Psychological Measurement*, 33, 413–418. <https://doi.org/10.1177/001316447303300224>
- Le, L. T. (2007, July). *Effects of item positions on their difficulty and discrimination: A study in PISA Science data across test language and countries*. Paper presented at the 72nd Annual Meeting of the Psychometric Society, Tokyo.
- Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research*, 55(3), 387–413. <https://doi.org/10.3102/00346543055003387>
- Li, F., Cohen, A., & Shen, L. (2012). Investigating the effect of item position in computer-based tests. *Journal of Educational Measurement*, 49(4), 362–379. <https://doi.org/10.1111/j.1745-3984.2012.00181.x>
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum.
- Magis, D., Beland, S., & Raiche, G. (2015). *Package 'difR'* (Version: 5.0). [Computer software manual]. Retrieved May 14, 2018. Retrieved from <https://cran.rproject.org/web/package/s/difR/difR.pdf>
- Magis, D., & Facon, B. (2012). Item purification does not always improve DIF detection: A counterexample with Angoff's delta plot. *Educational and Psychological Measurement*, 73(2), 293–311. <https://doi.org/10.1177/0013164412451903>
- Martin, M. O., Mullis, I. V. S., & Chrostowski, S. J. (2004). Item analysis and review. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 224–251). TIMSS & PIRLS International Study Center, Boston College.
- McNamara, T., & C. Roever (2006) *Language testing: The social dimension*. Blackwell.
- Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item position and item difficulty change in an IRT based common item equating design. *Applied Measurement in Education*, 22(1), 38–60. <https://doi.org/10.1080/08957340802558342>
- Ministry of National Education [MoNE], (2013). *2013-2014 Eğitim-öğretim yılı ortaöğretimi geçiş ortak sınavları e-klavuzu*. Ankara.
- Monahan, P. O., & Ankenmann, R. D. (2010). Alternative matching scores to control type I error of the Mantel–Haenszel procedure for DIF in dichotomously scored items conforming to 3PL IRT and nonparametric 4PBCB models. *Applied Psychological Measurement*, 34(3), 193–210. <https://doi.org/10.1177/0146621609359283>
- Muthén, L. K., & Muthén, B. O. (2010). *Mplus: Statistical analysis with latent variables user's guide 6.0*. Muthén & Muthén.
- Newman, D. L., Kundert, D. K., Lane Jr, D. S., & Bull, K. S. (1988). Effect of varying item order on multiple-choice test scores: Importance of statistical and cognitive difficulty. *Applied Measurement in Education*, 1(1), 89–97. [https://doi.org/10.1207/s15324818ame0101\\_8](https://doi.org/10.1207/s15324818ame0101_8)
- Ollennu, S. N. N., & Etsey, Y. K. A. (2015). The impact of item position in multiple-choice test on student performance at the basic education certificate examination (BECE) level. *Universal Journal of Educational Research*, 3(10), 718–723. <https://doi.org/10.13189/ujer.2015.031009>

- Özdemir, B. (2015). A comparison of IRT-based methods for examining differential item functioning in TIMSS 2011 mathematics subtest. *Procedia-Social and Behavioral Sciences*, 174, 2075-2083. <https://doi.org/10.1016/j.sbspro.2015.02.004>
- Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement Issues and Practice*, 19(3), 5–15. <https://doi.org/10.1111/j.1745-3992.2000.tb00033.x>
- Perlini, A. H., Lind, D. L., & Zumbo, B. D. (1998). Context effects on examinations: The effects of time, item order and item difficulty. *Canadian Psychology/Psychologie Canadienne*, 39(4), 299-307. <https://doi.org/10.1037/h0086821>
- Plake, B. S., Patience, W. M., & Whitney, D. R. (1988). Differential item performance in mathematics achievement test items: Effect of item arrangement. *Educational and Psychological Measurement*, 48(4), 885-894. <https://doi.org/10.1177/0013164488484003>
- Qian, J. (2014). An investigation of position effects in large-scale writing assessments. *Applied Psychological Measurement*, 38(7), 518-534. <https://doi.org/10.1177/0146621614534312>
- Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53(4), 495-502. <https://link.springer.com/article/10.1007/BF02294403>
- Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, 14(2), 197-207. <https://conservancy.umn.edu/bitstream/handle/11299/113559/v14n2p197.pdf?sequence=1>
- Rose, N., Nagy, G., Nagengast, B., Frey, A., & Becker, M. (2019). Modeling multiple item context effects with generalized linear mixed models. *Frontiers in Psychology*, 10,248. <https://doi.org/10.3389/fpsyg.2019.00248>
- Rosseel, Y., Jorgensen, T., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., Barendse, M., & Scharf, F. (2019). *Package 'lavaan'* (Version: 0.6-5) [Computer software manual]. <https://cran.r-project.org/web/packages/lavaan/lavaan.pdf>
- Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). Biased item detection techniques. *Journal of Educational Statistics*, 5, 213-233. <https://doi.org/10.2307/1164965>
- Ryan, K. E., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education*, 14(1), 73-90. [https://doi.org/10.1207/S15324818AME1401\\_06](https://doi.org/10.1207/S15324818AME1401_06)
- Salvucci, S., Walter, E., Conley, V., Fink, S., & Saba, M. (1997). *Measurement error studies at the National Center for Education Statistics (NCES)*. U.S. Department of Education.
- Schmitt, A. P., & Crone, C. R. (1991). Alternative mathematical aptitude item types: DIF issues. *ETS Research Report Series*, 1991(2), i-22. <https://doi.org/10.1002/j.2333-8504.1991.tb01409.x>
- Sümer, N. (2000). Yapısal eşitlik modelleri: Temel kavramlar ve örnek uygulamalar [Structural Equation Modeling: Basic Concepts and Applications]. *Türk Psikoloji Yazıları*, 3(6), 49-73. <https://psycnet.apa.org/record/2006-04302-005>
- Tal, I. R., Akers, K. G. & Hodge, K. G. (2008). Effect of Paper color and question order on exam performance. *Teaching of Psychology*, 35(1), 26-28. <https://doi.org/10.1080/00986280701818482>
- The West African Examinations Council [WAEC] (1993). *The effects of item position on performance in multiple choice tests*. Research Report, Research Division, WAEC, Lagos.

- Tippets, E., & Benson, J. (1989). The effect of item arrangement on test anxiety. *Applied Measurement in Education*, 2(4), 289-296. [https://doi.org/10.1207/s15324818ame0204\\_2](https://doi.org/10.1207/s15324818ame0204_2)
- Trendtel, M., & Robitzsch, A. (2018). Modeling item position effects with a Bayesian item response model applied to PISA 2009–2015 data. *Psychological Test and Assessment Modeling*, 60(2), 241-263. <https://bit.ly/3cQWkh5>
- Uysal, I., Ertuna, L., Ertas, F. G., & Kelecioğlu, H. (2019). Performances based on ability estimation of the methods of detecting differential item functioning: A simulation study. *Journal of Measurement and Evaluation in Education and Psychology*, 10(2), 133-148. <https://doi.org/10.21031/epod.534312>
- Verguts, T., & De Boeck, P. (2000). A Rasch model for detecting learning while solving an intelligence test. *Applied Psychological Measurement*, 24, 151-162. <https://doi.org/10.1177/01466210022031589>
- Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement*, 38, 535-548. <https://doi.org/10.1177/0146621614534955>
- Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement*, 41(2), 115-129. <https://doi.org/10.1177/0146621616676791>
- Wu, Q., Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2019). Predictors of individual performance changes related to item positions in PISA assessments. *Large-scale Assessments in Education*, 7(5), 1-21. <https://doi.org/10.1186/s40536-019-0073-6>
- Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Directorate of Human Resources Research and Evaluation, Department of National Defense.
- Zwick, R., Thayer, D. T., & Wingersky, M. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement*, 32(4), 341-363. <https://www.jstor.org/stable/1435217>

## 6. APPENDIX

### Appendix-1 Item parameter estimation for booklets.

Item	Booklet A			Booklet B			Booklet C			Booklet D		
	a	b	c	a	b	c	a	b	c	a	b	c
I1	1.14	2.43	.21	1.04	2.18	.23	1.20	2.23	.24	1.23	2.36	.21
I2	2.22	-.23	.14	2.03	-.19	.19	1.87	-.42	.15	2.12	-.24	.20
I3	1.37	-1.48	.05	1.61	-1.05	.33	1.47	-1.33	.28	1.28	-1.16	.40
I4	2.21	.53	.20	1.95	.62	.20	1.80	.60	.19	2.18	.59	.23
I5	2.39	.76	.19	1.64	.78	.18	2.47	.74	.21	2.01	.84	.24
I6	2.84	-.34	.30	1.66	-.78	.06	2.51	-.50	.30	2.01	-.58	.25
I7	1.73	-1.34	.00	1.76	-1.38	.00	1.82	-1.39	.00	1.55	-1.53	.00
I8	1.99	-.59	.13	1.74	-.85	.00	1.51	-.78	.05	1.66	-.60	.20
I9	1.32	.30	.20	1.57	.62	.24	1.69	.64	.29	1.47	.48	.25
I10	2.21	.29	.18	1.88	.36	.20	1.69	.30	.18	1.58	.32	.19
I11	2.27	-.08	.21	1.93	-.22	.15	2.04	-.21	.18	1.72	-.28	.13
I12	.94	.81	.21	1.21	.85	.22	1.00	.73	.23	.65	.15	.00
I13	2.11	.03	.25	1.85	-.20	.16	1.38	-.30	.17	1.92	-.10	.23
I14	1.34	.47	.21	1.19	.50	.16	1.30	.46	.22	1.07	.40	.17
I15	.92	-.19	.01	.72	-.44	.00	1.34	.25	.23	.99	.09	.12
I16	.97	.23	.12	1.35	.65	.33	1.67	.71	.34	1.31	.72	.31
I17	2.00	.77	.12	1.39	.81	.07	1.84	.80	.15	1.63	.67	.07
I18	2.26	.30	.22	2.25	.44	.24	1.82	.20	.18	2.11	.43	.25
I19	3.47	.81	.27	2.79	.88	.22	2.56	.77	.25	2.65	.85	.24
I20	2.10	1.16	.33	1.67	1.08	.30	2.33	1.12	.34	1.50	1.23	.31

## Validity and Reliability Evidence of Professional Obsolescence Scale According to Different Test Theories

Sadegul Akbaba Altun <sup>1</sup>, Sener Buyukozturk <sup>2</sup>, Merve Yildirim Seheriyeli <sup>2,\*</sup>

<sup>1</sup>Başkent University, Faculty of Education, Department of Educational Sciences, Division of Educational Administration, Ankara, Turkey

<sup>2</sup>Hasan Kalyoncu University, Faculty of Education, Department of Educational Sciences, Division of Educational Measurement and Evaluation, Gaziantep, Turkey

### ARTICLE HISTORY

Received: Aug. 27, 2020

Revised: Feb. 06, 2021

Accepted: Mar. 03, 2021

### Keywords:

School principals,

Professional obsolescence,

Scale development,

Professional development activities.

**Abstract:** This study aims to develop a scale that will determine the factors causing professional obsolescence in the field of education. In this context, the Professional Obsolescence Scale (POS) has been developed to determine the professional and organizational obsolescence of primary, secondary and high school administrators. In this scale development process, steps were followed in line with the suggestions of Crocker and Algina (2006) and Cronbach (1984). Firstly, 63 items were prepared and 991 school principals participated the study. R (version 4.0.1) software was used to analyze the data. Item and test parameters and information functions have been estimated using Samejima's Graded Response Model based on Item Response Theory. Principal Axis Analysis was performed for the construct validity of the scale, and four-dimensions structure with 47 items has been obtained. These dimensions are named as "Being Open to Professional Development", "Job-Ability Harmony in Profession", "Organizational Support in Professional Development", "Professional Burnout". The scores obtained from each dimension are evaluated within themselves. It has been observed that each dimension fulfills the conditions of unidimensionality, local independence, model-data fit and parameter invariance. According to the Classical Test Theory, Cronbach Alpha coefficients are between 0.807 and 0.945. The Stratified Alpha coefficient calculated for the whole scale is 0.94. According to the Item Response Theory, the marginal reliability coefficients were between 0.857 and 0.936 and the empirical reliability coefficients were found between 0.854 and 0.938.

## 1. INTRODUCTION

The concept of "Eskimişlik" is used as "Obsolescence" in English. In this context, it is generally explained as the concepts of professional or managerial obsolescence in studies related organizations. The Turkish Language Institute (TLI) dictionary does not represent the concept of "eskimişlik (Obsolescence)". However, it explains the concepts of old, obsolescence, aging and becoming outdated. The word "eski (old)" as an adjective covers expressions such as "long-

---

\*CONTACT: Merve YILDIRIM SEHERİYELİ ✉ [yldrm.mrv.7806@gmail.com](mailto:yldrm.mrv.7806@gmail.com) 📧 Hasan Kalyoncu University, Faculty of Education, Department of Educational Sciences, Division of Educational Measurement and Evaluation, Gaziantep, Turkey

standing, long-gone, anti-new, previous, non-valid, long-time working in any profession, specialized in the profession, experienced”. Again, as a name form for the word “old”, there are expressions of “a thing worn out, ruined, a word used in cases in which a person does not have his old respectability because he has lost his position or lost his status”. While the name form for the word “aging” is specified as “aging work”; to explain the word getting old, there are expressions such as “to become old, to be worn out, to be disgraced, to be worthless, to get old”. There are many idioms in the TLI Proverbs and Idioms Dictionary relating the word “eski (old)”. Some of these idioms are “old pines have turned into glasses, closing old notebooks, old baths old bowls, taking on the old identity, bringing a new tradition to the old village, not looking for the old one, if the old were in vogue (or reputation) it would rain light on the flea market”.

The concept of obsolescence has been tried to be defined by the organization and management employees. Burke (1969) studied skill obsolescence and, in his study with engineers, found that age was a factor in skill obsolescence. For example, Pazy (1994), who studied the cognitive scheme of professional obsolescence, interviewed 50 professionals and tried to understand the concept of obsolescence and found that professionals attribute different meanings to obsolescence. Pazy (1996), in his other study, also stated that there are three directions of research on obsolescence, the first of which is that the meaning of the concept of obsolescence differs; he stated that the other is about self-improvement and updating awareness, and another is about differences in career steps. Pazy (1996) sees the inability to adapt to change as an essential factor leading to obsolescence. Fossum et al. (1986) mentioned about skill obsolescence and discussed the concept of obsolescence in terms of human resources. Fossum et al. (1986) identified the factors affecting skill obsolescence as motivational, individual, organizational and extrinsic factors.

Shearer and Steger (1975) discussed workforce obsolescence and identified 12 factors leading to managerial and technical work obsolescence. They found that the factors that prevent obsolescence are the high need for success and participation in management. Besides, in their studies, managerial obsolescence, in contrast to professional obsolescence, was associated with more experience, but less with education. Warmington (1974) viewed obsolescence as a systems approach and explained it by taking into account the organization’s business/factory, process and output. If these processes do not meet the conditions of the day, they are considered obsolete. Başaran (2008) defined organizational obsolescence as the gradual insufficiency of an employee who was sufficient when he started his work. He explained the personal reasons for obsolescence as a) being prone to obsolescence, b) emotional disturbance, c) unsuitable working habits, d) inappropriate management style.

Mohan et al. (2001) identified the factors that cause obsolescence and listed them as follows: It has been stated that obsolescence is due to its superior attitude, followed by organizational climate and organizational support. He inferred that superiors played an important role in the development of the administrators of the organization (Chauhan & Chauhan 2005), and they also found that the organizational climate and the superiors’ attitude contributed to managerial obsolescence. Murillo (2011) deals with the concept of obsolescence with its technical and economic obsolescence dimensions. It is atrophy in skills because of the physical weakening of the employee due to the technological age and illness of the employee and not using his skills sufficiently. These are the losses that result from the change of technology and the new skills required by the organization as a result of economic obsolescence, changing depending on the sector and the company, and being unable to keep up with these changes. Other studies take the age variable of the employee as a factor in obsolescence. Burke (1969) surveyed 50 engineers. It has been observed that elderly engineers react less and are not fully equipped to deal with the work. Similarly, in the studies of van Loo et al. (2001), older workers were seen as a high-risk



group for skill obsolescence. Toner (2011) concluded that the quality and quantity of employee skills are crucial to innovation and economic performance.

Van Loo et al. (2001) investigated the relationship between risk factors and skill obsolescence and the role of measures. It was expected that risk factors would cause skill obsolescence and that the measures taken would prevent skill obsolescence. As expected, obsolescence is related to business conditions. Older workers were seen as a high-risk group for skill obsolescence. Improving the conditions in the workplace was seen as a preventive factor. It was stated that developments in technological, organizational and demographic changes also caused skill obsolescence. Contrary to the age of the employee, the age of the organization is a factor in aging and different results have been obtained in some studies. For example, according to Sorensen and Stuart (2000), an organization can innovate and prevent it from obsolescence with aging. For example, they can discover new developments in the field of biotechnology, increase the number of patents and continue as a leading company.

Searching the literature, we realize that the concept of managerial obsolescence has been studied for many years (see Bařaran, 2008; Burke, 1969; Fossum et al., 1986; van Loo et al., 2001; Warmington, 1974), as the renewal of the concept of obsolescence (see Knight, 1998; Rothman & Perrucci, 1971; Sorensen & Stuart, 2000; Shaffer, 1969); change (see Chauhan & Chauhan, 2004, 2008) lifelong learning, labor aging, knowledge obsolescence, human resources (see Fossum et al., 1986; Murillo, 2011; Pazy, 1996; Toner, 2011). It seems that both qualitative (see Pazy, 1994) and quantitative research (see, Shearer & Steger, 1975; Rothman & Perrucci, 1971) have been done in some fields except education. However, the fact that most of these studies are in business (see, Jones, Chanko, Roberts, 2004; Mohan, Chauhan & Chauhan, 2001; Chauhan & Chauhan, 2004, 2005; 2008; 2009) and technical fields (see. Sorensen & Stuart, 2000) should be taken into consideration. Chauhan & Chauhan (2009) say that it is necessary to combat obsolescence. As it is seen above, organizational and managerial obsolescence have been studied in different fields. Only two studies related to education were on pedagogical obsolescence (see, MacNeill & Cavanagh, 2006) and IT related concept of obsolescence. MacNeill and Cavanagh (2006) criticized New Public Management (NPM) reform as the managerial reforms that accompanied accountability affected schools negatively. As a result of the NPM movement pedagogical obsolescence occurred and restricted school principals' pedagogical leadership. Another topic related to obsolescence occurred at schools is planned obsolescence which is related to IT used at schools. Wandera (2015) also mentioned this threat, how schools respond to this and its effects on the teaching and learning process.

The fight against obsolescence should be at both an organizational and individual level. The problem of obsolescence should be shared between the two stakeholders: Individual and organization. Self-improvement and self-improvement initiatives can be done at an individual level. At the organizational level, employees can improve themselves through continuous training (Chauhan & Chauhan, 2009). Up to now, antiquity has been studied in different sectors, but no tool has been developed to measure the level and dimensions of professional obsolescence in educational organizations. Therefore, this study aims to develop a scale that will determine the factors causing professional obsolescence in the field of education, to determine the dimensions that lead to obsolescence, and to determine the antiquity levels of these dimensions.

## 2. METHOD

### 2.1. Working Group

A total of 1001 school principals were reached within the scope of the research. Three people who did not respond to five consecutive items were excluded from the analysis. Of the 998 participants taking part in the analyses, 151 (15%) were women and 847 (85%) were men. In

the Ministry of National Education system, school sizes are symbolized by the letters A, B and C, depending on the number of students. Although it varies depending on the school level, A type schools are relatively large ones, B type schools are medium-size and C type schools are small ones (MEB Eğitim Kurumları, 2009). 654 (66%) of these participants work in A-type schools, 196 (20%) in B and 146 (15%) in C-type schools, two people did not specify. Again, 452 (45%) of the participants work as principals in primary school, 268 (27%) in secondary school, 268 (27%) in high school, 4 (0.4%) in both primary and secondary schools, 6 (0.6%) the person did not specify. Of the participants, 36 (4%) hold associate degrees, 771 (77%) bachelor's degree, while 153 (15%) completed their postgraduate education, and 34 (4%) stated the other option.

With regards to the seniority of the participants as managers, 3 (0.3%) people did not respond. There are 211 (21%) people with seniority of fewer than three years as a manager, 210 (21%) people with seniority of 4-6 years, 375 (38%) people with seniority of 7-18 years, 170 (17%) people with seniority of 19-30 years, 29 (3%) people with a seniority of 30 or more years.

The number of participants who stated that “they took a management course after 1998 and were appointed after the exam”, that is to say, those who attended in-service seminars before becoming managers is 406 (41%); those who stated that “they were appointed before 1998 and took management courses and seminars” is 165 (17%). The number of participants who stated that “they did not take courses and seminars related to management” is 299 (30%); and participants stating that “they took courses and seminars or graduated after becoming a manager” is 56 (6%). 67 (7%) people chose “the other” option.

The duration of participation of the participants in professional development activities in the last 18 months varies between 0 and 130 days with an average of 12 days. The number of days on which these activities are compulsory varies between 0 and 120, with an average of 8 days. Again, the effectiveness levels of these activities ranged from 1: not at all effective to 5: very effective and the average was 4. While 798 (80%) of those participating in these events stated that they did not pay at all, 134 (13%) stated that they paid some and 66 (7%) paid the whole price. Similarly, 729 (73%) of 921 (92%) people who stated that these activities took place during regular working hours said that the activities were organized in a way that they would allow them to participate. 76 out of 77 (8%) people who stated that events were organized outside of regular working hours stated that they received additional payment to participate in these activities. The number of those who want to participate in more activities than the ones available in the last 18 months is 523 (52%), and the number of those who do not is 474 (47%), one person did not specify. The reasons stated by those who did not participate although they wished to participate are as follows.

**Table 1.** *Reasons for participants not to attend in-service activities.*

Reasons	Frequency
I did not have the prerequisites for participation (e.g., qualifications, experience, seniority).	60
Professional development activity was too expensive / I could not afford it.	61
I could not get the necessary support from my higher institution.	80
Activity hours coincided with my work schedule.	223
I did not have time because of my family responsibilities.	105
There was no professional development activity suitable for me.	249
Other	62

As [Table 1](#) shows, it is seen that the most frequently stated reasons are that the activity is not suitable for the participant, the activity time is not suitable for the participant, and they do not

have time due to family responsibilities. The frequency of choosing the expressions that define the attitudes of the participants to support the professional development of their superiors is given in the table below.

**Table 2.** *Participants' views on the attitudes of their superiors to support their professional development.*

Views	Frequency
He is insensitive to the education needs of its subordinates. If subordinates are sent for training, they do not offer a separate time slot for this training.	214
He thinks of his job only as applying corporate goals and policies. It does not take into account the professional needs of subordinates.	2
He thinks that his subordinates' development needs are important. However, he believes that the initiative in this is with the employee. It sees no harm in opportunities to continue their professional development.	214
He is aware of the training needs of its subordinates. At a certain level, it tries to provide opportunities and create environments that will renew them professionally.	346
He is very sensitive to the training needs of his subordinates. Considering their potentials and interests, it creates new opportunities besides the current opportunities according to the needs of their subordinates to provide career development.	255

As seen in Table 2, those who did not find the attitudes of their superior superiors supportive stated 216 opinions, while those who found supportive expressed 815 opinions.

## 2.2. Data Collection Tool

Professional Obsolescence Scale (POS) given in the appendix has been developed to determine the professional and organizational obsolescence levels of primary, secondary and high school principals. In this scale development process, the following steps were followed in line with the suggestions of Crocker and Algina (2006) and Cronbach (1984).

### 1. Defining the feature to be measured and writing the items

While developing the scale, the items of the scale in the article titled "Are you on the verge of antiquity?" published by Chauhan and Chauhan in 2009 and obtained permission for use were used. Also, the theoretical framework, which is included in Bařaran's (2008) Organizational Behavior book and the following dimensions, was taken into consideration. In addition to these, items expressing information and communication technologies were written by the researchers and added to these items.

Bařaran's (2008) definition of organizational obsolescence was taken into consideration and the four sub-dimensions determined there were taken as a basis. These dimensions are summarized below.

- Being suitable for obsolescence: Under this dimension, items were written about the decrease in the motivation of the employee for organizational goals, the negative attitude towards learning, the hardening of actions, especially in attitudes, the decrease in physical strength and aging.
- Emotional Disorder: Items were written for the employees having difficulty in the workplace and prolonged frustration, having feelings of inferiority and guilt, having a constant headache, being stuck with personal problems, taking alcohol continuously and getting vaccinated, smoking and using drugs.

- **Inappropriate Working Habit:** Items expressing the incompatibility of the employee's working method that has turned into a habit in the workplace and the working style required by the organization and the task were written.
- **The inappropriateness of the Management Style:** Items were written that the management style of the organization was not suitable for the personality characteristics of the employee, the employee was forced to carry out a task beyond the competence of the employee, the objectives of the organization were uncertain, the superiors ignored the obsolescence, the counseling was insufficient and the employee was not assigned to a position where the employee would be effective.

Chauhan and Chauhan (2009) stated that there are two main factors in the scale of professional obsolescence that they developed, these are individual and organizational factors, and there are sub-factors under them. This scale consists of 34 items and measures obsolescence in 8 dimensions. Four of them are organizational factors (organizational climate, organizational support, superior attitude, on-the-job development activities) and four are individual factors. Individual factors are listed as professional knowledge and skills, development motivation, attitude towards learning, taking the initiative in self-development. These dimensions are briefly summarized below.

- a. **Organizational climate:** Organizational climate encourages autonomy, innovation, and reward for high performance.
- b. **Organizational support:** The organization develops and supports the training and career plan for its employees.
- c. **The attitude of superiors:** Provides support for the development of subordinates.
- d. **On-the-job development activities:** Employees perceive on-the-job activities that are suitable for improving/updating themselves as appropriate activities for self-improvement.
- e. **Professional knowledge and skills:** Employees perceive that their knowledge and skills are appropriate for their job.
- f. **Development/update motivation:** employees are motivated to improve themselves.
- g. **Attitude towards learning:** Positive or negative attitude towards learning.
- h. **Taking the initiative in self-development:** Employees take the initiative in self-improvement.

Sixty-three items were written first on the scale. Although all of the items are scored between 1 and 5, the attributes indicated by the numbers differ. For example; Some items are scored between 1: Not at all effective and 5: Very effective, while some items are scored between 1: Not at all important and 5: Very important. No item requires reverse coding. In addition, low scores from each of the first three sub-dimensions of the scale and high scores from the fourth sub-dimension indicate that the professional obsolescence is high.

## 2. Expert opinion and revisions of the items

The draft form was prepared by adding demographic information together with the prepared items, and it was examined by five experts in the field and two measurement and evaluation experts. The appropriateness, scoring, and correction suggestions, if any, of each item were requested from the experts. There was no change in the number of items in line with the recommendations. The draft form was finalized by making textual corrections.

## 3. Pre-plot

The draft form prepared at this stage was applied to 10 school principals in terms of clarity, understandability and determination of the implementation period. Five items were corrected textually in line with the verbal feedback received at the end of the application.

#### 4. Application and determination of psychometric properties

The application was carried out both with paper-pencil and using Google Forms. To determine the psychometric properties of the measurements obtained, validity and reliability evidence was presented.

#### 2.3. Data Analysis

First of all, the data set was examined and found out that no variable has more than 10% missing values. But based on all cells, the proportion of missing data was found to be 0.1%. The assignment was made with EM (Expectation - Maximization) algorithm for the missing data. Seven people marking all items as five were excluded from the analysis.

For the evidence of the construct validity of the scale, the responses of 991 individuals were first performed with 63 items and Principal Axis Analysis (PA) using the varimax orthogonal rotation method using the polychoric correlation matrix. In PA, the minimum factor loading was .50 (Todman & Dugard, 2007), and for the detection of overlapping items, the difference between the factor loading of the same item on two factors was taken as .10 (Büyüköztürk, 2019). In addition, the "Checklist for Reporting EFA" proposed by Akbaş et al. (2019) was used for reporting. Although the responses to the items are graded between 1 and 5, the qualifiers for the grades differ. For example; Some items are scored between 1: Not at all effective and 5: Very effective, while some items are scored between 1: Not at all important and 5: Very important. For this reason, Item Response Theory, which examines the validity and reliability evidence at an item level, was taken as a basis.

IRT is a measurement theory that models the relationship between individuals' response patterns to items and their abilities. The property measured in the Classical Test Theory is the sum of the responses given by individuals to the items of a test. Therefore, this observed score is the sum of the person's true score and measurement errors (Crocker & Algina, 2006). For this reason, test and item statistics cannot be calculated independently from the group or the items. In IRT, on the other hand, item parameters can be estimated independently from the sample and ability parameters from the items. This feature is called parameter invariance (DeMars, 2010).

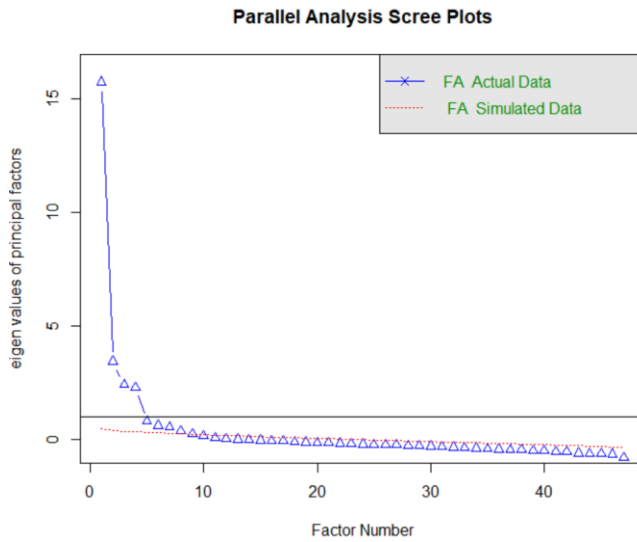
Models used for items scored in two and multi-categories in IRT differ. While Rasch, 1PL, 2PL and 3PL models are used for dichotomously scored items, Partial Credit Model (adjacent category approach) is used as an extension of 1PL model, and Graded Response Model (cumulative category approach) is used as an extension of 2PL model for polytomous scored items (Tang, 1996; Yürekli, 2010). In this study, since the categories are scored cumulatively, the item and test parameter estimates were examined with Samejima's Graded Response Model. IRT assumptions dimensionality, local independence and model-data fit were tested, and parameter invariances were examined. PA for dimensionality, C2 statistics for model-data fit were examined; for parameter invariance, item and study group were randomly divided into two and correlation values were examined (DeMars, 2010). In addition to item discrimination and threshold parameters, item and test information functions, marginal and empirical reliability coefficients. Also, the Stratified Alpha coefficient recommended by Cronbach et al. (1965) to be used in multidimensional scales was also examined. Because even in scales where unidimensionality assumption is not provided, Cronbach's Alpha coefficient can give very high values (Tan, 2009).

### 3. FINDINGS

The KMO (Kaiser Mayer Olkin) value calculated using the polychoric correlation matrix to examine the suitability of the data for factor analysis was found to be 0.94 and the Bartlett sphericity test was found to be significant ( $\chi^2 = 34742.74$ ;  $sd = 1081$ ;  $p < 0.05$ ). In the first stage

of the Principal Axis Analysis, which was carried out with 63 items, it was seen that there were eight dimensions with eigenvalues greater than one. The scree plots obtained as a result of PA and parallel analysis are given below.

**Figure 1.** Scree plots obtained as a result of PA and parallel analysis.



When Figure 1 was examined, it was seen that the structure was four-dimensional according to the results of PA and parallel analysis. As a result of Varimax rotation items with a factor loading of less than 0.50 (7, 8, 9, 10, 14, 15, 17, 18, 32, 33, 34, 46, 51) and overlapping items (13, 16, 45) were removed from the scale. PA results for this structure are given in the table below.

**Table 3.** PA Varimax rotation results in Professional Obsolescence Scale factor loadings, eigenvalues, explained variance proportions.

Items	F1: Being Open to Professional Development	F2: Job-Ability Harmony in Profession	F3:Organizational Support in Professional Development	F4: Professional Burnout
M54	<b>0.850</b>	0.080	0.090	-0.020
M55	<b>0.820</b>	0.120	0.080	0.020
M58	<b>0.810</b>	0.200	0.080	0.060
M63	<b>0.810</b>	0.170	0.130	0.080
M53	<b>0.780</b>	0.080	0.230	-0.050
M57	<b>0.770</b>	0.150	0.160	0.050
M59	<b>0.770</b>	0.180	0.140	0.100
M61	<b>0.760</b>	0.250	0.030	0.070
M62	<b>0.740</b>	0.160	0.170	0.160
M56	<b>0.730</b>	0.090	0.140	0.060
M60	<b>0.700</b>	0.140	0.210	0.010
M50	<b>0.700</b>	0.210	0.050	0.110
M52	<b>0.680</b>	0.060	0.290	0.020
M48	<b>0.680</b>	0.260	0.120	0.000
M49	<b>0.660</b>	0.070	0.210	0.090
M26	<b>0.640</b>	0.100	0.170	0.010
M47	<b>0.620</b>	0.120	0.270	0.140



M24	<b>0.610</b>	0.100	0.040	0.150
M25	<b>0.610</b>	0.230	0.150	0.010
M39	<b>0.590</b>	0.290	0.080	0.160
M27	<b>0.580</b>	0.200	0.230	-0.010
M40	<b>0.540</b>	0.190	0.130	0.130
M38	<b>0.540</b>	0.220	-0.010	0.060
M4	0.190	<b>0.810</b>	0.080	0.050
M2	0.190	<b>0.800</b>	0.060	0.000
M3	0.190	<b>0.800</b>	0.180	-0.050
M1	0.170	<b>0.800</b>	0.080	-0.060
M6	0.350	<b>0.620</b>	0.140	-0.110
M11	0.400	<b>0.610</b>	0.010	-0.010
M5	0.240	<b>0.570</b>	0.280	-0.040
M12	0.320	<b>0.460</b>	0.230	-0.060
M29	0.260	0.140	<b>0.780</b>	-0.010
M30	0.150	0.120	<b>0.780</b>	0.010
M31	0.040	0.000	<b>0.720</b>	0.010
M28	0.280	0.200	<b>0.710</b>	-0.080
M36	0.130	0.060	<b>0.670</b>	-0.020
M35	0.290	0.190	<b>0.650</b>	-0.020
M37	0.360	0.130	<b>0.550</b>	0.120
M22	-0.010	-0.100	-0.090	<b>0.700</b>
M21	0.000	0.000	-0.050	<b>0.690</b>
M43	0.190	0.130	0.090	<b>0.640</b>
M19	-0.140	-0.160	-0.060	<b>0.610</b>
M42	0.170	0.080	0.150	<b>0.600</b>
M23	0.010	-0.080	-0.010	<b>0.570</b>
M20	0.000	-0.110	0.030	<b>0.560</b>
M41	0.230	0.120	0.040	<b>0.550</b>
M44	0.200	0.040	-0.060	<b>0.490</b>
Eigenvalues*	12.53	4.77	4.25	3.50
Explained Variance	27%	10%	9%	7%
Cronbach's Alpha	0.945	0.872	0.858	0.807
Stratified Alpha	0.940			

\* Four dimensions were verified as a result of the parallel analysis.

As seen in Table 3, the 47 items got arranged in a four-dimensional structure explaining 53% of the total variance. The items in each dimension were examined and the first dimension consisting of 23 items was named as “Being Open to Professional Development”, the second dimension consisting of eight items was named as “Job-Ability Harmony in Profession”, the third dimension consisting of seven items was named as “Organizational Support in Professional Development” and the fourth dimension of nine items was named as “Professional Burnout”. When the internal consistency of each dimension was examined, it was seen that the Cronbach Alpha coefficients were over 0.81 and the Stratified Alpha coefficient of the four-dimensional structure was 0.94. Therefore, it can be said that the scores have high reliability, according to Classical Test Theory (CTT). Definitions and sample items for each dimension are given below.

**Table 4.** Description of dimensions and sample items.

Dimensions	Description	Sample items
F1: Being Open to Professional Development	School administrators' institutions for their professional development should be open to participating in internal or external activities determined by their superiors or themselves for their professional development, and they plan for these activities.	M54: Your participation in a working group for professional development Very effective ... Not effective at all  M55: Conducting individual or joint research on a subject that interests you professionally Very effective ... Not effective at all
F2: Job-Ability Harmony in Profession	It expresses to what extent school administrators are aware of their learning abilities and how they use these skills in their professional development. Also, it indicates the compatibility of the knowledge, skills and abilities of school administrators with their job.	M1: How appropriate is your current professional knowledge for the job you are doing? Very suitable... Not suitable  M4: How do you compare the skills you have with the requirements of your job? Above what the job requires ... Below what the job requires
F3: Organizational Support in Professional Development	It expresses to what extent school administrators' superiors or institutions support their professional development, and to what extent they encourage administrators' high performance and innovative status.	M30: To what extent does your institution make long-term career planning of its managerial personnel? Very much ... Not at all  M29: To what extent do the policies at your institution encourage you to study at a more advanced level? Very much ... Not at all
F4: Professional Burnout	It refers to the personal, managerial and organizational factors that will cause the professional burnout of school administrators.	M21: How does the incompatibility of your working style with your organization's working style affect your job performance? Very much ... Not at all  M22: To what extent does your organization's management style conflict with your personality affect your job performance? Very much ... Not at all

The names given to the dimensions in Table 4 are based on the definitions and these definitions are made to cover all the items in the relevant dimension. Pearson's correlation coefficients between dimensions calculated based on raw scores are given below.

**Table 5.** Correlations between dimensions.

	F1	F2	F3	F4
F1	1.000			
F2	0.491*	1.000		
F3	0.410*	0.335*	1.000	
F4	0.112*	-0.047	-0.003	1.000

\* $p < 0.05$ .

When Table 5 is examined, it is seen that the Pearson correlation coefficients between the first three dimensions are greater than 0.30. Relationships between these dimensions are medium and positive. As the scores from one of these dimensions increase, the scores from the others

also increase. The relationship between the fourth dimension and the first dimension is significant and at a low level (Field, 2009). In that case, a single score cannot be obtained from the whole scale; each dimension will be evaluated within itself.

Before examining item and test parameters according to the Samejima's Graded Response Model based on IRT, dimensionality, local independence, model-data fit assumptions should be tested. Eigenvalues and ratios for dimensionality, variance proportion explained by a single factor, and C2 statistics for model-data fit were examined. Findings regarding these tests are given in the table below.

**Table 6.** *Statistics on assumptions for parameter estimates.*

	First eigenvalue	Second eigenvalue	Proportion of eigenvalues	Variences explained by the single factor (%)	C2	df	<i>p</i>
F1	12.919	1.295	9.97	56	2499.568	230	0.000
F2	4.944	0.910	5.43	61	348.682	20	0.000
F3	4.324	0.738	5.86	61	284.131	14	0.000
F4	4.016	1.505	2.67	44	1105.835	27	0.000

When Table 6 is examined, it is seen that the proportion of the first and second eigenvalues are more than 2.5 and the explained variances are greater than 30%. Therefore, each dimension is one-dimensional in itself (Çokluk et al., 2014). Parallel analysis results support this finding. The size number of the test equals the number of latent features that can be locally independent. On the other hand, local independence is the condition in which the score to be obtained from an item by individuals with the same ability level is not affected by other items (Embretson and Reise, 2000). Therefore, it can be said that since the unidimensionality assumption is met, the local independence assumption is also met (Crocker and Algina, 2006; Hambleton and Swaminathan, 1985). According to the results of the C2 statistics, it was also observed that the model-data fit was achieved according to Samejima's Graded Response Model ( $p < 0.05$ ). Parameters related to the items are given in the table below. Discrimination (a) parameters are classified as very high when it is 1.70 and above, according to Baker (2001). Although threshold parameters (b) are scaled between -3 and +3 in practice, theoretically, they take values between  $-\infty$  and  $+\infty$ . As the b value increases, the probability of an individual to mark the item in one of the higher categories increases with a 50% probability.

**Table 7.** *Discrimination and threshold parameters for the Graded Response Model.*

Dimensions	Items	Item parameters				
		a	b1	b2	b3	b4
F1	M54	3.439	-2.889	-2.223	-1.066	0.111
	M55	3.195	-3.057	-2.306	-1.089	0.118
	M63	3.285	-3.229	-2.481	-1.284	-0.053
	M53	2.690	-2.599	-2.025	-0.984	0.238
	M57	2.654	-2.655	-2.036	-0.901	0.326
	M59	2.863	-2.838	-2.375	-1.223	0.054
	M62	2.491	-3.222	-2.461	-1.157	0.167
	M56	2.247	-2.491	-1.966	-0.901	0.161
	M50	2.056	-3.286	-2.231	-0.929	0.501
	M58	3.331	-3.046	-2.785	-1.199	0.009
	M60	2.194	-3.687	-2.904	-1.468	-0.098
	M61	2.680	-3.654	-2.988	-1.497	-0.135
	M48	1.930	-2.924	-2.070	-0.855	0.585
	M52	2.110	-3.785	-3.039	-1.565	-0.063
M49	1.859	-2.982	-2.122	-0.789	0.539	

	M26	1.667	-2.941	-2.099	-0.973	0.256
	M47	1.737	-3.356	-2.493	-1.080	0.613
	M25	1.646	-4.003	-2.976	-1.519	-0.103
	M24	1.443	-3.705	-2.949	-1.663	-0.157
	M39	1.538	-4.527	-3.579	-1.819	-0.128
	M38	1.303	-3.225	-2.861	-1.733	-0.310
	M27	1.560	-2.839	-2.329	-1.262	0.273
	M40	1.261	-3.730	-3.442	-2.339	-0.803
	mean	2.225	-3.247	-2.554	-1.274	0.091
F2	M4	3.102	-3.217	-2.541	-1.380	0.332
	M2	3.027	-3.112	-2.498	-1.236	0.453
	M3	2.869	-3.278	-2.627	-1.444	0.025
	M1	2.986	-3.121	-2.511	-1.618	-0.032
	M6	1.730	-3.785	-2.864	-1.616	-0.023
	M11	1.790	-4.642	-3.897	-2.039	0.043
	M5	1.524	-4.194	-2.661	-1.397	0.427
	M12	1.239	-5.119	-3.718	-1.617	0.441
	mean	2.283	-3.809	-2.915	-1.543	0.208
F3	M29	3.585	-1.861	-1.177	-0.221	0.957
	M30	2.629	-1.672	-0.932	0.153	1.317
	M28	2.771	-2.081	-1.321	-0.345	0.848
	M31	1.582	-1.381	-0.485	0.476	1.858
	M36	1.513	-1.062	-0.699	0.017	0.824
	M35	1.874	-2.458	-1.655	-0.650	0.821
	M37	1.522	-3.234	-2.068	-0.585	1.098
	mean	2.211	-1.964	-1.191	-0.165	1.103
F4	M22	2.343	-1.687	-0.878	0.006	0.992
	M21	2.304	-1.829	-0.940	-0.082	0.916
	M43	1.351	-2.836	-1.675	-0.218	1.338
	M42	1.215	-2.683	-1.615	-0.089	1.459
	M19	1.319	-1.848	-0.456	0.705	2.027
	M23	1.382	-2.267	-1.065	0.178	1.627
	M20	1.329	-1.642	-0.705	0.260	1.456
	M41	1.108	-3.601	-2.239	-0.641	1.115
	M44	0.894	-2.489	-1.927	-0.752	0.314
	mean	1.472	-2.320	-1.278	-0.070	1.249

When [Table 7](#) is examined, it is seen that in the first dimension, item discrimination, in other words, the information given by the items about the structure varies between 1.261 (M40) and 3.439 (M54), and the average discrimination is 2.225. When the threshold parameters are examined, with a 50% probability, the average ability level required for individuals to mark categories higher than 1 instead of 1 is -3.247; the average ability level required for individuals to mark categories higher than 2 instead of 2 is -2.554; the average ability level required for individuals to mark categories higher than 3 instead of 3 is -1.274; the average ability level required for individuals to mark categories for 5 instead of 4 is 0.091. The marginal reliability coefficient for this dimension is 0.937; the empirical reliability coefficient is 0.938.

In the second dimension, it is seen that the information given by the items varies between 1.239 (M12) and 3,102 (M4), and the average discrimination is 2.283. When the threshold parameters are examined, with a 50% probability, the average ability level required for individuals to mark categories higher than 1 instead of 1 is -3.809; the average ability level required to mark categories higher than 2 instead of 2 is -2.915; the average ability level required to mark categories higher than 3 instead of 3 is -1.543; the average ability level required for individuals to mark categories for 5 instead of 4 is 0.208. The marginal reliability coefficient for this dimension is 0.857; the empirical reliability coefficient is 0.867.

In the third dimension, it is seen that the information given by the items varies between 1.513 (M36) and 3.585 (M29), and the average discrimination is 2.221. When the threshold parameters are examined, with a 50% probability, the average ability level required for individuals to mark categories higher than 1 instead of 1 is -1.964; the average ability level required for marking categories higher than 2 instead of 2 is -1.191; the average ability level required to mark categories higher than 3 instead of 3 is -0.165; the average ability level required for individuals to mark categories for 5 instead of 4 is 1.103. The marginal reliability coefficient for this dimension is 0.893; the empirical reliability coefficient is 0.894.

In the fourth dimension, it is seen that the information given by the items varies between 0.894 (M44) and 2.343 (M22), and the average discrimination is 1.472. When the threshold parameters are examined, with a 50% probability, the average ability level required for people to mark categories higher than 1 instead of 1 is -2.320; the average ability level required to mark categories higher than 2 instead of 2 is -1.278; the average ability level required to mark categories higher than 3 instead of 3 is -0.070; the average ability level required for individuals to mark categories for 5 markings instead of 4 is 1.249. The marginal reliability coefficient for this dimension is 0.850; the empirical reliability coefficient is 0.854.

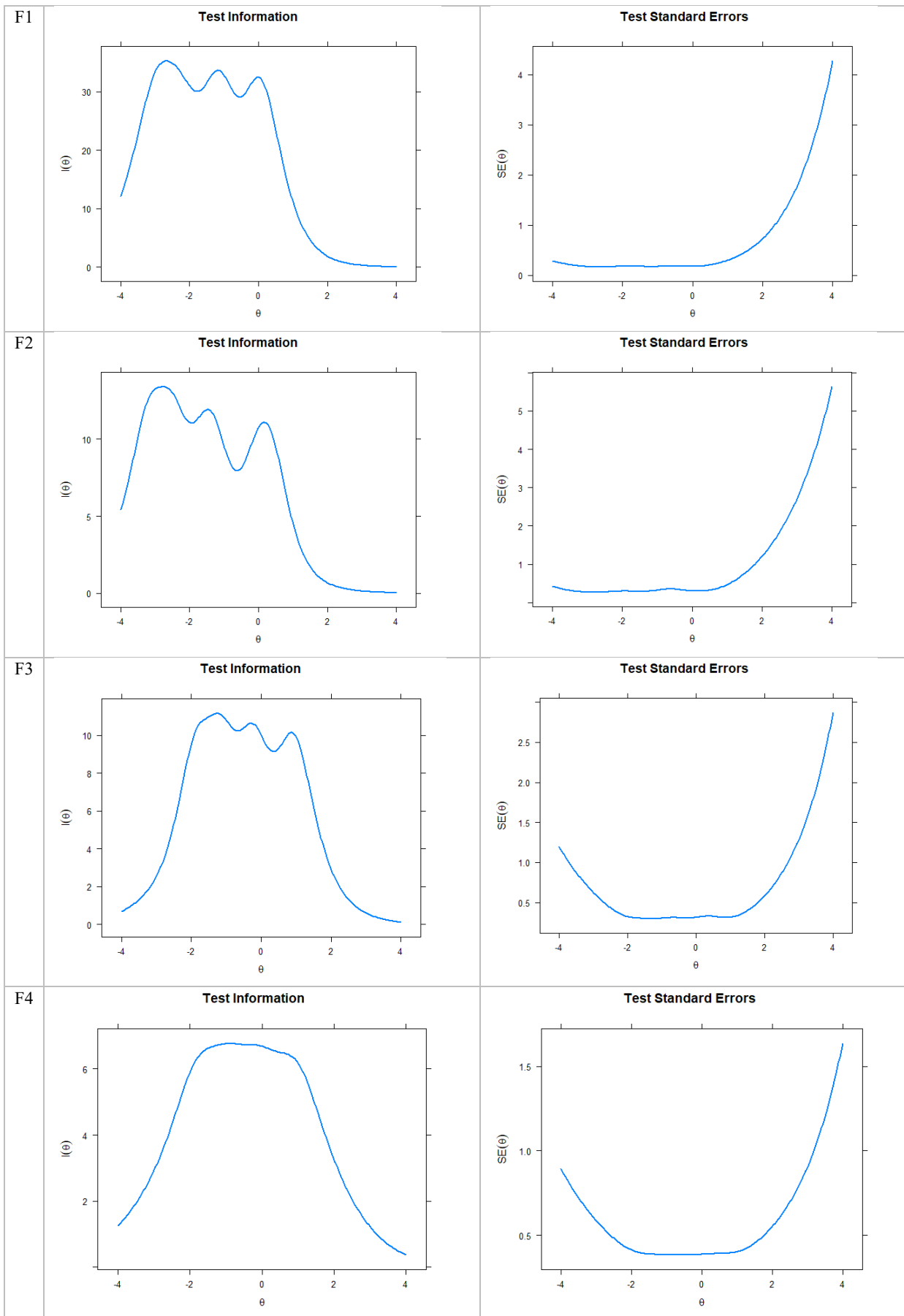
Test information function and standard errors for these dimensions are given below. When [Figure 2](#) is examined, it can be said that the structures measured in the first and second dimensions measure individuals in the  $-\infty$  to -1 ability range, and the structures measured with the third and fourth dimensions measure the individuals in the -2 to +2 ability range with relatively little error. As we move away from these ability ranges, it is seen that the standard errors of measurement increase.

Apart from this information, one of the superior features of IRT is parameter invariance. In other words, it is the estimation of item parameters independent from the sample, and ability parameters ( $\theta$ ) independently from item parameters. Providing parameter invariance is, therefore, a proof of validity (DeMars, 2010; Baker, 2016). For this purpose, the rows in the data set were divided into two groups as odd and even, and the correlation between item parameters obtained from these groups was calculated. Besides, the columns belonging to the items in the data set were divided into two groups as odd and even, and the correlation between the ability ( $\theta$ ) parameters obtained from these groups was calculated. Correlation coefficients for item and ability parameter invariance are given in [Table 8](#) below.

**Table 8.** Correlation coefficients for item and ability parameter invariance.

Dimensions	a	mean b	theta
F1	0.950	0.929	0.904
F2	0.948	0.978	0.781
F3	0.949	0.986	0.752
F4	0.833	0.989	0.608

**Figure 2.** Test information functions and standard errors regarding dimensions.





When the table is examined, it is seen that all correlation coefficients have medium and high correlation above 0.60. This finding indicates that item parameters can be estimated independently from the group and ability parameters can be predicted independently from the items. Finally, the number of materials, the lowest and highest scores and average values of the four-dimensional structure reached in Table 9 are given.

**Table 9.** Descriptive statistics on POS sub-dimensions.

Dimensions	Item numbers	Min.	Max.	Mean	Mean (Five point likert scale)
F1: Being Open to Professional Development	23	23	115	97.442	4.237
F2: Job-Ability Harmony in Profession	8	8	40	34.308	4.288
F3: Organizational Support in Professional Development	7	7	35	24.231	3.461
F4: Professional Burnout	9	9	45	30.806	3.423

When Table 9 is examined, it is seen that the average scores obtained from each dimension vary between 3.423 and 4.288 out of five. The averages of the third and fourth dimensions are relatively lower than the first two dimensions. In that case, it can be stated that managers perceive more professional obsolescence in “Organizational Support in Professional Development” and “Professional Burnout”.

#### 4. DISCUSSION and CONCLUSION

This research aimed to develop a scale that will determine the factors leading to professional obsolescence in the field of education. As a result of the analyses made in this context, a four-dimensional scale consisting of 47 items was developed to show the professional antiquity of school administrators. The dimensions of the scale items that will determine the professional obsolescence of school administrators were found as “Being Open to Professional Development”, “Job-Ability Harmony in Profession”, “Organizational Support in Professional Development”, “Professional Burnout”. A total score is not obtained from these dimensions, and each dimension is evaluated descriptively in itself. Low scores in the first three dimensions and high scores in the last dimension show that professional obsolescence is high. In other words, not being open to professional development, lack of or low adaptation to work skills in the profession, lack of or low organizational support in professional development, high Professional Burnout indicate professional obsolescence. It is seen that the items in “Professional Burnout”, one of the dimensions that emerged in this study, are the personal factors in organizational obsolescence stated by Bařaran (2008), in a single group. Again, one of the dimensions in this scale, “Organizational Support in Professional Development” was also proposed as a dimension in the scale developed by Chauhan and Chauhan (2009), and a similar grouping was observed here. Another dimension that emerged in this study is the dimension of “Job -Ability Harmony in Profession”, again Chauhan and Chauhan (2009) show similarities with the dimension of “professional knowledge and skills”, which is a factor in professional obsolescence. The dimension of “Being Open to Professional Development” comprises a combination of several dimensions related to professional development in the scale developed by Chauhan & Chauhan (2009).

Professional development activities are an important factor that prevents knowledge and ability obsolescence. The averages of the dimensions show that school administrators are quite good in the dimension of “Being Open to Professional Development” and it is consistent with the findings of participating in professional development activities above. In the dimension of “Job-

Ability Harmony in Profession”, again, it can be said that managers find their knowledge and ability suitable for their job. However, it can be said that “Organizational Support in Professional Development” is not perceived very high among these dimensions. The fact that the “Professional Burnout” is above average may indicate that school administrators are prone to professional obsolescence.

As a result, this four-dimensional scale developed to measure the professional antiquity of school administrators can be used both to determine whether school administrators are open to professional development and to determine the level of competence of their knowledge and ability while performing their profession. Also, the scale developed can be used both to determine the organizational support of school administrators in professional development and to determine the factors that will cause them to become obsolete. Moreover, the variables (age, education level, innovation, change, etc.) that are stated in the professional obsolescence scale developed and other variables that are related to professional obsolescence can be studied.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s). **Ethics Committee Number:** Başkent University/Faculty of Education, 17162298.600-291.

### Authorship Contribution Statement

**Sadegül Akbaba Altun:** Investigation, Resources, Introduction, Methodology, Findings, Discussion, Supervision, and Validation. **Şener Büyüköztürk:** Investigation, Resources, Methodology, Supervision, and Validation. **Merve Yıldırım Seheryeli:** Method, Software, Analysis, Findings.

### ORCID

Sadegül Akbaba Altun  <https://orcid.org/0000-0001-5690-6088>

Şener Büyüköztürk  <https://orcid.org/0000-0002-0898-1697>

Merve Yıldırım Seheryeli  <https://orcid.org/0000-0002-1106-5358>

## 5. REFERENCES

- Akbaş, U., Karabay, E., Yıldırım-Seheryeli, M., Ayaz, A., & Demir, Ö. O. (2019). Comparison of Exploratory Factor Analysis Studies in Turkish Measurement Tools Index According to Parallel Analysis Results. *Journal of Theoretical Educational Science*, 12(3), 1095-1123. <https://doi.org/10.30831/akukeg.453786>
- Baker, F. B. (2016). *Madde Tepki Kuramının Temelleri [Fundamentals of Item Response Theory]*. (N. Güler, Ed., & M. İlhan, Çev.) Pegem Akademi.
- Başaran, İ.E. (2008). *Örgütsel Davranış [Organizational Behavior]*. Ekinoks Yayınları.
- Burke, R.J. (1969). Effects of aging on engineer’s satisfactions and mental health: Skill obsolescence, *Academy of Management Journal*, 12(4), 479-486. <https://doi.org/10.5465/254736>
- Büyüköztürk, Ş. (2019). *Sosyal Bilimler İçin Veri Analizi El Kitabı [Manual of Data Analysis for Social Sciences]* (26. baskı). Pegem Akademi.
- Chauhan, S.P., & Chauhan, D. (2004). Professional Obsolescence: Causes and Preventive Measures. *Indian Journal of Industrial Relations*, 39(3), 347-363. <http://www.jstor.org/stable/27767911>
- Chauhan, S.P., & Chauhan, D. (2005). Overcoming Managerial Obsolescence: The key to human development. Pritam Singh, Ajay Singh, and Daisy Chauhan, (Eds.), *in Creating Value Through People* (pp. 321-38). Excel Books.

- Chauhan, S.P., & Chauhan, D. (2008). Human Obsolescence: A Wake-up Call to Avert a Crisis. *Global Business Review*, 9(1), 85-100. <https://doi.org/10.1177/097215090700900106>
- Chauhan, S.P., & Chauhan, D. (2009). Are you on the verge of obsolescence? *Indian Journal of Industrial Relations*, 44(4), 646-659.
- Crocker, L., & Algina, J. (2006). *Introduction to Classical and Modern Test Theory*. Cengage Learning.
- Cronbach, L. J., Schonemann, P., & McKie, D. (1965). Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement*, 25, 291-312. <https://doi.org/10.1177/001316446502500201>
- Cronbach, L. J. (1984). *Essentials of Psychological Testing* (4th ed). Harper Row.
- Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2014). *Sosyal Bilimler İçin Çok Değişkenli İstatistik SPSS ve LISREL Uygulamaları [Multivariate Statistics for Social Sciences: SPSS and LISREL Applications]*. Pegem Akademi.
- DeMars, C. (2010). *Item Response Theory*. Oxford University Press.
- Embretson, S.E., & Reise, S.P. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates.
- Field, A. (2009). *Discovering Statistics Using SPSS* (3rd ed.). Sage Publications Ltd.
- Fossum, J. A., Arvey, R.D., Paradise, C.A., & Robbins, N. E. (1986). Modeling the Skills Obsolescence Process: A Psychological/Economic Integration, *Academy of Management Review*, 11(2), 362-374.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Kluwer Nijhoff Publishing.
- Jones, E., Chonko, L.B., & Roberts, J.A. (2003). Sales for obsolescence: Perceptions from sales and marketing executives of individual, organizational, and environmental factors. *Industrial Marketing Management*, 33, 439-456.
- Knight, P. (1998). Professional obsolescence and continuing professional development in higher education, *Innovations in Education and Teaching International*, 35(3), 248-256. <https://doi.org/10.1080/1355800980350309>
- MacNeill, N., & Cavanagh, R. (2006). Principals' pedagogic obsolescence: re-assessing what is important in schools. *Curriculum and Leadership Journal*, 4(29). [http://www.curriculum.edu.au/leader/principals\\_pedagogic\\_obsolescence\\_re-assessing\\_wh,15968.html?issueID=10421](http://www.curriculum.edu.au/leader/principals_pedagogic_obsolescence_re-assessing_wh,15968.html?issueID=10421)
- Millî Eğitim Bakanlığı (MEB) Eğitim Kurumları Yöneticilerinin Atama ve Yer Değiştirmelerine İlişkin Yönetmelik. [Regulation on Appointment and Relocation of Educational Institutions Administrators.] (2009, Ağustos 3). *Resmi Gazete* (Sayı 27318). <https://www.resmigazete.gov.tr/eskiler/2009/08/20090813-2.htm>
- Mohan, V., Chauhan, S.P., & Chauhan, D. (2001). Are you aware how your personality effects your behavior? *Global Business Review*, 2(2), 289-304. <https://doi.org/10.1177/097215090100200209>
- Murillo, P. (2011). Human capital obsolescence: some evidence for Spain, *International Journal of Manpower*, 32(4), 426-445. <https://doi.org/10.1108/01437721111148540>
- Pazy, A. (1994). Cognitive Schemata of Professional Obsolescence, *Human Relations*, 47(10), 1167.
- Pazy, A. (1996). Concept and career-stage differentiation in obsolescence research. *Journal of Organizational Behavior*, 17(1), 59-78.
- Rothman, R.A., & Perrucci, R. (1971). Vulnerability to Knowledge Obsolescence: Among Professionals, *The Sociological Quarterly*, 12, 147-138.
- Shaffer, J. D. (1969). On institutional obsolescence and innovation- Background for Professional dialogue on public policy. *American Journal of Agricultural Economics*, 51(2), 245-267.

- Shearer, R., & Steger, J. A. (1975). Manpower Obsolescence: A New Definition and Empirical Investigation of Personal Variables, *Academy of Management Journal*, 18, 263- 275.
- Sorensen, J. B., & Stuart, T. E. (2000). Aging, obsolescence, and organizational innovation. *Administrative Science Quarterly*, 45(1), 81-112.
- Tan, Ş. (2009). Misuses of KR-20 and Cronbach's alpha reliability coefficients. *Journal of Education and Science*, 34(152), 101-112.
- Tang, K. (1996). *Polytomous Item Response Theory (IRT) Models and Their Applications in Large-Scale Testing Programs: Review of Literature*. Educational Testing Service.
- Todman, J., & Dugard, P. (2007). *Approaching multivariate statistics: an introduction for psychology*. Psychology Press.
- Toner, P. (2011), *Workforce Skills and Innovation: An Overview of Major Themes in the Literature*, OECD Science, Technology and Industry Working Papers, 2011/1. OECD Publishing. <https://doi.org/10.1787/5kgkdgdkc8tl-en>
- van Loo, J., Grip, A., & Steur, M. (2001). Skills obsolescence: causes and cures. *International Journal of Manpower*, 22(1/2), 121-138.
- Wandera, D. B. (2014). The threat of obsolescence: teaching and learning responding to technology. *Technology, Pedagogy and Education*, 24(2), 279-281. <https://doi.org/10.1080/1475939X.2014.913533>
- Warmington, A. (1974). Obsolescence As An Organizational Phenomenon, *The Journal of Management Studies*, 11 (2), 96-114.
- Yürekli, H. (2010). *The Relationship Between Parameters from Some Polytomous Item Response Theory Models*. (Unpublished Master Thesis). The Florida State University, USA. [http://purl.flvc.org/fsu/fd/FSU\\_migr\\_etd-1104](http://purl.flvc.org/fsu/fd/FSU_migr_etd-1104)

## 6. APPENDIX

The items and the dimensions of the scale developed to determine the Professional and Organizational Obsolescence levels of primary, secondary and high school principals are given below (Turkish version of the POS).

Mesleki Eskimişlik Ölçeği					
<b>F1: Mesleki Gelişime Açık Olmak</b>					
M24	Sizin belirleyeceğiniz bir zaman aralığında eğitim almak				
	Çok etkili				Hiç etkili değil
	5	4	3	2	1
M25	Mesleğinizle ilgili kitap ve dergileri okumak				
	Çok etkili				Hiç etkili değil
	5	4	3	2	1
M26	Mesleğinizle ilgili Sivil Toplum Kuruluşları (Eğitim Yönetimi Derneği vb.) toplantılarına katılmak				
	Çok etkili				Hiç etkili değil
	5	4	3	2	1
M27	Beş yıl sonrası için mesleğinizle ilgili planlama yapma				
	Çok etkili				Hiç etkili değil
	5	4	3	2	1
M38	Kurumunuzda mesleki becerileri artıracak bir gelişim planının olması sizin kendinizi geliştirmenizi ne oranda motive eder?				
	Çok fazla				Hiç
	5	4	3	2	1
M39	Sizin kaydettiğiniz mesleki gelişim kurumumuz tarafından takdir edilmesi sizi ne oranda motive eder?				
	Çok fazla				Hiç
	5	4	3	2	1
M40	Mesleğinizi daha cazip hale getirmeye yönelik değişiklikler sizin mesleki gelişiminizi ne oranda etkiliyor?				
	Çok fazla				Hiç
	5	4	3	2	1
M47	Bir üst yöneticinizin sizin becerilerinizi nasıl geliştirebileceğiniz konusunda öneriler getirmesi sizin becerilerinizi ne oranda geliştirir?				
	Çok fazla				Hiç
	5	4	3	2	1
M48	Kurumu içi eğitim programları				
	Çok etkili				Hiç etkili değil
	5	4	3	2	1
M49	Meslekle ilgili kurum dışı eğitim programları				
	Çok etkili				Hiç etkili değil
	5	4	3	2	1
M50	Diğer kurumlara kendi kurumu adına ziyarette bulunmak				
	Çok etkili				Hiç etkili değil
	5	4	3	2	1
M52	İş başında problem çözme				
	Çok etkili				Hiç etkili değil
	5	4	3	2	1
M53	Kurumunuz tarafından düzenlenen ya da desteklenen seminer ve konferanslara katılma				
	Çok etkili				Hiç etkili değil
	5	4	3	2	1
M54	Sizin mesleki gelişim için oluşturulmuş bir çalışma grubuna katılmanız				
	Çok etkili				Hiç etkili değil

	5	4	3	2	1
M55	Mesleki anlamda ilginizi çeken bir konuda bireysel ya da ortak araştırma yapmak				
	Çok etkili				Hiç etkili değil
	5	4	3	2	1
M56	Diploma/sertifika programlarına (Örneğin lisansüstü programlar) katılmak				
	Çok etkili				Hiç etkili değil
	5	4	3	2	1
M57	Kurum gelişimi çalışmalarında (örneğin, müfredat geliştirme grubunda, okul gelişim ekibinde) görev alma				
	Çok etkili				Hiç etkili değil
	5	4	3	2	1
M58	Kendinizi geliştirme etkinlikleri mesleğinizi etkili olarak yürütmeniz için ne kadar önemlidir?				
	Çok önemli				Hiç önemli değil
	5	4	3	2	1
M59	Kendini geliştirmeye yönelik etkinlikleri sizin kariyer gelişiminiz için ne kadar önemlidir?				
	Çok önemli				Hiç önemli değil
	5	4	3	2	1
M60	Yönetimde Bilgi ve İletişim Teknolojilerini Kullanma becerilerinin öğretilmesi sizin işinizi etkili olarak yürütmenize ne oranda katkı sağlar?				
	Çok katkı sağlar				Hiç katkı sağlamaz
	5	4	3	2	1
M61	Çalışanların mesleki gelişim ihtiyaçlarının belirlenmesi çalışanların kendilerini geliştirmelerin ne kadar önemlidir?				
	Çok önemli				Hiç önemli değil
	5	4	3	2	1
M62	Kurumunuzun, sizin veya diğer personelin mesleki alandaki zayıf yönlerini iyileştirmeye yönelik gelişim planı oluşturmaları sizin kariyer gelişiminizde ne kadar önemlidir?				
	Çok önemli				Hiç önemli değil
	5	4	3	2	1
M63	Kurumunuzun, size meslekî gelişim etkinliklerine katılım imkânını sağlaması ne kadar önemlidir?				
	Çok önemli				Hiç önemli değil
	5	4	3	2	1
<b>F2: Meslekte İş Yetenek Uyumu</b>					
M1	Şu anda sahip olduğunuz mesleki bilgi yaptığınız iş için ne kadar uygun?				
	Çok uygun				Uygun değil
	5	4	3	2	1
M2	Sahip olduğun bilgiyi yaptığınız işle karşılaştırdığınızda nasıl görüyorsunuz?				
	İşin gerektirdiğinin üstünde				İşin gerektirdiğinin altında
	5	4	3	2	1
M3	Şu anda sahip olduğunuz becerilerin yaptığınız iş ile uygunluk derecesi nedir?				
	Çok uygun				Hiç uygun değil
	5	4	3	2	1
M4	Sahip olduğunuz becerileri yaptığınız işin gerektirdikleriyle karşılaştırdığınızda nasıl görüyorsunuz?				
	İşin gerektirdiğinin üstünde				İşin gerektirdiğinin altında
	5	4	3	2	1
M5	Şu andaki işinizde mesleki becerilerinizden/yeteneklerinizden ne oranda yararlanılıyor?				
	Çok fazla				Hiç
	5	4	3	2	1
M6	Kendinizi güncel tutmaya ilişkin motivasyonunuzu nasıl görüyorsunuz?				



	Çok yüksek				Çok düşük
	5	4	3	2	1
M11	Genel olarak, işinizle ilgili bilgi/beceri öğrenme yeteneğinizi nasıl görüyorsunuz?				
	Çok yüksek				Çok düşük
	5	4	3	2	1
M12	Öğrenme yeteneğiniz son beş yılda ne oranda değişti?				
	Çok gelişti				Hiç gelişmedi
	5	4	3	2	1
<b>F3: Mesleki Gelişimde Örgütsel Destek</b>					
M28	Kurumunuz kendinizi geliştirmenize yönelik eğitim almanızı kolaylaştırıyor mu?				
	Çok Fazla				Hiç
	5	4	3	2	1
M29	Kurumunuzdaki politikalar, sizin daha ileri bir düzeyde eğitim almanızı ne oranda teşvik ediyor?				
	Çok Fazla				Hiç
	5	4	3	2	1
M30	Kurumunuz kendi yönetici personelinin uzun süreli kariyer planlamasını ne oranda yapıyor?				
	Çok Fazla				Hiç
	5	4	3	2	1
M31	Yüksek performans kurumunuz tarafından ne oranda ödüllendiriliyor?				
	Çok Fazla				Hiç
	5	4	3	2	1
M35	Kurumunuz sizi daha yenilikçi olmanız konusunda ne oranda cesaretlendiriyor?				
	Çok Fazla				Hiç
	5	4	3	2	1
M36	Size göre, işinize ilgili olsun veya olmasın, bir üst amiriniz gelişiminize ne kadar ilgi gösteriyor?				
	Çok Fazla				Hiç
	5	4	3	2	1
M37	Kurumunuzda yöneticilerin mesleki ilerlemelerinin belirlenmesi, yöneticilerin kendilerini geliştirmelerini ne oranda sağlar?				
	Çok Fazla				Hiç
	5	4	3	2	1
<b>F4: Mesleki Tükenmişlik</b>					
M19	Yaşadığımız duygusal sorunlar (stres, hayal kırıklığı vb.) sizin mesleki gelişiminizi ne oranda etkilemektedir?				
	Çok Fazla				Hiç
	5	4	3	2	1
M20	Sağlık durumunuz sizin mesleğinizi etkili yapmada ne oranda etkilidir?				
	Çok Fazla				Hiç
	5	4	3	2	1
M21	Sizin çalışma biçiminizin kurumunuzun çalışma biçimiyle uyumsuzluğu, sizin iş performansınızı nasıl etkiler?				
	Çok Fazla				Hiç
	5	4	3	2	1
M22	Kurumunuzun yönetim biçiminin sizin kişiliğinizle çatışması sizin iş performansınızı ne oranda etkiler?				
	Çok Fazla				Hiç
	5	4	3	2	1
M23	Yeteneğinizin üstünde kurumunuzda size bir görev verilmesi sizin performansınızı ne oranda etkiler?				
	Çok Fazla				Hiç
	5	4	3	2	1

M41	Kurumunuzun amaçlarının belirsiz olması, sizin işinize karşı tutumunuzu ne oranda etkiliyor?				
	Çok Fazla				Hiç
	5	4	3	2	1
M42	Sizin üstünüzün sizin bilgi ve becerinizdeki eskimişliği önemsememesi sizin kendinizi geliştirmenizi ne oranda etkiliyor?				
	Çok Fazla				Hiç
	5	4	3	2	1
M43	Kurumunuzda sizin işinizle ilgili performansınıza yönelik dönüt mekanizmasının yetersiz olması sizin işinizi etkili yapmanızı ne oranda etkiler?				
	Çok Fazla				Hiç
	5	4	3	2	1
M44	Kurumunuzda, etkili olacağınız bir göreve atanmamış olmanız sizlerin bilgi ve becerilerinizi geliştirmenizi ne oranda etkiler?				
	Çok Fazla				Hiç
	5	4	3	2	1

## Point and Interval Estimators of an Indirect Effect for a Binary Outcome

Hyung Rock Lee <sup>1,\*</sup>, Jaeyun Sung <sup>2</sup>, Sunbok Lee <sup>3</sup>

<sup>1</sup>University of Central Arkansas, Department of Exercise & Sport Science, Conway, AR U.S.A.

<sup>2</sup>Lyon College, Department of Political Science, Batesville, AR U.S.A.

<sup>3</sup>Ewha Womans University, Department of Education, Seoul, South Korea

### ARTICLE HISTORY

Received: July 24, 2020

Revised: Feb. 15, 2021

Accepted: Mar. 05, 2021

### Keywords:

Indirect effects,  
Binary outcome,  
Confidence intervals,  
Bootstrap,  
Delta methods.

**Abstract:** Conventional estimators for indirect effects using a difference in coefficients and product of coefficients produce the same results for continuous outcomes. However, for binary outcomes, the difference in coefficient estimator systematically underestimates the indirect effects because of a scaling problem. One solution is to standardize regression coefficients. The residual from a regression of a predictor on a mediator, which we call the residualized variable in this paper, was used to address the scaling problem. In simulation study 1, different point estimators of indirect effects for binary outcomes are compared in terms of the means of the estimated indirect effects to demonstrate the scaling problem and the effects of its remedies. In simulation study 2, confidence and credible intervals of indirect effects for binary outcomes were compared in terms of powers, coverage rates, and type I error rates. The bias-corrected (BC) bootstrap confidence intervals performed better than did other intervals.

## 1. INTRODUCTION

Mediation analysis tests hypotheses about the mechanism through which a focal independent variable influences an outcome of interest. In mediation analysis, a third intermediate variable named a mediator accounts for the relationship between the independent variable and the outcome, and the effect of the independent variable on the outcome via the mediator is referred to as an indirect effect (Baron & Kenny, 1986). In the literature, two estimators have been widely used to estimate the indirect effect: the difference in coefficients of two nested regression models (Clogg, Petkova, & Shihadeh, 1992; Freedman & Schatzkin, 1992) and the product of coefficients in a path model (Alwin & Hauser, 1975; Bollen, 1987; Sobel, 1982).

For a given sample, the estimates of the two estimators are exactly the same when the outcome is continuous (MacKinnon, Warsi, & Dwyer, 1995). However, when the outcome is binary, the estimates from the two estimators are not the same (Breen, Karlson, & Holm, 2013; MacKinnon, Lockwood, Brown, Wang, & Hoffman, 2007). For a binary outcome, the difference in coefficients estimator underestimates the indirect effect because the regression coefficients of two nested probit or logit models are estimated in different scales (Allison, 1999; Karlson, Holm, & Breen, 2012; Winship & Mare, 1983). One solution to the scaling problem

---

\*CONTACT: Hyung Rock LEE ✉ [rlee@uca.edu](mailto:rlee@uca.edu) 📧 University of Central Arkansas, Department of Exercise & Sport Science, Conway, AR U.S.A

is to use standardized regression coefficients for the difference in coefficients estimator (MacKinnon & Dwyer, 1993; Winship & Mare, 1983). Breen et al. (2013) proposed another solution in which a residualized variable was used to address the scaling issue in the use of the difference in coefficients estimator for a binary outcome. Traditionally, confidence or credible intervals based on the delta (Sobel, 1982), bootstrap (Bollen & Stine, 1990; MacKinnon, Lockwood, & Williams, 2004), and Bayesian methods (Yuan & MacKinnon, 2009) have been widely used to make statistical inference about indirect effects. Previous studies on the indirect effect for a continuous outcome showed that the normality assumption about the sampling distribution might not be valid in small samples in which the true sampling distribution is asymmetric (Bollen & Stine, 1990; MacKinnon et al., 2004). Given the various methods researchers may choose for testing indirect effects for a binary outcome, the performances of those methods are not fully compared yet.

This study aims to compare various point and interval estimators of the indirect effect for a binary outcome using Monte Carlo simulation studies. The point estimators in our study include the conventional difference in coefficients estimator, the difference in coefficients estimator with standardized regression coefficients, the difference in coefficients estimator with residualized variables, and the product of coefficients estimator. Also, the interval estimates or confidence intervals of the indirect effects obtained using the delta, bootstrap, and Bayesian methods are also of interest. In simulation study 1, the point estimators were compared in terms of the means of estimated indirect effects across replications. In simulation study 2, the confidence intervals based on the delta, bootstrap, and Bayesian methods were compared in terms of powers, type I error rates, and coverage rates. We first present the scaling problem using the difference in coefficients estimator for a binary outcome. Then, we describe two solutions to the scaling problem. The delta, bootstrap, and Bayesian methods are briefly introduced before the method section, in which more details on the Monte Carlo simulation studies are presented.

### **1.1. The Scaling Problem in Estimating an Indirect Effect for a Binary Outcome**

The indirect effects for binary outcomes are frequently of interest in social science. For example, in prevention studies, the outcomes of interest are often binary variables such as heart disease or drug use incidence. Since prevention programs are typically designed to change some mediating constructs that are assumed to be related to the outcomes of interest, the success of prevention programs can be evaluated by testing the indirect effect of prevention programs on binary outcomes via mediating constructs (MacKinnon & Dwyer, 1993; MacKinnon et al., 2007).

Binary outcomes in mediation analysis can be modeled using probit or logit regressions. However, the indirect effect estimated by the difference in probit or logit regression coefficients can be inaccurate because the coefficients of two nested probit or logit regressions are measured in different scales and therefore are not directly comparable (Allison, 1999; Karlson et al., 2012; Winship & Mare, 1983). For a more detailed discussion of the scaling issue, let us consider the following simple mediation model in which the latent response variable  $y^*$  is used to model a binary outcome:

$$y^* = \beta_1 + \beta_{yx}x + e_1, \tag{1}$$

$$y^* = \beta_2 + \beta_{yx.m}x + \beta_{ym.x}m + e_2, \tag{2}$$

$$m = \beta_3 + \beta_{mx}x + e_3, \tag{3}$$

where  $y^*$  is a continuous latent response variable,  $\beta_{yx}$  is the total effect of  $x$  on  $y^*$ ,  $\beta_{yx.m}$  is the direct effect of  $x$  on  $y^*$  net of  $m$ ,  $\beta_{ym.x}$  is the direct effect of  $m$  on  $y^*$  net of  $x$ , and  $e_1$ ,  $e_2$ , and

$e_3$  represent error terms. In the latent response variable formulation, a continuous latent response variable  $y^*$  is introduced to represent the propensity of the occurrence of a certain category in a categorical outcome. Then, a categorical outcome is considered an observed categorical indicator of an unobserved continuous latent response variable (Muthén, 1979, 1984). For a binary outcome  $y$ , a continuous latent response variable  $y^*$  is related to the binary outcome  $y$  via a threshold  $\tau$  as follows:

$$y = 1 \text{ if } y^* > \tau \text{ or } 0 \text{ if } y^* \leq \tau, \tag{4}$$

where the threshold  $\tau$  is typically assumed to be zero for an identification purpose. Note that, in Equations 1 and 2, the specific form of the model for a binary outcome, i.e., a probit or logit model, is determined by the distribution of an error term: normally distributed error terms result in probit models, and logistically distributed error terms result in logit models (Winship & Mare, 1983).

The scaling issue in estimating indirect effects using probit or logit models can be illustrated by considering the relationship between the regression coefficients in a latent response variable formulation and those in probit or logistic models (Allison, 1999; Breen et al., 2013; Karlson et al., 2012). To examine the relationship, let us assume that the error distributions in Equations 1 and 2 follow normal distributions. That is,  $e_1 = \sigma_1 u$  and  $e_2 = \sigma_2 u$ , where  $u$  is a random variable following a standard normal distribution, and  $\sigma_1$  and  $\sigma_2$  are scale factors. Then, the probit model for Equation 1 can be described as follows:

$$g[\Pr(y = 1|x)] = g[\Pr(y^* > 0|x)] \tag{5}$$

$$= g \left[ \Phi \left( \frac{E(y^*|x)}{\sqrt{V(y^*|x)}} \right) \right] \tag{6}$$

$$= \frac{E(y^*|x)}{\sqrt{V(y^*|x)}} \tag{7}$$

$$= \frac{\beta_1 + \beta_{yx}x}{\sigma_1} \tag{8}$$

$$= b_1 + b_{yx}x, \tag{9}$$

where  $g$  is the probit link function or the inverse of the cumulative distribution function of a standard normal distribution,  $\Phi$  is the cumulative distribution function of a standard normal distribution, and  $b_1$  and  $b_{yx}$  are the regression coefficients in the probit model. Similarly, the probit model for Equation 2 can be expressed as the following equation:

$$g[\Pr(y = 1|x, m)] = \frac{\beta_2 + \beta_{yx.m}x + \beta_{ym.m}m}{\sigma_2} = b_2 + b_{yx.m}x + b_{ym.m}m. \tag{10}$$

Then, by comparing the regression coefficients from the latent response variable formulation in Equations 1 and 2, and those from the probit model in Equations 9 and 10, we obtain the following equations:

$$b_{yx} = \frac{\beta_{yx}}{\sigma_1}, \tag{11}$$

$$b_{yx.m} = \frac{\beta_{yx.m}}{\sigma_2}, \tag{12}$$

$$b_{yx} - b_{yx.m} = \frac{\beta_{yx}}{\sigma_1} - \frac{\beta_{yx.m}}{\sigma_2} \neq \beta_{yx} - \beta_{yx.m}. \tag{13}$$

By using logistically distributed error terms and logit link functions, it can be shown that Equations 11, 12, and 13 are also valid for a logit model. Notice that the probit or logit regression coefficients  $b_{yx}$  and  $b_{yx.m}$  in Equations 11 and 12 involve different scale factors  $\sigma_1$  and  $\sigma_2$ , which implies that the coefficients of two nested probit or logit models are not directly comparable because they are measured in different scales. Furthermore, because the model in Equation 2 has an additional variable  $m$ , the residual variance of the model in Equation 2 should be smaller than that of the model in Equation 1, i.e.,  $\sigma_2 \leq \sigma_1$ . Therefore, the difference in probit or logit regression coefficients in Equation 13, i.e.,  $b_{yx} - b_{yx.m}$ , would underestimate the true amount of an indirect effect, i.e.,  $\beta_{yx} - \beta_{yx.m}$  (Breen et al., 2013; MacKinnon et al., 2007).

## 1.2. The Solutions to the Scaling Problem

One solution to the scaling problem is to make the scale equivalent across nested models by standardizing regression coefficients before estimating indirect effects (MacKinnon, 2008; MacKinnon & Cox, 2012; Winship & Mare, 1983). A residualized variable was recently used to address the scaling problem (Breen et al., 2013; Karlson et al., 2012). Those two approaches are briefly illustrated in this section.

### 1.2.1. Standardized Coefficients

In order to make the scale equivalent or comparable across two nested probit or logit models, Winship and Mare (1983) suggested to standardize regression coefficients using the variance of a latent response variable  $y^*$ . For probit models, the variances of  $y^*$  in Equations 1 and 2 can be obtained using the following equations (MacKinnon, 2008):

$$Var[y^*] = b_{yx}^2 Var[x] + 1. \tag{14}$$

$$Var[y^*] = b_{yx.m}^2 Var[x] + b_{ym.x}^2 Var[m] + 2b_{yx.m}b_{ym.x}Cov[x, m] + 1 \tag{15}$$

For logit models, the constant 1 in Equations 14 and 15 needs to be replaced by  $\pi^2/3$ , which is the variance of the standard logistic distribution. Then, the standardized coefficients can be obtained by dividing probit or logit coefficients in Equations 9 and 10 by the square root of the variances of  $y^*$  in Equations 14 and 15, and the indirect effect using the standardized coefficients can be expressed as follows:

$$\bar{b}_{yx} - \bar{b}_{yx.m} = \frac{b_{yx}}{\sqrt{Var[y^*]}} - \frac{b_{yx.m}}{\sqrt{Var[y^*]}}, \tag{16}$$

where  $\bar{b}_{yx}$  and  $\bar{b}_{yx.m}$  represent the standardized regression coefficients for  $b_{yx}$  and  $b_{yx.m}$ , respectively. The standard error of  $\bar{b}_{yx} - \bar{b}_{yx.m}$ , which is needed for statistical inferences, can be expressed as the following equation

$$SE[\bar{b}_{yx} - \bar{b}_{yx.m}] = \sqrt{SE[\bar{b}_{yx}]^2 + SE[\bar{b}_{yx.m}]^2 - 2Cov[\bar{b}_{yx}, \bar{b}_{yx.m}]}, \tag{17}$$

where  $SE[\bar{b}_{yx}] = SE[b_{yx}]/\sqrt{Var[y^*]}$  and  $SE[\bar{b}_{yx.m}] = SE[b_{yx.m}]/\sqrt{Var[y^*]}$ . For logit models,  $Cov[\bar{b}_{yx}, \bar{b}_{yx.m}]$  can be obtained using the formula described in Freedman and Schatzkin (1992). However, we are unaware of any analytical method for calculating the covariance between coefficients of two nested probit models. In our Monte Carlo simulation study, therefore, the bootstrap method was used to construct the confidence intervals of  $\bar{b}_{yx} - \bar{b}_{yx.m}$ .



### 1.2.2. Residualized Variables

Breen et al. (2013) proposed another solution to the scaling problem in estimating the indirect effect for a binary outcome. In their method, the mediator  $m$  in Equation 2 is replaced by the residualized variable  $\tilde{m}$  as shown in the following equation:

$$y^* = \beta_2 + \beta_{yx.\tilde{m}}x + \beta_{y\tilde{m}.x}\tilde{m} + e_4, \tag{18}$$

where  $\tilde{m}$  is the residual from a regression of  $m$  on  $x$ ,  $e_4 = \sigma_4u$ ,  $u$  is a standard normal distribution for probit models and a standard logistic distribution for logit models, and  $\sigma_4$  is a scale factor. Since  $x$ -residualized  $\tilde{m}$  is uncorrelated with  $x$ , adding  $\tilde{m}$  will not change the coefficient of  $x$ , which gives the following equation:

$$\beta_{yx} = \beta_{yx.\tilde{m}} \tag{19}$$

Also, it can be shown that the model in Equation 18 can be obtained by reparameterizing the model in Equation 2, which implies that the residuals in Equations 2 and 18 should be the same:

$$\sigma_2 = \sigma_4 \tag{20}$$

Given Equations 19 and 20, we have the following equation:

$$b_{yx.\tilde{m}} - b_{yx.m} = \frac{\beta_{yx.\tilde{m}}}{\sigma_4} - \frac{\beta_{yx.m}}{\sigma_2} = \frac{\beta_{yx} - \beta_{yx.m}}{\sigma_2} = \frac{\beta_{mx}\beta_{ym.x}}{\sigma_2} = b_{mx}b_{ym.x}, \tag{21}$$

where  $b_{mx}$  is used to represent  $\beta_{mx}$  for notational consistency. Note that, unlike in Equation 13 in which  $\beta_{yx}$  and  $\beta_{yx.m}$  are measured on difference scales,  $\beta_{yx.\tilde{m}}$  and  $\beta_{yx.m}$  in Equation 21 are measured on the same scale. Therefore, Equation 21 implies that  $b_{yx.\tilde{m}} - b_{yx.m}$  measures the change in the coefficient of  $x$  due to the inclusion of  $m$ , or an indirect effect, on the same scale (Karlson et al., 2012). Another implication of Equation 21 is that it provides the exact decomposition of the total effect  $b_{yx.\tilde{m}} = \beta_{yx}/\sigma_2$  into the direct  $b_{yx.m} = \beta_{yx.m}/\sigma_2$  and indirect  $b_{mx}b_{ym.x} = \beta_{mx}\beta_{ym.x}/\sigma_2$  effects.

### 1.3. Confidence Intervals for Indirect Effects

Confidence intervals have been widely used as interval estimators for indirect effects because they are more informative than hypothesis tests. Confidence intervals can provide information about the variability and direction of the true effect as well as the binary decision on the statistical significance (Gardner & Altman, 1986; Harlow, Mulaik, & Steiger, 2013). Three types of confidence intervals for indirect effects have been discussed in the literature: confidence intervals based on the delta method (Sobel, 1982), the bootstrap method (Bollen & Stine, 1990; MacKinnon et al., 2004), and the Bayesian method (Yuan & MacKinnon, 2009).

A delta method is a general approach for approximating asymptotic standard errors of the non-linear function of statistics. Once the standard error is obtained using the delta method, confidence intervals can be constructed by assuming that the sampling distribution of the non-linear function of statistics follows a normal distribution. However, the normality assumption on the sampling distribution may not be valid in practice. For example, confidence intervals using the delta method performed poorer than did confidence intervals using the bootstrap method for small samples in which the true sampling distribution of the indirect effect deviates from the normal distribution (Bollen & Stine, 1990; MacKinnon et al., 2004).

Unlike the delta method, the bootstrap method does not assume any specific form of the sampling distribution. In the bootstrap method, the analytical derivation of the sampling distribution in the asymptotic theory is replaced with the sampling distribution's empirical

construction. Bootstrap samples of the same size as the original sample are randomly drawn from the original sample with replacement, and then the statistic of interest is calculated for the bootstrap samples to construct the empirical sampling distribution of the statistic. Because no distributional assumption is required in the bootstrap method, confidence intervals using the bootstrap method can be asymmetric to reflect the asymmetric nature of the true sampling distribution. The asymmetric confidence intervals using the bootstrap method may perform better than the symmetric confidence intervals using the delta method when the actual sampling distribution deviates from a normal distribution.

One of the simplest bootstrap confidence intervals is the percentile bootstrap confidence interval in which the lower and upper bounds of  $100(1-\alpha)\%$  confidence intervals are defined as  $\alpha$  and  $1-\alpha/2$  percentiles of the values of the statistic calculated from the bootstrap samples. Note that the justification for the percentile bootstrap confidence interval requires the existence of a monotone transformation of the statistic such that the transformed statistic on the transformed scale is symmetrical and centered on the observed statistic. However, such transformation rarely exists in practice, and therefore the percentile bootstrap confidence intervals are often incorrect. This limitation led to the development of the bias-corrected (BC) bootstrap confidence intervals in which bias in the sampling distribution of the statistic is adjusted using a correction factor (Davison & Hinkley, 1997). More specifically, let  $\hat{\theta}$  and  $\hat{\theta}_{(b)}$  be the statistics that are calculated from the original and  $b$ -th bootstrap sample respectively, where  $b = 1, \dots, B$  and  $B$  is the total number of bootstrap samples. In the BC bootstrap confidence intervals, the estimated is defined as the  $z$  score of the percentile of the observed  $\hat{\theta}$ . That is,  $\hat{z}_0 = \Phi^{-1}(p/B)$ , where  $p$  is the number of  $\hat{\theta}_{(b)}$ s that are less than  $\hat{\theta}$  and  $\Phi^{-1}$  is the inverse cumulative distribution function for a standard normal distribution. Then, the upper and lower bounds of  $100(1-\alpha)\%$  confidence intervals are defined as  $2\hat{z}_0 + z_{1-\alpha/2}$  and  $2\hat{z}_0 + z_{\alpha/2}$ , respectively (MacKinnon et al., 2004; Carpenter & Bithell, 2000).

Confidence intervals using the delta method and the bootstrap method are based on the frequentist approach in which an unknown parameter is treated as an unknown fixed value. In the frequentist approach, a confidence interval gives an estimated range of values that are likely to include the unknown fixed value of the parameter. On the contrary, the Bayesian approach treats an unknown parameter as a random variable with a probability distribution. In the Bayesian approach, prior information on the parameter of interest is quantified as a prior distribution, and the Bayes theorem is used to update the prior distribution to the posterior distribution by incorporating the observed data. All knowledge and uncertainty about the unknown parameter can be inferred from the posterior distribution. A credible interval in the Bayesian approach is the counterpart of the confidence interval in the frequentist approach, and the 95% credible interval is defined as the range between 0.025 and 0.975 percentiles of the posterior distribution. Yuan and MacKinnon (2009) pointed out that the Bayesian method is appealing for studies with complex mediation models and small samples because the Bayesian method does not impose restrictive normality assumptions on the sampling distribution of estimates.

So far, we have discussed various point and interval estimators of indirect effects for binary outcomes. As we mentioned earlier, this study aims to compare various point and interval estimators using Monte Carlo simulation studies. In the following sections, more detail on the Monte Carlo simulation studies are discussed.

## 2. Simulation Study

### 2.1. Simulation Study 1

#### 2.1.1. Simulation Description

Simulation study 1 was designed to demonstrate the difference between various point estimators of indirect effects for binary outcomes in terms of averages of estimated indirect effects across replications. Data sets for a simulation were generated based on Equations 2 and 3. The effect sizes of  $\beta_{yx.m}$ ,  $\beta_{ym.x}$ , and  $\beta_{mx}$  in Equations 2 and 3 were set equal to one another for simplicity and set at 0.14, 0.39, and 0.59 to represent small, medium, and large effect sizes, respectively (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002). The independent variable  $x$  was sampled from a standard normal distribution. Given  $x$ , the mediator  $m$  was generated based on Equation 3 under the assumption that the error term  $e_3$  follows a standard normal distribution. Then, a continuous latent response variable  $y^*$  was generated based on Equation 2 by setting the error term  $e_2$  as a standard normal distribution for a probit model and a standard logistic distribution for a logit model. Note that the scale factor  $\sigma_2$  was defined as  $e_2 = \sigma_2 u$ , where  $u$  is a standard normal distribution for a probit model and a standard logistic distribution for a logit model. Therefore, the scale factor  $\sigma_2$  in this simulation study was one because  $e_2 = u$  in our study. Sample sizes were set at 50, 100, 200, 500, 1000, and 5000 following MacKinnon et al. (2007). In all, 3 (effect sizes)  $\times$  6 (sample sizes) = 18 conditions were simulated for both probit or logit models, and each simulation condition was replicated 1000 times. The R software package (R Core Team, 2014) was used to generate data sets and to estimate the probit or logistic regression coefficients. For each simulation condition, the averages of estimated indirect effects from five different estimators were calculated: two product of coefficients estimators and three difference in coefficients estimators. To be more specific, an example showing how to calculate five different estimates for indirect effects is presented below. The data set in the example was generated based on the previously described procedure.

#### 2.1.2. An Example

Given a generated data set, the probit regressions described in Equations 9 and 10, and the linear regression described in Equation 3 were fitted to give the following regression coefficients:

$$g[\text{Pr}(y = 1|x)] = 0.0508(0.0433) + 0.7536(0.0524)x, \quad (22)$$

$$g[\text{Pr}(y = 1|x, m)] = 0.0346(0.0461) + 0.5656(0.0586)x + 0.5927(0.0521)m, \quad (23)$$

$$m = 0.0391(0.0309) + 0.5249(0.0311)x, \quad (24)$$

where the numbers in parentheses indicate the standard errors for the coefficients. The conventional product of coefficients estimate and the difference in coefficients estimate are  $\hat{b}_{mx} \hat{b}_{ym.x} = 0.5249 \times 0.5927 = 0.3111$  and  $\hat{b}_{yx} - \hat{b}_{yx.m} = 0.7536 - 0.5656 = 0.1880$ , respectively. Note that the difference in coefficients estimator underestimates the indirect effect because of the scaling issue.

On the other hand, the variances of  $y^*$  described in Equations 14 and 15 can be obtained as the following:

$$\text{Var}[y^*] = b_{yx}^2 \text{Var}[x] + 1 = (0.7536)^2(0.9839) + 1 = 1.5588 \quad (25)$$

$$\text{Var}[y^*] = b_{yx.m}^2 \text{Var}[x] + b_{ym.x}^2 \text{Var}[m] + 2b_{yx.m}b_{ym.x} \text{Cov}[x, m] + 1 \quad (26)$$

$$= (0.5656)^2(0.9839) + (0.5927)^2(1.2229) + 2(0.5656)(0.5927)(0.5165) + 1 \quad (27)$$

$$= 2.0906. \quad (28)$$

With the variances of  $y^*$ , the standardized regression coefficients can be obtained by dividing the regression coefficients by the square root of the variances of  $y^*$ :  $\bar{b}_{yx} = 0.7536/\sqrt{1.5588} = 0.6036$ ,  $\bar{b}_{yx.m} = 0.5656/\sqrt{2.0906} = 0.3912$ , and  $\bar{b}_{ym.x} = 0.5927/\sqrt{2.0906} = 0.4099$ . Then, the rescaled product of coefficients and difference in coefficients estimates can be obtained using the standardized regression coefficients:  $\hat{b}_{mx}\bar{b}_{ym.x} = 0.5249 \times 0.4099 = 0.2152$  and  $\bar{b}_{yx} - \bar{b}_{yx.m} = 0.6036 - 0.3912 = 0.2124$ . Note that two rescaled estimates are very similar, but not the same.

Finally, the difference in coefficients estimate can be obtained using the residualized  $m$ , which is the residual of  $m$  in Equation 24. In our study, the residualized  $m$  is denoted as  $\tilde{m}$ . Given  $\tilde{m}$ , another probit regression can be fitted to the data to give the following regression coefficients:

$$g[\Pr(y = 1|x, m)] = 0.0577(0.0461) + 0.8767(0.0596)x + 0.5927(0.0521)\tilde{m}. \quad (29)$$

Then, the difference in coefficients estimate is  $\hat{b}_{yx.\tilde{m}} - \hat{b}_{yx.m} = 0.8767 - 0.5656 = 0.3111$ . Note that  $\hat{b}_{yx.\tilde{m}} - \hat{b}_{yx.m}$  and  $b_{mx}\hat{b}_{ym.x}$  are exactly the same. In this example, we have demonstrated how to calculate five different estimates of indirect effects for binary outcomes:  $\hat{b}_{mx}\hat{b}_{ym.x}$ ,  $\hat{b}_{yx} - \hat{b}_{yx.m}$ ,  $\hat{b}_{yx}\hat{b}_{ym.x}$ ,  $\bar{b}_{yx} - \bar{b}_{yx.m}$ , and  $\hat{b}_{yx.\tilde{m}} - \hat{b}_{yx.m}$ . The results of simulation study 1 are presented below.

### 2.1.3. Results

The averages of estimated indirect effects from the five different estimators for probit and logit models are presented in Tables 1 and 2, respectively. Note that true parameter values for some estimators were unknown. In this simulation, the effect sizes of  $\beta_{yx.m}$ ,  $\beta_{ym.x}$ , and  $\beta_{mx}$  were manipulated. Therefore, the true values of  $b_{yx.m}$  and  $b_{ym.x}$  can be calculated using Equation 10 in which the scale factor  $\sigma_2$  can be set to one because we have used the standard normal and the standard logistic distributions as error distributions. Also, the true value of  $b_{mx}$  was known because  $\beta_{mx}$  was just relabeled as  $b_{mx}$  for notational consistency, i.e.,  $\beta_{mx} = b_{mx}$ . However, the true values of other coefficients,  $\beta_{yx}$  and  $\beta_{yx.m}$ , were unknown. Therefore, following MacKinnon et al. (2007), the averages of estimated indirect effects for samples of 106 were calculated across 1000 replications and were considered as true values. The estimated true values were exactly the same as the known true values up to four decimal points. For example, when  $\beta_{yx.m} = \beta_{ym.x} = \beta_{mx} = 0.14$ , the true value of  $b_{mx}b_{ym.x}$  was  $0.14 \times 0.14 = 0.0196$ , which was exactly the same as the value obtained for samples of 106.

For probit models, several trends can be identified from the results presented in Table 1. First, the conventional difference in coefficients estimator,  $\hat{b}_{yx} - \hat{b}_{yx.m}$ , underestimated the indirect effect compared to the conventional product of coefficients estimator,  $\hat{b}_{mx}\hat{b}_{ym.x}$ . For example, when  $\beta_{yx.m} = \beta_{ym.x} = \beta_{mx} = 0.59$  and the sample size is 106, the means of estimated indirect effects from  $\hat{b}_{mx}\hat{b}_{ym.x}$  and  $\hat{b}_{yx} - \hat{b}_{yx.m}$  were 0.3481 and 0.2180, respectively. The result showed the scaling problem in directly comparing regression coefficients of two nested probit models. As can be seen from Equation 13, the regression coefficients of two nested probit models are measured in different scales, and therefore are not directly comparable. The difference in estimated indirect effects from the two estimators increased as the effect size of coefficients increased, which is consistent with MacKinnon et al. (2007).

Second, the difference in coefficients estimator and the product of coefficients estimator with standardized coefficients, which are  $\hat{b}_{mx}\bar{b}_{ym.x}$  and  $\bar{b}_{yx} - \bar{b}_{yx.m}$ , yielded very similar, but not identical, results across all simulation conditions. This result showed that the use of the standardized regression coefficients can reduce the difference in estimated indirect effects from

the difference in coefficients and the product of coefficients estimators, but the difference still exists. That is, the total effect can not be exactly decomposed into the direct and indirect effects with the estimators using the standardized regression coefficients, which might cause some problems in calculating the proportion of the indirect effect in the total effect or the effect size of the indirect effect.

Lastly, the conventional product of coefficients estimator,  $\hat{b}_{mx}\hat{b}_{ym.x}$  and the difference in coefficients estimator with the residualized variable,  $\hat{b}_{yx.\tilde{m}} - \hat{b}_{yx.m}$ , produced exactly the same estimate for the indirect effect. The result indicates that the total effect can be exactly decomposed into the direct and indirect effects by using the residualized variable (Breen et al., 2013). The results for logit models in Table 2 also show similar patterns, as shown in Table 1.

**Table 1.** Averages of Estimated Indirect Effects from Different Estimators (Probit Models).

n	$\beta$	Products		Differences		
		$\hat{b}_{mx}$	$\hat{b}_{ym.x}$	$\hat{b}_{mx}\bar{\hat{b}}_{ym.x}$	$\hat{b}_{yx} - \hat{b}_{yx.m}$	$\bar{\hat{b}}_{yx} - \bar{\hat{b}}_{yx.m}$
50	0.14	0.0223	0.0203	0.0180	0.0208	0.0223
100	0.14	0.0204	0.0192	0.0171	0.0192	0.0204
200	0.14	0.0196	0.0189	0.0173	0.0189	0.0196
500	0.14	0.0195	0.0189	0.0177	0.0189	0.0195
1000	0.14	0.0196	0.0191	0.0180	0.0191	0.0196
5000	0.14	0.0198	0.0194	0.0182	0.0194	0.0198
10 <sup>6</sup>	0.14	0.0196	0.0192	0.0181	0.0192	0.0196
50	0.39	0.1675	0.1341	0.1185	0.1287	0.1675
100	0.39	0.1593	0.1278	0.1151	0.1271	0.1593
200	0.39	0.1596	0.1279	0.1169	0.1300	0.1596
500	0.39	0.1546	0.1260	0.1143	0.1276	0.1546
1000	0.39	0.1523	0.1263	0.1148	0.1266	0.1523
5000	0.39	0.1526	0.1266	0.1149	0.1267	0.1526
10 <sup>6</sup>	0.39	0.1521	0.1265	0.1150	0.1265	0.1521
50	0.59	0.4075	0.2383	0.2076	0.2422	0.4075
100	0.59	0.3714	0.2362	0.2189	0.2379	0.3714
200	0.59	0.3598	0.2346	0.2178	0.2349	0.3598
500	0.59	0.3518	0.2333	0.2184	0.2335	0.3518
1000	0.59	0.3515	0.2339	0.2185	0.2338	0.3515
5000	0.59	0.3492	0.2336	0.2183	0.2336	0.3492
10 <sup>6</sup>	0.59	0.3481	0.2332	0.2180	0.2332	0.3481

Notes. The number in each cell represents the averages of estimated indirect effects across 3000 replications for a given condition. The effect sizes of coefficients were set to be equal, i.e.,  $\beta = \beta_{mx} = \beta_{ym.x} = \beta_{yx.m}$ . The bar over a coefficient indicates that the coefficient is standardized using the variance of a latent response variable  $y^*$ . The hat over a coefficient indicates that it is a estimated value.  $\tilde{m}$  represents the x-residualized  $m$  variable, i.e., the residual of  $m$  when  $m$  is regressed on  $x$ .

**Table 2.** Averages of Estimated Indirect Effects from Different Estimators (Logit Models).

n	$\beta$	Products		Differences		
		$\hat{b}_{mx}$	$\hat{b}_{ym.x}$	$\hat{b}_{yx} - \hat{b}_{yx.m}$	$\bar{\hat{b}}_{yx} - \bar{\hat{b}}_{yx.m}$	$\hat{b}_{yx.\hat{m}} - \hat{b}_{yx.m}$
50	0.14	0.0200	0.0102	0.0154	0.0096	0.0200
100	0.14	0.0213	0.0113	0.0185	0.0108	0.0213
200	0.14	0.0204	0.0110	0.0187	0.0107	0.0204
500	0.14	0.0202	0.0110	0.0190	0.0108	0.0202
1000	0.14	0.0192	0.0105	0.0183	0.0103	0.0192
5000	0.14	0.0196	0.0107	0.0188	0.0105	0.0196
10 <sup>6</sup>	0.14	0.0196	0.0107	0.0188	0.0106	0.0196
<hr/>						
50	0.39	0.1675	0.1341	0.1185	0.1287	0.1675
100	0.39	0.1593	0.1278	0.1151	0.1271	0.1593
200	0.39	0.1596	0.1279	0.1169	0.1300	0.1596
500	0.39	0.1546	0.1260	0.1143	0.1276	0.1546
1000	0.39	0.1523	0.1263	0.1148	0.1266	0.1523
5000	0.39	0.1526	0.1266	0.1149	0.1267	0.1526
10 <sup>6</sup>	0.39	0.1521	0.1265	0.1150	0.1265	0.1521
<hr/>						
50	0.59	0.4075	0.2383	0.2076	0.2422	0.4075
100	0.59	0.3714	0.2362	0.2189	0.2379	0.3714
200	0.59	0.3598	0.2346	0.2178	0.2349	0.3598
500	0.59	0.3518	0.2333	0.2184	0.2335	0.3518
1000	0.59	0.3515	0.2339	0.2185	0.2338	0.3515
5000	0.59	0.3492	0.2336	0.2183	0.2336	0.3492
10 <sup>6</sup>	0.59	0.3481	0.2332	0.2180	0.2332	0.3481

Notes. The number in each cell represents the averages of estimated indirect effects across 3000 replications for a given condition. The effect sizes of coefficients were set to be equal, i.e.,  $\beta = \beta_{mx} = \beta_{ym.x} = \beta_{yx.m}$ . The bar over a coefficient indicates that the coefficient is standardized using the variance of a latent response variable  $y^*$ . The hat over a coefficient indicates that it is a estimated value.  $\hat{m}$  represents the x-residualized  $m$  variable, i.e., the residual of  $m$  when  $m$  is regressed on  $x$ .

### 2.2. Simulation Study 2

In simulation study 2, confidence and credible intervals of the product of coefficients estimator,  $\hat{b}_{mx} \hat{b}_{ym.x}$ , were constructed using the delta, bootstrap, and Bayesian methods, and their performance were compared in terms of powers, type I error rates, and coverage rates. As can be seen from the simulation study 1, the values of  $\hat{b}_{yx.\hat{m}} - \hat{b}_{yx.m}$  were exactly the same as the values of  $\hat{b}_{mx} \hat{b}_{ym.x}$ . Also, testing  $H_0: \hat{b}_{mx} \hat{b}_{ym.x} = 0$  and  $H_0: \hat{b}_{mx} \bar{\hat{b}}_{ym.x} = 0$  are equivalent.

Mplus (L. K. Muth'en & Muth'en, 2010) was used to construct confidence and credible intervals. Note that the bootstrap confidence interval implemented in Mplus is the BC bootstrap confidence interval. In this simulation, date sets were generated using only the probit model because Mplus limits the data generation to the Probit model. In order to model binary outcomes, the weighted least square estimation should be used in Mplus. However, the weighted least square estimation allows only the probit link. Also, the Bayesian estimation in Mplus only allows the probit link. Therefore, data sets in this simulation were generated using the probit model by setting the error term  $e_2$  in Equation 2 as a standard normal distribution, and then the probit model was used to estimate relevant coefficients to calculate estimates for indirect effects.



Data sets were generated using the latent response variable as described in simulation study 1. However, unlike in the simulation study 1, only the effect sizes of  $\beta_{mx}$  and  $\beta_{ym.x}$  were set equal to one another, and the effect sizes of  $\beta_{mx}$  and  $\beta_{yx.m}$  were set at 0.14, 0.39, and 0.59. Sample sizes were set at 50, 100, 200, 500, and 1000. In all,  $32$  (effect sizes)  $\times$   $5$  (sample sizes) =  $45$  conditions were simulated to calculate powers and coverage rates. For type I error rates,  $\beta_{mx}$  and  $\beta_{ym.x}$  are set equal to zeros, and  $\beta_{yx.m}$  were set at 0.14, 0.39, and 0.59. Therefore,  $3$  (effect sizes)  $\times$   $5$  (sample sizes) =  $15$  conditions were simulated for type I error rates. Each simulation condition was replicated 3000 times.

Given data sets, confidence or credible intervals were constructed using Mplus. In constructing confidence intervals using the delta and bootstrap methods, the weighted least square estimation was used by setting the ESTIMATOR = WLSMV option, and the indirect effects were defined using the MODEL CONSTRAINT command. In constructing credible intervals, the Bayesian estimation was used by setting the ESTIMATOR = BAYES option. Because the option for bootstrap confidence intervals was not compatible with the built-in Monte Carlo facility in Mplus, author-written R code was used to automatically run Mplus and extract relevant estimates from generated output files. The coverage rate was evaluated using the criteria suggested by Bradley (1978); the confidence interval is considered to be liberally, moderately, or strictly robust if the coverage rate falls within the range [.925, .975], [.940, .960], or [.945, .955], respectively.

### 2.2.1. Results

Confidence and credible intervals for the indirect effect estimator,  $\hat{b}_{mx}$   $\hat{b}_{ym.x}$ , were constructed using the delta, bootstrap, and Bayesian methods, and their powers, coverage rates, and type I error rates are presented in Tables 3, 4, and 5. Several trends can be identified from the tables. First, the powers of the BC bootstrap confidence intervals were higher than the powers of other confidence or credible intervals in almost every simulation condition. The differences in powers among methods are prominent when the sample sizes and effect sizes are small. For example, the powers of the confidence and credible intervals using the delta, bootstrap, and Bayesian methods were 0.048, 0.217, and 0.125 respectively when the sample size is 200, and  $\beta_{mx} = \beta_{ym.x} = \beta_{yx.m} = 0.14$ . The only exceptions were the conditions in which sample sizes were 50, and  $\beta_{mx} = \beta_{ym.x} = 0.59$ . In those conditions, the powers of the credible intervals using the Bayesian method were little bit higher than the powers of the confidence intervals using the bootstrap method.

Second, the BC bootstrap confidence intervals performed better than did other intervals in terms of coverage rates. In our study, confidence and credible intervals were constructed with a 95% confidence level. Therefore, the nominal coverage rate of confidence and credible intervals is .95. In Tables 3, 4, and 5, The values marked with \*, \*\*, and \*\*\* indicate that the coverage rates are liberally [.925, .975], moderately [.94, .96], and strictly [.945, .955] robust based on the criteria suggested by Bradley (1978). As can be seen from the tables, the BC bootstrap confidence intervals were more robust than other intervals. The coverage rates of the intervals seemed to become close to the nominal level of .95 as the sample sizes and effect sizes increase. However, the pattern is less clear for sample sizes.

Third, as shown in Table 3, the type I error rates of the delta, bootstrap, and Bayesian methods were very close to zero in all simulation conditions. Because confidence and credible intervals were constructed with a 95% confidence level in this study, the nominal type I error rate is .05. Therefore, our results indicate that the tests of indirect effects using confidence and credible intervals are very conservative.

**Table 3.** Powers, Coverage Rates, and Type I Error Rates of Confidence and Credible Intervals for  $\hat{b}_{mx}$   $\hat{b}_{ym.x}$

n	$\beta_{mx}$ $\beta_{ym.x}$			Delta			Bootstrap Pow			Bayesian Pow		
	$n$	$\hat{\beta}_{mx}$	$\beta_{ym.x}$ $\beta_{yx.m}$	Pow	Cov	Typ	$\bar{P}ow$	Cov	Typ	$\bar{P}ow$	Cov	Typ
50	.14(0)	.14	(0).14	.002	.965*	.000	.024	.976	.002	.020	.990	.002
100	.14(0)	.14	(0).14	.008	.929*	.000	.072	.936*	.005	.027	.986	.003
200	.14(0)	.14	(0).14	.048	.908	.000	.217	.938*	.004	.125	.946***	.001
500	.14(0)	.14	(0).14	.353	.929*	.000	.640	.960**	.003	.623	.939*	.002
1000	.14(0)	.14	(0).14	.844	.936*	.001	.930	.958**	.002	.873	.954**	.001
50	.14(0)	.14	(0).39	.002	.972*	.000	.021	.980	.002	.014	.989	.002
100	.14(0)	.14	(0).39	.008	.929*	.000	.071	.929*	.006	.024	.982	.002
200	.14(0)	.14	(0).39	.044	.907	.000	.190	.927*	.004	.114	.946***	.002
500	.14(0)	.14	(0).39	.345	.922	.000	.627	.956**	.002	.597	.939*	.001
1000	.14(0)	.14	(0).39	.831	.937*	.000	.920	.954**	.001	.862	.954***	.001
50	.14(0)	.14	(0).59	.001	.964*	.000	.021	.979	.004	.014	.993	.001
100	.14(0)	.14	(0).59	.007	.940**	.000	.069	.931*	.005	.022	.980	.002
200	.14(0)	.14	(0).59	.042	.913	.000	.192	.928*	.004	.112	.940**	.001
500	.14(0)	.14	(0).59	.314	.930*	.000	.601	.958**	.003	.580	.940**	.002
1000	.14(0)	.14	(0).59	.805	.942**	.000	.921	.960**	.003	.849	.964*	.001

Notes. Pow=powers, Cov=coverage rates, and Typ=type I error rates. Each condition was replicated 3000 times. Type I error rates were calculated by setting  $\beta_{mx} = \beta_{ym.x} = 0$  as indicated by zeros within the parentheses. Bootstrap indicates the BC bootstrap confidence intervals. The values marked with \*, \*\*, and \*\*\* indicate that the coverage rates are liberally [.925, .975], moderately [.94, .96], and strictly [.945, .955] robust (Bradley,1978).

**Table 4.** Powers and Coverage Rates of Confidence and Credible Intervals for  $\hat{b}_{mx}$   $\hat{b}_{ym.x}$

n	$\beta_{mx}$ $\beta_{ym.x}$			Delta		Bootstrap		Bayesian	
	$n$	$\hat{\beta}_{mx}$	$\beta_{ym.x}$ $\beta_{yx.m}$	$\bar{P}ow$	Cov	Pow	Cov	$\bar{P}ow$	Cov
50	.39	.39	.14	.158					
50	.39	.39	.14	.158	.922	.401	.955***	.398	.942**
100	.39	.39	.14	.645	.941**	.831	.962*	.750	.946***
200	.39	.39	.14	.979	.936*	.990	.953***	.990	.938*
500	.39	.39	.14	1.000	.951***	1.000	.953***	1.000	.924*
1000	.39	.39	.14	1.000	.942**	1.000	.948***	1.000	.955***
50	.39	.39	.39	.129	.925*	.383	.959**	.375	.949***
100	.39	.39	.39	.614	.938*	.810	.960**	.723	.935*
200	.39	.39	.39	.968	.936*	.980	.951***	.983	.934*
500	.39	.39	.39	1.000	.942**	1.000	.944**	1.000	.926*
1000	.39	.39	.39	1.000	.946***	1.000	.950***	1.000	.954***
50	.39	.39	.59	.119	.925*	.368	.962*	.364	.944**
100	.39	.39	.59	.553	.938*	.779	.955***	.691	.933*
200	.39	.39	.59	.961	.940**	.977	.954***	.982	.935*
500	.39	.39	.59	1.000	.947***	1.000	.947***	1.000	.930*
1000	.39	.39	.59	1.000	.953***	1.000	.955***	1.000	.960**

Notes. Pow=powers and Cov=coverage rates. Each condition was replicated 3000 times. Bootstrap indicates the BC bootstrap confidence intervals. The values marked with \*, \*\*, and \*\*\* indicate that the coverage rates are liberally [.925, .975], moderately [.94, .96], and strictly [.945, .955] robust (Bradley, 1978).

**Table 5.** Powers and Coverage Rates of Confidence and Credible Intervals for  $\hat{b}_{mx}$   $\hat{b}_{ym.x}$

n	$\beta_{mx}$	$\beta_{ym.x}$		Delta		Bootstrap		Bayesian	
<i>n</i>	$\beta_{mx}$	$\beta_{ym.x}$	$\beta_{yx.m}$	Pow	Cov	Pow	Cov	Pow	Cov
50	.39	.39	.14	.158					
50	.59	.59	.14	.606	.950	.811	.954***	.842	.943**
100	.59	.59	.14	.979	.947***	.987	.950***	.750	.946***
200	.59	.59	.14	1.000	.952***	1.000	.947***	.990	.938*
500	.59	.59	.14	1.000	.953***	1.000	.951***	1.000	.924*
1000	.59	.59	.14	1.000	.950***	1.000	.946***	1.000	.955***
50	.59	.59	.39	.555	.943**	.769	.954***	.808	.936*
100	.59	.59	.39	.974	.952***	.984	.953***	.981	.935*
200	.59	.59	.39	1.000	.948***	1.000	.953***	1.000	.932*
500	.59	.59	.39	1.000	.947***	1.000	.946***	1.000	.926*
1000	.59	.59	.39	1.000	.954***	1.000	.954***	1.000	.955***
50	.59	.59	.59	.723	.948***	.754	.948***	.791	.931*
100	.59	.59	.59	.954	.934*	.971	.944**	.961	.923
200	.59	.59	.59	1.000	.944**	1.000	.944**	1.000	.917
500	.59	.59	.59	1.000	.953***	1.000	.954***	1.000	.932*
1000	.59	.59	.59	1.000	.954***	1.000	.954***	1.000	.959**

Notes. Pow=powers and Cov=coverage rates. Each condition was replicated 3000 times. Bootstrap indicates the BC bootstrap confidence intervals. The values marked with \*, \*\*, and \*\*\* indicate that the coverage rates are liberally [.925, .975], moderately [.94, .96], and strictly [.945, .955] robust (Bradley, 1978).

### 3. DISCUSSION

The indirect effect has been estimated in two ways: the difference in coefficients or the product of coefficients. Unlike for continuous outcomes, the difference in coefficients estimator for binary outcomes systematically underestimates the indirect effect because the estimator compares regression coefficients that are measured in different scales. To address the scaling issue, it was proposed to use standardized regression coefficients (Winship & Mare, 1983; MacKinnon, 2008) or residualized variables (Breen et al., 2013). The simulation study 1 was designed to contrast those estimators of indirect effects for binary outcomes in terms of the averages of estimated indirect effects. On the other hand, confidence or credible intervals have been widely used to test indirect effects. In the simulation study 2, confidence or credible intervals of the product of coefficients estimator for binary outcomes were constructed using the delta, bootstrap, and Bayesian methods, and their performance were compared in terms of powers, coverage rates, and type I error rates.

In the simulation study 1, five different point estimators were compared in terms of the averages of estimated indirect effects. The results in Tables 1 and 2 showed that the conventional difference in coefficients estimator ( $\hat{b}_{yx} - \hat{b}_{yx.m}$ ) systematically underestimated the indirect effects compared to the conventional product of coefficients estimator ( $\hat{b}_{mx} \hat{b}_{ym.x}$ ). The discrepancy between the two estimators can be reduced by using the standardized regression coefficients. That is, estimated indirect effects from  $\hat{b}_{mx} \hat{b}_{ym.x}$  and  $\hat{b}_{yx} - \hat{b}_{yx.m}$  were similar but not identical. The estimated indirect effects from the conventional product of coefficients estimator ( $\hat{b}_{mx} \hat{b}_{ym.x}$ ) and the difference in coefficients estimator using the residualized variable ( $\hat{b}_{yx.\hat{m}} - \hat{b}_{yx.m}$ ) were exactly the same, which indicates the exact decomposition of the total effect into the direct and indirect effects. In all, the conventional difference in

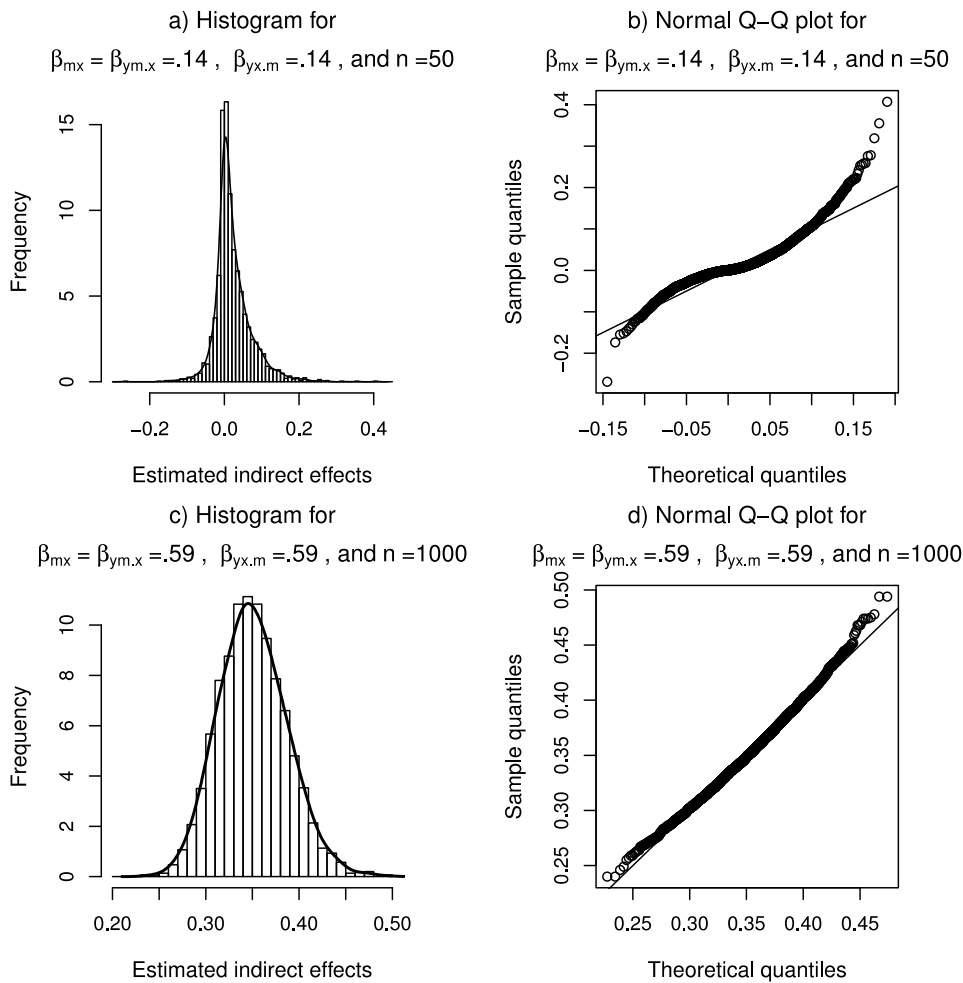
coefficients estimator ( $\hat{b}_{yx} - \hat{b}_{yx.m}$ ) should not be used for binary outcomes. Also, the counterparts of  $\hat{b}_{yx.m} - \hat{b}_{yx}$  and  $\hat{b}_{yx} - \hat{b}_{yx.m}$  are the  $\hat{b}_{mx}\hat{b}_{ym.x}$  and  $\hat{b}_{mx}\bar{\hat{b}}_{ym.x}$ , respectively.

Here, it may be interesting to discuss the effect size measure of indirect effects for a binary outcome. At first, it may seem that the exact decomposition is necessary to interpret the effect size measures that are defined as the ratio of the indirect effect to the total effect,  $b_{mx}b_{ym.x}/(b_{yx.m} + b_{mx}b_{ym.x})$  as proportion. However, as Preacher and Kelley (2011) pointed out,  $b_{mx}b_{ym.x}/(b_{yx.m} + b_{mx}b_{ym.x})$  is not a proportion, and can exceed one and even be negative in some cases. On the other hand, in our previous example, the values of  $b_{mx}b_{ym.x}/(b_{yx.m} + b_{mx}b_{ym.x})$  that were calculated using the original and standardized regression coefficients were exactly the same. The value of  $b_{mx}b_{ym.x}/(b_{yx.m} + b_{mx}b_{ym.x})$  was  $(0.5249 \times 0.5927)/(0.5656 + 0.5249 \times 0.5927) = 0.3548$  for the original coefficients, and the value of  $b_{mx}\bar{\hat{b}}_{ym.x}/(\bar{\hat{b}}_{yx.m} + b_{mx}\bar{\hat{b}}_{ym.x})$  was  $(0.5249 \times 0.5927 \sqrt{2.0906})/(0.5656 \sqrt{2.0906} + 0.5249 \times 0.5927/\sqrt{2.0906}) = 0.3548$  for the standardized coefficients. It is apparent that the two values should be the same because the numerator and denominator in the effect size for original coefficients are divided by the same scaling factor. Moreover, Breen et al. (2013) also suggested to use  $b_{mx}b_{ym.x}/(b_{yx.m} + b_{mx}b_{ym.x})$  as the effect size measures for the indirect effect with a residualized variable, which gives exactly the same effect size of 0.3548.

Confidence and credible intervals have been widely used as interval estimators for indirect effects. In the literature, different methods for constructing interval estimators have been compared for indirect effects with continuous outcomes (MacKinnon et al., 2004; Yuan & MacKinnon, 2009). In this study, confidence and credible intervals using the delta, bootstrap, and Bayesian methods were compared for indirect effects with binary outcomes. The results in Tables 3, 4, and 5 showed that the BC bootstrap confidence intervals performed better than did other intervals in terms of powers, coverage rates, and type I error rates, especially when the sample sizes and effect sizes are small. This result is expected because the sampling distributions of estimators tend to deviate from the normal distribution in small samples (MacKinnon et al., 2004; Bollen & Stine, 1990; Yuan & MacKinnon, 2009). In Figure 1, histograms and normal Q-Q plots of estimated indirect effects for the worst- and best- case scenarios in our study are presented to demonstrate how much the sampling distribution of indirect effects could deviate from the normal distribution depending on the sample sizes and effect sizes. For the worst case scenario, the histogram and normal Q-Q plot for  $\beta_{mx} = \beta_{ym.x} = .14$ ,  $\beta_{yx.m} = .14$ , and  $n = 50$  are presented in Figures 1a and 1b, which show clear deviation from the normal distribution. On the contrary, the histogram and normal Q-Q plot for the best case scenario, where  $\beta_{mx} = \beta_{ym.x} = .59$ ,  $\beta_{yx.m} = .59$ , and  $n = 1000$ , were very close to those for the true normal distribution. Therefore, the poor performance of the confidence intervals using the delta method seems to be reasonable because the assumption of the normal sampling distribution in the delta method is not valid in small samples.

On the other hand, the comparison between the BC bootstrap confidence intervals and the Bayesian credible intervals is interesting. Both methods do not require any specific form of the sampling distribution. The sampling distribution in the bootstrap method is empirically constructed, and the posterior distribution in the Bayesian method is updated from the prior distribution. With the flexibility in the form of the sampling distribution, the two methods can capture the possible asymmetric nature of the true sampling distribution in small samples. Therefore, the better performance of the two methods over the delta method can be understood as the result of the flexible assumption about the sampling distribution.

Figure 1. Histograms and Normal Q-Q plots.



Note. The values of estimated indirect effects for histograms and normal Q-Q plots come from the bootstrap method. The estimated indirect effects were exactly the same for both bootstrap and delta methods. The estimated indirect effects from the Bayesian estimation were little bit different from those from the bootstrap and delta methods, but produced very similar histogram and normal Q-Q plot. In Figures (a) and (c), solid lines indicate density plots for the corresponding histograms.

Interestingly, the BC confidence intervals showed better performance than did the Bayesian credible intervals in most simulation conditions. One of the possible explanations may be the use of the default non-informative prior in Mplus. In general, inferences in the Bayesian method are made based on the posterior distribution, which is proportional to the product of the prior and likelihood distributions. The non-informative prior, which is the default prior in Mplus, is typically used when there is no prior knowledge on the parameter of interest. In such a case, the likelihood distribution is the only dominant factor for estimating the posterior distribution. Note that, even in the Bayesian method, we still need an assumption about the form of the likelihood distribution. Therefore, the use of non-informative prior may make the estimation procedure less flexible in capturing the asymmetric nature of the true sampling distribution because the estimation procedures heavily rely on the pre-specified form of the likelihood distribution. In our study, the average widths of intervals using the delta, bootstrap, and Bayesian methods were 0.193, 0.276, and 0.250 respectively when  $\beta_{mx} = \beta_{ym.x} = .14$ ,  $\beta_{yx.m} = .14$ , and  $n = 50$ , and were 0.144, 0.144, and 0.149 respectively when

$\beta_{mx} = \beta_{ym.x} = .59$ ,  $\beta_{yx.m} = .59$ , and  $n = 1000$ . In small samples, the BC confidence intervals were most wide, whereas all intervals were quite similar in their average widths in large

samples. This might indicate that the BC confidence intervals are most flexible in capturing possible asymmetric nature of the true sampling distribution.

This study compared various point and interval estimators of the indirect effect for a binary outcome. The conventional difference in coefficients estimator should be avoided in estimating the indirect effect for a binary outcome because of the scaling problem. For interval estimations, the BC bootstrap confidence intervals seem to perform better than the intervals based on the delta and Bayesian methods. In this study, only non-informative prior was used in the Bayesian method. It would be interesting to compare the bootstrap method with the Bayesian methods with different prior distributions. Also, this study did not consider the case where the moderator is binary, which would be another interesting study.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

### Authorship contribution statement

**Hyung Rock Lee:** Investigation, Resources, Visualization, Software, Formal Analysis, and Writing - original draft. **Jaeyun Sung:** Methodology, Visualization, Supervision, and Validation. **Sunbok Lee:** Methodology, Software, Formal Analysis, Supervision, and Validation.

### ORCID

Hyung Rock Lee  <https://orcid.org/0000-0002-7415-9466>

Jaeyun Sung  <https://orcid.org/0000-0001-7461-3123>

Sunbok Lee  <https://orcid.org/0000-0002-0924-7056>

## 5. REFERENCES

- Allison, P. D. (1999). Comparing logit and probit coefficients across groups. *Sociological Methods & Research*, 28, 186-208.
- Alwin, D. F., & Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review*, 40, 37-47.
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173-1182.
- Bollen, K. A. (1987). Total, direct, and indirect effects in structural equation models. *Sociological Methodology*, 17, 37-69.
- Bollen, K. A., & Stine, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology*, 20, 15-140.
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.
- Breen, R., Karlson, K. B., & Holm, A. (2013). Total, direct, and indirect effects in logit and probit models. *Sociological Methods & Research*, 42, 164-191.
- Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in Medicine*, 19, 1141-1164.
- Clogg, C. C., Petkova, E., & Shihadeh, E. S. (1992). Statistical methods for analyzing collapsibility in regression models. *Journal of Educational and Behavioral Statistics*, 17, 51-74.
- Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press.



- Freedman, L. S., & Schatzkin, A. (1992). Sample size for studying intermediate endpoints within intervention trials or observational studies. *American Journal of Epidemiology*, *136*, 1148-1159.
- Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than p values: estimation rather than hypothesis testing. *British Medical Journal (Clinical Research)*, *292*, 746-750.
- Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (2013). *What if there were no significance tests?* Psychology Press.
- Karlson, K. B., Holm, A., & Breen, R. (2012). Comparing regression coefficients between same-sample nested models using logit and probit a new method. *Sociological Methodology*, *42*, 286-313.
- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Routledge.
- MacKinnon, D. P., & Cox, M. C. (2012). Commentary on mediation analysis and categorical variables: The final frontier by dawn iacobucci. *Journal of Consumer Psychology: the official journal of the Society for Consumer Psychology*, *22*, 600-602.
- MacKinnon, D. P., & Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review*, *17*, 144-158.
- MacKinnon, D. P., Lockwood, C. M., Brown, C. H., Wang, W., & Hoffman, J. M. (2007). The intermediate endpoint effect in logistic and probit regression. *Clinical Trials*, *4*, 499-513.
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*, 83-104.
- MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*, 99-128.
- MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, *30*, 41-62.
- Muthen, B. (1979). A structural probit model with latent variables. *Journal of the American Statistical Association*, *74*, 807-811.
- Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115-132.
- Muthen, L. K., & Muthen, B. O. (2010). *Mplus: Statistical analysis with latent variables: User's guide*. Muthen & Muthen.
- Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: quantitative strategies for communicating indirect effects. *Psychological Methods*, *16*, 93-115.
- R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <http://www.R-project.org/>
- Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, *13*, 290-312.
- Winship, C., & Mare, R. D. (1983). Structural equations and path analysis for discrete data. *American Journal of Sociology*, *89*, 54-110.
- Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, *14*, 301-322.

## Examining the Dimensionality and Monotonicity of an Attitude Dataset based on the Item Response Theory Models

Seval Kula Kartal <sup>1,\*</sup>, Ezgi Mor Dirlik <sup>2</sup>

<sup>1</sup>Pamukkale University, Faculty of Education, Department of Educational Sciences, Measurement and Evaluation in Education, Denizli, Turkey

<sup>2</sup>Kastamonu University, Faculty of Education, Department of Educational Sciences, Measurement and Evaluation in Education, Kastamonu, Turkey

### ARTICLE HISTORY

Received: Apr. 20, 2020

Revised: Jan. 12, 2021

Accepted: Mar. 12, 2021

### Keywords:

Dimensionality,  
Monotonicity,  
Generalized graded  
unfolding model,  
Non-parametric item  
response theory.

**Abstract:** In the current study, the factor structure of an attitude scale was analyzed by using the two different item response theory models that allow modeling non-monotonic item response curves. The current study utilized the two models to examine whether the two-factor solution of factor analysis may be caused by method effect, or by the failure of the analysis in describing and fitting the dataset because of the monotonicity assumption. This study was conducted on a dataset obtained from 355 undergraduate students who were studying at the Middle East Technical University. The data were obtained by carrying out the Attitude Scale Towards Foreign Languages as Medium of Instruction, which was developed by Kartal and Gülleroğlu (2015). The fit of the scale items to the generalized graded unfolding model was examined based on the item response curves, item parameters, item fit statistics and fit graphics. For Mokken scaling, scalability coefficients were calculated, dimensionality analyzes were conducted by using the Automated Item Selection Procedure. The monotonicity assumption was investigated based on the rest-score group methods. The results of the current study revealed that items of the attitude scale fit to the unidimensional models that do not assume monotone increasing item response curves for all items, while the factor analysis suggested a two-factor solution for the data. Researchers are recommended to utilize statistical techniques that can identify any possible violation of the monotonicity assumption and model items having non-monotonic response curves to examine dimensionality of their data.

## 1. INTRODUCTION

Behaviors of individuals, which are among the fundamental research areas of education and psychology, are mostly observed indirectly based on the measurement tools that have been developed to observe specific behaviors of people depending on their answers to the scale or test items. Measurement tools generally include both negatively and positively worded items to prevent possibility of response bias. However, inclusion of negatively and positively worded items on the measurement tool may cause respondents' answers to be affected by wording

---

CONTACT: Seval Kula Kartal ✉ [kulasevaal@gmail.com](mailto:kulasevaal@gmail.com) 📍 Pamukkale University, Faculty of Education, Department of Educational Sciences, Measurement and Evaluation in Education, 20070, Denizli, Turkey

ISSN-e: 2148-7456 /© IJATE 2021

direction of items (DiStefano & Motl, 2009; Tomas & Oliver, 1999).

As stated by Brown (2006), in addition to effects of the main dimensions measured by the scale, items may also be affected by the method that is used to collect the data. Researchers may obtain high correlations among items because of wording direction. As a result, items may constitute two separate factors one of which includes only negatively worded items and the other one includes only positively items, while there is actually one dominant latent dimension underlying the scale items. The related researches also support that wording direction of items affects how respondents answer to the scale items and causes spurious factors because of the method effect (Gu et al., 2015; Wang et al., 2001; Wang et al., 2018; Wouters et al., 2012).

Whether there is any method effect on the data is an important question that researchers should answer during their dimensionality analyses. One of the dimensionality analyses mostly utilized to detect presence of item direction factors is the confirmatory factor analysis (Horan et al., 2003; Supple & Plunkett, 2011; Tomas & Oliver, 1999). In the confirmatory factor analysis framework, researchers analyze if the scale items constitute two distinct factors each including items written in one direction. In case of confirming the two-factor structure caused by the wording direction of items there is another important point about the monotonicity assumption of the factor analysis that researchers should take into consideration to make correct decisions concerning the dimensionality of the data.

A monotonic relation between the latent trait and item response is one of the fundamental assumptions of the factor analysis. The factor analysis assumes that the values of observed variables are linearly (or even monotonically) related to values on the underlying latent variables. The monotonicity assumption is an essential point that researchers should consider while analyzing the dimensionality of their data. The main reason of this is that the monotonicity assumption may affect predictions of different dimensionality analysis techniques concerning the size and sign of the inter-item correlations. For example, the factor analysis accepts that all scale items have linear and monotonically increasing item response curves. It may be correct to assume monotone response curves for the extreme scale items. However, moderate items are more likely to have bell-shaped response curves. This means that factor analysis may not be able to describe the correlations among moderate items appropriately because of the monotonicity assumption (Van Schuur & Kiers, 1994).

Furthermore, the factor analysis expects high and positive correlations among scale items, measuring one dominant dimension, after all negatively worded items are reverse coded. In contrast, several techniques that can model nonmonotonic item response curves, such as the generalized graded unfolding model, expects high correlations only among items that are close together along the latent dimension, since respondents will tend to show similar reactions to those items. As stated by Davison (1977), as the distance between items increases, the correlation between them decreases, and then may begin to increase again this time with a negative sign. Thus, a correlation matrix of a dataset fitting the generalized graded unfolding model, will have high correlations along the diagonal, lower correlations downward and to the left, and negative correlations in the lower-left corner. Since such a correlation matrix includes both negative and positive correlations, factor analysis of this matrix may confirm a two-factor structure, while there is in fact one –not two- latent dimension underlying scale items (Davison, 1977; Spector et al., 1997, Tay & Drasgow, 2012; Van Schuur & Kiers, 1994). If the data does not hold the main assumption of the factor analysis (linear and monotonically increasing item response curves), the factor analysis may suggest erroneous factor solutions. When the dimensionality of a dataset is analyzed based on the factor analysis, oppositely worded items may form distinct item direction factors. However, before making any decision concerning the presence of method effect caused by item wording direction, it is necessary to evaluate if the dataset holds the assumptions of the factor analysis (especially the monotonicity assumption).

Thus, the utilization of mathematical models that does not assume monotonically increasing item response curves gains importance to detect possible violations of the monotonicity assumption.

One of the measurement models not assuming monotonicity is the generalized graded unfolding model (GGUM) that was developed by Roberts (1995) based on the parametric item response theory framework. This model expects an individual who has a neutral attitude toward any attitude object to strongly disagree with an extremely positive or negative item because extreme items are located far from the individual's position on the attitude continuum. When the item is much more negative than the person's attitude, then the person strongly disagrees from above the item. In contrast, if the item is much more positive than the person's attitude, then the person strongly disagrees from below the item. Therefore, there are two possible responses associated with the single observable response of strongly disagree. Thus, the model assumes that there are two latent response categories underlying an observable response category. The model estimates one discrimination parameter, one location parameter and the threshold parameter equal to the number of the response categories minus 1 for each item (Roberts, 1995; Roberts et al., 1999).

The other way of analyzing the monotonicity of item response functions (IRF) is the Mokken models based on the Nonparametric Item Response Theory (NIRT). These models included in NIRT, unlike parametric ones, do not require any restrictive assumptions about the shape of the IRFs (Sijtsma & Molenaar, 2002). The NIRT models do not provide alternatives to parametric ones, rather than they allow studying the minimum assumptions that have to be met. Thanks to these minimum assumptions, the IRFs estimated by the NIRT models may be much closer to the "true response functions". Therefore, it is useful to estimate IRFs by utilizing a NIRT model before estimating them based on parametric approach that has strict assumptions for IRFs (van Linden & Hambleton, 1999).

The Mokken models that are accepted as probabilistic forms of Guttman scaling approach estimate the relationship between the measured latent variable and the possibility of giving correct answers based on an explanatory approach rather than a deterministic way adopted by the Guttman scaling. The Mokken scaling aims to develop unidimensional scales and, in this process, the assumptions of unidimensionality and local independence, which are valid for the IRT, are required to be met. The uni-dimensionality assumption requires scale items to measure one dominant latent dimension. The local independence assumption means that the possibility of test-takers' giving corrects answer to an item is not affected by the other test items. In other words, all items of the test should be answered independently by the test-takers (Hambleton et al., 1985). In addition to these assumptions, the Mokken scaling requires the monotonicity of the IRF, but this monotonicity assumption is different from the one required by parametric models of IRT. Mokken (1999) stated this type of monotonicity as "simple monotonicity" and defined this assumption related with the local independence. Under the assumption of monotonicity, all item pairs are non-negatively correlated for all subgroups of subjects and all subsets of items.

As mentioned before, the Mokken scaling, which is different from the classical factorial analyses such as explanatory and confirmatory factor analyses, allows developing unidimensional scales. The Mokken scaling based on the NIRT approach provides several advantages to researchers (Wismeijer et al., 2008). For example, it gives not only an opportunity to investigate the dimensionality of the latent structure but also allows analyzing psychometric qualities of unidimensional scales based on more basic and less restrictive assumptions (Sijtsma & Molenaar, 2002).

It is important to select appropriate measurement models and statistical techniques that fit the data structure, because, as stated by Tay and Drasgow (2012), inappropriate measurement

models may affect inferences of construct dimensionality. The factor analysis of the attitude scale, which was utilized in the current study, suggested two factors each including items written in one direction. However, it is necessary to examine whether the two-factor solution of the factor analysis may be caused by the method effect, or by the failure of the analysis in describing and fitting the dataset because of the monotonicity assumption. Thus, this study aims to investigate the effects of violations from the assumption of monotonicity on the determination of factor structure of a scale. In the current study, the data obtained from answers provided by the students to the Attitude Scale Towards Foreign Languages as Medium of Instruction was examined to reveal to what extent the data meet the monotonicity assumption of the factor analysis. Accordingly, the current study examines the fit of the scale items to the two-item response theory models (the generalized graded unfolding model and the Mokken model of the non-parametric item response theory) that allow modeling non-monotonic item response curves.

## **2. METHOD**

The current study is a fundamental one that aims to investigate the effects of violations from the assumption of monotonicity on the determination of factor structure of a scale. While doing this, the two IRT models were utilized and the results of the analyses were compared.

### **2.1. Participants**

The present study was conducted on the data obtained from 355 students who were studying at the Faculties of Education (73 students), Arts and Science (139 students), and Economic and Administrative Sciences (143 students) of the Middle East Technical University (METU) during the 2012-2013 academic year. The reason of selecting the participants from this university was that the METU is one of the oldest universities that have been using English as the medium of instruction. 88 students were freshmen, 133 of them were sophomores, 68 students were juniors, and lastly, 66 of them were seniors. 243 out of 355 students were female, while 112 of them were male.

### **2.2. Research Instruments**

The data were obtained by conducting the Attitude Scale Towards Foreign Languages as Medium of Instruction, which was developed by Kartal and Gülleroğlu (2015). The scale included 10 positively and 10 negatively worded items. Students gave answers to the scale items on a five-point Likert scale. The item-total correlations calculated for the items varied between 0.43 and 0.76. The t-tests values of the total scores of bottom 27% and top 27% of participants for each item were significant and high. The exploratory factor analysis was carried out to examine the construct validity of the scale. The eigenvalues suggested a three-factor structure, but the scree plot revealed that the scale had a two-factor structure. To make a decision on the factor numbers of the scale, the distribution of the items into the factors were examined. As a result, it was found that only one item belonged to the third factor, while the positively and negatively worded items belonged to the first and the second factor, respectively. The Cronbach alpha correlation coefficient of the scale was calculated as 0.92. It is over the accepted lower boundary for the reliability, which is 0.70-0.80 (Reise & Revicki, 2015).

### **2.3. Data Analysis**

The fit of the scale items to the generalized graded unfolding model (GGUM) was examined based on the item response curves, item parameters, item fit statistics and fit graphics. The adjusted  $\chi^2/df$  ratios were analyzed to evaluate item level model data fit (Carter et al., 2015; Studts, 2008; Speer et al., 2016). The adjusted  $\chi^2/df$  ratio lower than 3 was accepted as an evidence for item fit (Chernyshenko et al., 2007). The researchers recommend to utilize the statistical and graphical techniques together to examine item level model data fit



(Chernyshenko et al., 2001). Thus, the fit of the GGUM to the scale items were evaluated based on the fit graphics in addition to the fit statistics. To obtain item fit graphs, respondents are ranked order according to their trait levels and homogeneous clusters of approximately equal size are formed. Then, the mean estimated trait level values in each cluster are plotted against both the average observed and average expected item response for that cluster (Roberts, 2016). In addition, as recommended by Roberts (2016), the fit between the content of each item and item location determined by the location parameters estimated by the GGUM was examined. The MODFIT1.1 statistical program developed by Stark (2001) was utilized to estimate the adjusted  $\chi^2/df$  ratios and to plot item fit graphics. The “GGUM” package, developed by Tendeiro and Castro-Alvarez (2019), on the R program was utilized to estimate the item parameters.

In order to analyze the fit of the scale items to the Mokken models, firstly, the suitability of the data set for the Mokken model analyses was checked. The outliers and extreme values were investigated. The number of Guttman errors was calculated to control outliers (Zijtstra et al., 2011), and then scalability coefficients were calculated at the scale, ( $H$ ), item ( $H_i$ ), and item-pair level ( $H_{ij}$ ) levels. For scalability coefficients, the lower bound was accepted as 0.3. The related researches strongly emphasize to select items with scalability coefficients higher than 0.3 (Meijer et al., 2015). However, Egberink and Meijer (2011) stated that very high  $H_i$  coefficients may not be accepted, too. Items with too high  $H_i$  coefficients may be the results of repeating the same items and deteriorated validity of the scales. Therefore, the  $H_i$  coefficients should be interpreted carefully. The Automated Item Selection Procedure (AISP) was conducted to investigate the unidimensionality of the data. The conditional covariance values were analyzed and then the monotonicity analyses were conducted by composing the IRF with nonparametric regression method to examine the local independence assumption. In addition to the graphical analyses, the monotonicity of the IRFs was investigated with the significance tests. To determine the model-data fit, the last assumption of Mokken models, invariant item ordering, was analyzed for the data set. For this assumption, the P-matrix and the rest-score method were used. In addition, the  $H^T$  coefficient proposed by Ligtoet (2010) showing the accuracy of item ordering was calculated. The critical values in evaluating the violations from the invariant item ordering and monotonicity assumptions was accepted as 80, which is called as *Crit* values. The *Crit* values lower than 40 indicate no serious violations. The *Crit* values between 40-80 indicate minor violation, and they are acceptable. However, the *Crit* values higher than 80 indicate serious violations, and the items with higher *Crit* values than 80 are omitted from the scale (Junker & Sijtsma, 2001; Molenaar & Sijtsma, 2000). The researchers stated that the *Crit* values should be interpreted carefully by taking into consideration the results obtained from other methods (Meijer et al., 2015). Accordingly, in the current study, the results from the P-matrix method, the rest-score method and the  $H^T$  coefficients were used together to evaluate the assumption of invariant item ordering. The “mokken” package, developed by Van der Ark (2007), on the R program was utilized to analyze the fit of Mokken models.

### 3. FINDINGS

Item response curves and item parameters were estimated to examine the fit of the scale items to the GGUM. When item response curves were analyzed, it was found that 7 (item number 2, 3, 5, 7, 12, 15, 18) out of 10 negatively worded items had monotonic response curves, while all of positively worded scale items had non-monotonic response curves. Thus, the findings revealed that 13 out of 20 items had non-monotonic response curves. This finding indicated that there were non-monotonic relations between item responses and respondent’s trait levels on most of the scale items. Since the GGUM can model non-monotonic relations between the item response and the latent trait, it can be stated that the GGUM is an appropriate alternative to model the item responses provided by the respondents to the scale items.



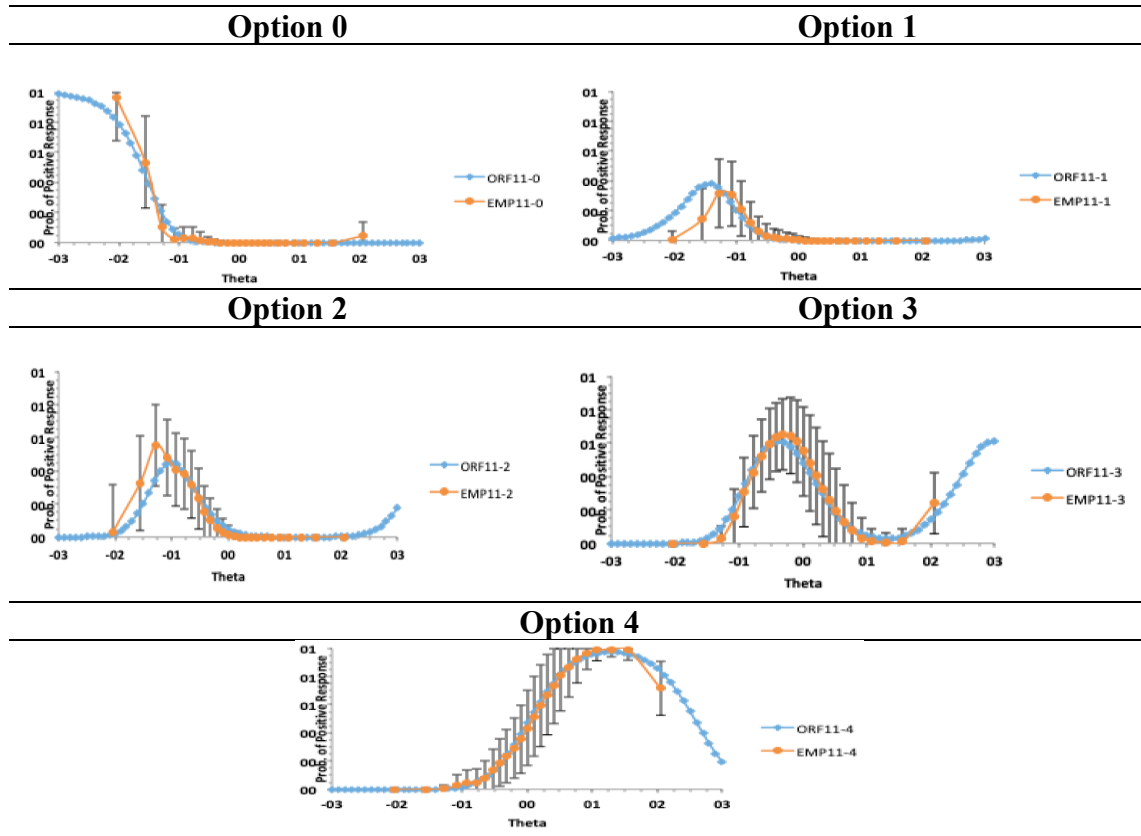
Item location parameters estimated for the items by the GGUM were examined to evaluate item level model data fit. As stated by Roberts (2016), the most basic diagnostic of GGUM performance with a data set is simply to rank items according to their location parameters and then evaluate whether the content of each item makes sense with the associated item location. Item contents should flow from very negative, moderately negative, neutral, moderately positive, and very positive expressions with respect to attitude object. Accordingly, the results revealed that item location parameters estimated for negatively worded items varied between -4.98 and -2.48. The location parameters of positively worded items varied between 0.99 and 1.49. Item location parameters indicated that negative items located on the more extreme end of the attitude continuum, while positive items located on an area representing more moderate positive attitude. Furthermore, it was found that the item contents were in line with the item location parameters. For example, negatively worded items generally had more extreme expressions and represented very negative attitude towards using English as medium of instruction. However, the positively worded items had more moderate expressions. In addition, the researchers expected to have no items whose location parameter were between -1 and +1, because there was not any item representing the neutral attitude towards the attitude object. In parallel with this expectation, it was found that there was no item, which had a location parameter between -1 and +1. The location parameters provided evidences for that the GGUM was able to estimate item parameters that were consistent with the item contents.

The item level model data fit was examined based on the both the statistical and the graphical techniques. As mentioned before, the adjusted  $\chi^2/df$  ratios were analyzed to evaluate item fit. The findings indicated that 13 out of 20 items had ratios lower than 2, and 2 items had ratios lower than 3. The adjusted  $\chi^2/df$  ratios of the remaining 5 items were higher than 3. The GGUM provided fit to the 15 out of 20 scale items. Item fit graphics were also examined to determine the item level fit of the GGUM. In line with the statistical findings, fit graphs plotted for the items, having ratios lower than 3 supported that the GGUM provided fit to 15 scale items. In addition, the fit graphs of the 5 items accepted as unfit based on their adjusted  $\chi^2/df$  ratios revealed that these items also fitted to the GGUM. To provide an example, the fit plots for five response categories of item 11, which had the highest the adjusted  $\chi^2/df$  ratio (21.23) are given in [Figure 1](#).

In the fit plots given in [Figure 1](#), vertical lines correspond to the 95% confidence interval for the observed response ratios. If the response ratios estimated by the GGUM do not overlap with the confidence interval for the observed ratios, then, this indicates that the GGUM does not fit to this specific scale item (Chernyshenko et al., 2001). As [Figure 1](#) indicates, the GGUM provided consistent estimations with the observed response ratios. Except for only one response category (option 1), the estimated response ratios of the remaining response categories overlap with the confidence interval of the observed response ratios.

In addition to the GGUM analysis, the NIRT analyses were also conducted to investigate monotonicity and dimensionality of the scale. The first step of the NIRT application is the estimation of scalability coefficients. The scalability coefficients were estimated at three levels; item, item pairs and scale. Firstly, the item-pair scalability coefficients ( $H_{ij}$ ) were analyzed, and it was found that all of them were positive. This finding is a pre-requisite and the very first step of the Mokken scaling. If there is any negative value among the  $H_{ij}$  coefficients, the scale is evaluated as not suitable for the Mokken models (Sijtsma & Van der Ark, 2017). Secondly, the item level scalability coefficients,  $H_i$ , were estimated and these values are given in [Table 1](#).

Figure 1. The Fit Plots for Item 11.



When the item scalability coefficients given in Table 1 were investigated, it was found that 19 out of 20 items had higher values than the cut off value, which was 0.30. The item 12 was the only item having coefficient lower than 0.30. Based on the item scalability coefficients, 19 items were found suitable for the Mokken scaling. These coefficients provided information about the item discrimination levels. Items with higher  $H_i$  coefficients are more discriminative than the items having lower coefficients. Accordingly, items with a  $H_i$  value between 0.3 and 0.4 are considered weak, items with a value between 0.4 and 0.5 are considered to be medium and items with a value greater than 0.5 are accepted as high discriminative (Sijtsma & Molenaar, 2002; Sijtsma & van der Ark, 2017). Based on these values, it was revealed that only one item (12) had low, six items had weak, 12 items had moderate, and only one item had high level of discrimination power.

Table 1. The Item Scalability Coefficients.

Item Number	$H_i$	Item Number	$H_i$
1	0.44	11	0.50
2	0.47	12	0.27
3	0.49	13	0.45
4	0.35	14	0.37
5	0.30	15	0.41
6	0.33	16	0.36
7	0.43	17	0.39
8	0.34	18	0.44
9	0.40	19	0.45
10	0.44	20	0.45

Thirdly, the scale level of scalability coefficient was calculated, and this value was also evaluated based on the aforementioned cut off values. The  $H$  value of the scale estimated as 0.41. This value indicated that the scale was moderately adapted to the Mokken scaling. The  $Z$  statistics were calculated for the significance of scalability coefficients. As a result, it was found that all of the  $Z$  values were greater than 0. Therefore, it was concluded that the scalability coefficients were greater than 0 and significant not only for the sample but also for the population.

The second step of the Mokken scaling is to check the unidimensionality of the data. The Automated Item Selection Procedure (AISP) was used for this analysis (Sijtsma & Van der Ark, 2017). As a result of the AISP, it was determined that there was a single factor underlying the data, but items 5 and 12 did not fit to the unidimensional structure proposed by the AISP. It was previously determined that the  $H_i$  coefficient of the item 12 was lower than the cutoff value, and the  $H_i$  coefficient of item 5 was at the boundary value level. According to these results, it can be concluded that the scale is compatible with unidimensional structure, except for the two items. The monotonicity assumption was examined for the scale to provide extra evidences for the dimensionality of the data. This assumption was investigated based on the graphical and statistical methods. The graphical analyses were conducted based on the item step functions and item response functions which were formed depending on the rest-score groups method. In addition, violations from the assumption of monotonicity were also investigated based on the statistical tests. The results obtained from the monotonicity analyses are given in [Table 2](#).

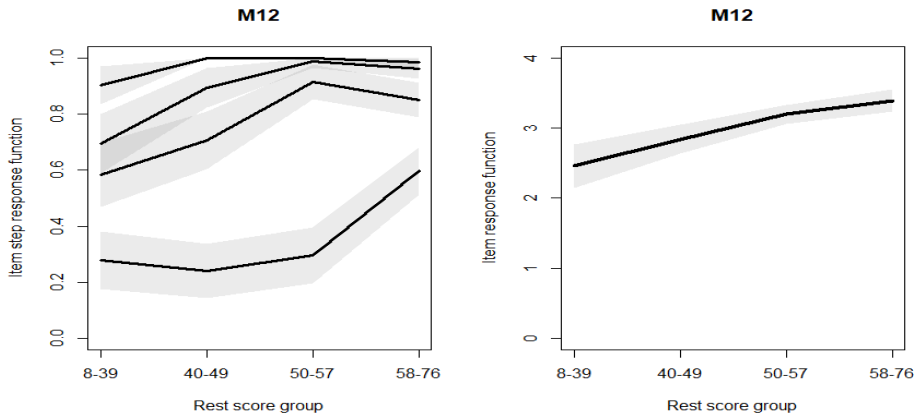
**Table 2.** *The Results of Monotonicity Analyses.*

Items	$H_i$	#vi	zsig	crit
1	0.44	0	0	0
2	0.47	0	0	0
3	0.49	0	0	0
4	0.35	2	0	38
5	0.30	0	0	0
6	0.33	0	0	0
7	0.43	0	0	0
8	0.34	0	0	0
9	0.40	0	0	0
10	0.44	1	0	9
11	0.50	0	0	0
12	0.27	2	0	32
13	0.45	0	0	0
14	0.37	0	0	0
15	0.41	0	0	0
16	0.36	0	0	0
17	0.39	0	0	0
18	0.44	0	0	0
19	0.45	0	0	0
20	0.45	0	0	0

In [Table 2](#), the  $H$  values correspond to the item level scalability coefficients, #vi indicates the number of violations from the monotonicity assumption, and the  $zsig$  values display the significance of the violation. The last value is the  $crit$ , and it indicates the significance levels of the violation from the assumption. When the values in [Table 2](#) were analyzed, it was found that items 4, 10 and 12 had some violations from the monotonicity assumptions. However, the  $crit$  values of these violations indicated that these violations were below the critical value of 80.

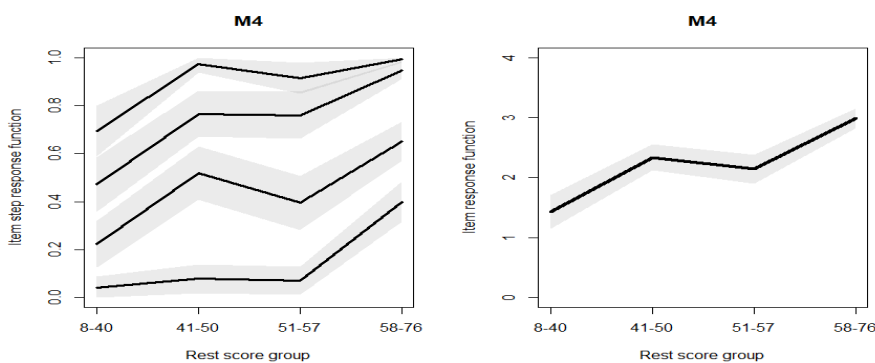
Therefore, it can be concluded that the monotonicity assumption was met for the scale items. The item step and response functions were also examined. The item step and response functions of item 12, which was detected as having minor violation from the monotonicity, are given in Figure 2.

Figure 2. The Item Step Function and Item Response Function of Item 12.



When Figure 2 was analyzed, it was found that there were some violations from the monotonicity in both functions. According to the item step function, the function increased monotonously in the transition from the first category to the second category, but the function decreased especially in the high scores in the second and third step functions. When item response function was examined, a decrease in the total score ranging from 50 to 57 was observed, but the decrease did not continue throughout the all score groups. As the level of having the measured trait increased, the probability of answering the item as “5-totally agree” increased, too, as it was expected. Consequently, for this item, these graphical analyses supported the statistical findings of the monotonicity analyses, and the graphs indicated that there were some violations, but these violations were negligible. The item step and response functions of 4, which had some violations from the monotonicity assumption were given in Figure 3.

Figure 3. The Item Step Function and Item Response Function of Item 4.



In Figure 3, it is clear that there are several decreases in both functions. According to the item step and response function, the function decreased in the second and fourth step functions, and this finding indicated that the item violated the monotonicity assumption. However, the results of the statistical test revealed that these violations were negligible. Considering both the statistical and graphical analyses, it was found that even if there were several violations from the monotonicity assumption, the assumption was met for most of the scale items, and this scale can be scaled based on the Monotone Homogeneity Model (MHM), which allows a flexible and unidimensional scaling in the NIRT approach.

To determine whether the scale fit to the MHM, the Mokken scale investigation was continued with the last assumption of the NIRT models, which is invariant item ordering. Invariant item ordering is a prerequisite for the strict model of the Mokken scaling, which is the Double Monotonicity Model (DMM). This model allows ordering not only the person regarding to their traits, but also the items regarding to their difficulty levels. This assumption was checked based on the P-matrix method. In addition, the  $H^T$  coefficient was estimated in order to check the accuracy of item ordering. The results of the analyses are presented in [Table 3](#).

**Table 3.** *The Results of the Analysis of Invariant Item Ordering Assumption.*

Items	$H_i$	#vi	t-sig	Crit
18	0.44	1	0	21
5	0.30	1	0	28
9	0.40	2	0	34
11	0.50	4	1	95
3	0.49	7	5	176
1	0.44	3	2	100
13	0.45	5	2	109
6	0.33	3	2	126
8	0.34	4	3	117
17	0.39	2	0	49
16	0.36	4	1	100
7	0.43	3	1	62
2	0.47	2	0	35
19	0.45	3	1	72
14	0.37	3	0	47
4	0.35	5	3	114
20	0.45	3	1	82
15	0.41	2	1	76
10	0.44	0	0	0

In [Table 3](#), the item scalability coefficients- $H_i$ , the number of violations -#vi, the critical values of violations-*Crit* and the t values estimated for violations were presented. Item 12 was excluded from the scale as it had been found as misfit to the Mokken scaling. After the exclusion of item 12 from the scale, the scalability coefficient of item 5 increased (0.30). Hence, there was no need to remove this item from the scale. The critical value was accepted as 80 in the evaluations of violation. Accordingly, 9 out of 19 items were detected violating the assumption seriously. These violations were higher than the critical value, therefore, it was concluded that invariant item ordering assumption was not met for the items. It was concluded that the scale items may not be scaled based on the Double Monotonicity Model. In addition to the results provided by the P-matrix, the  $H^T$  coefficient was calculated as 0.207, which was lower than the boundary level. This finding supported the P-matrix results and it was concluded that the scale items did not have the feature of invariant item ordering.

After item 12 excluded from the scale, all Mokken scaling analyses were repeated for the revised form of the scale, and it was found that there was an increase in the scalability coefficients both at the item and at the scale levels. The H coefficient increased from 0.40 to 0.42, while no improvements were found for the monotonicity and invariant item ordering assumptions. Consequently, the 19-item scale was found suitable to be scaled based on the MHM. The last examination of the 19-item scale was the estimation of reliability coefficients. Four different coefficients were estimated for the reliability of the scale, and the results are presented in [Table 4](#).

**Table 4.** *The Reliability Coefficients.*

Coefficients	MS	Cronbach Alfa	Lambda2	LCRC
	0.923	0.921	0.924	0.929

In **Table 4**, the MS (Molenaar- Sijtsma) coefficient is a coefficient utilized in the Mokken scaling. The Lambda2 is a coefficient that is related to the Guttman errors. The third one is the LCRC (Latent Class Reliability Coefficient), and it gives information about the accuracy of the latent classification. When the values were analyzed, it was found that all of the coefficients were higher than 0.90. Based on the findings, it was concluded that the reliability of measurement was high, since all of the coefficients were higher than 0.70, which is widely accepted lower boundary for the reliability.

#### **4. DISCUSSION and CONCLUSION**

The current study examined the fit of the scale items to the item response theory models that do not assume monotone increasing item response curves for items. Accordingly, the dimensionality of the scale data was analyzed based on the generalized graded unfolding model and the Mokken Model of non-parametric item response model. Based on the item parameters, item response curves, item fit graphics and statistics estimated by the GGUM, it was concluded that the scale items fit to the model. The exploratory factor analysis, which assumes monotonic relations between the trait levels of individuals and their item responses, suggested a two-factor structure for the scale items. The results provided by the factor analysis indicated that individuals' item responses were affected not only by their attitude towards using English as medium of instruction but also the wording direction of the scale items. However, the current study revealed that the scale items provided fit to the GGUM, which is a unidimensional item response theory model.

The GGUM that takes account the non-monotonic item characteristic curves suggested a unidimensional structure for the data. Supportively, based on the results provided by the non-parametric item response theory, it was concluded that the attitude scale items fit to the MHM and there is one latent dimension underlying the responses given to the scale items. This finding is in line with the results provided by the GGUM. The non-parametric item response model and the GGUM confirmed that the data has a unidimensional structure, while the factor analysis suggested a two-factor-structure for the same data. It was found that the data fit to a unidimensional model if that model allows modeling non-monotonic response curves.

The results of the studies carried out on different scales measuring various affective traits are in line with the findings of the current study. For example, Van Schuur and Kiers (1994) revealed that the correlations matrices provided by the non-monotonic and monotonic measurement models differ from each other. The researchers state that the differences observed on the matrices affect the findings concerning the dimensionality of the data, and because of monotonicity assumption, researchers have results supporting multidimensionality for a data set that is actually unidimensional. Supportively, Spector et al., (1997) stated that monotonic analyses such as the factor analysis may suggest multidimensional structures for the data that is, in fact, explained by one dominant dimension. Tay and Drasgow (2012) examined the effect of the monotonicity assumption on the dimensionality analysis. The researchers carried out the principal components factor analysis on the data simulated based on the GGUM. As a result, the factor analysis suggested a two-factor-structure for the data, which is, in fact, unidimensional. The researchers accepted this finding as an evidence for that the utilization of a measurement model that cannot model the possible non-monotonicity observed in the data may cause incorrect inferences concerning the dimensionality of the data. The researchers recommend to reexamine the structure of the data by taking into consideration the monotonicity



assumption when the application of the factor analysis yields two dimensions that are defined by the conceptual ends of the unipolar construct (i.e., nonoccurrence and frequent occurrence; nonexistent and extreme).

The related studies (Spector et al., 1997; Tay & Drasgow, 2012; Van Schuur & Kiers, 1994) revealed that when a scale includes both positively and negatively worded items, the factor analysis may sometimes suggest two separate factors one of which includes only negatively worded items and the other one includes only positively worded items, while there is actually one dominant latent dimension underlying the scale items. Supportively, the results of the current study indicated that the items of the Attitude Scale Towards Foreign Languages as Medium of Instruction fit to the unidimensional models that do not assume monotone increasing item response curves, while the factor analysis suggested a two-factor solution for the same data. Based on this finding, it is necessary to note that the dimensionality analyses assuming monotonic relations between the latent trait and item responses may not always provide the best description for the structure of the data. Therefore, researchers are recommended to utilize statistical techniques that can identify any possible violation of the monotonicity assumption and model items having non-monotonic response curves, especially when they aim to examine dimensionality of the data obtained from a measurement tool containing both negatively and positively worded items.

#### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

#### Authorship contribution statement

**Seval Kula Kartal:** Investigation, Resources, Analysis based on the generalized graded unfolding model, Writing the original draft. **Ezgi Mor Dirlik:** Investigation, Analysis based on the non-parametric item response theory model, Writing the original draft.

#### ORCID

Seval Kula Kartal  <https://orcid.org/0000-0002-3018-6972>

Ezgi Mor Dirlik  <https://orcid.org/0000-0003-0250-327X>

#### 5. REFERENCES

- Carter, N. T., & Dalal, D. K. (2010). An ideal point account of the JDI work satisfaction scale. *Personality and Individual Differences, 49*, 743-748.
- Chernyshenko, O. S., Stark, S. E., Drasgow, F., & Roberts, J. S. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment, 19*(1), 88-106.
- Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*(4), 523-562.
- DiStefano, C., & Motl, R. W. (2006) Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling, 13*(3), 440-464.
- Gorsuch, R. L. (1983). *Factor analysis*. Saunders.
- Gu, H., Wen, Z., & Fan, X. (2015). The impact of wording effect on reliability and validity of the Core Self-Evaluation Scale (CSES): A bi-factor perspective. *Personality and Individual Differences, 83*, 142-147.

- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1985). Principles and applications of item response theory. SAGE Publications, Inc.
- Horan, P. M., DiStefano, C., & Motl, R. W. (2003) Wording effects in self-esteem scales: Methodological artifact or response style?. *Structural Equation Modeling*, 10(3), 435-455.
- Junker, B. (2000). *Some topics in nonparametric and parametric IRT, with some thoughts about the future*. Unpublished manuscript. Carnegie Mellon University.
- Junker, B. W., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement*, 25(3), 211-220.
- Ligtvoet, R., Van der Ark, L. A., Te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement*, 70(4), 578-595.
- Meijer, R. R., & Egberink, I. J. (2011). Investigating invariant item ordering in personality and clinical scales: some empirical findings and a discussion. *Educational Testing and Measurement*, 20(10), 589-607.
- Meijer, R. R., Tendeiro, J. N., & Wanders, R. B. (2014). The use of nonparametric item response theory to explore data quality. In *Handbook of Item Response Theory Modeling* (pp. 103-128). Routledge.
- Reise, S. P., & Revicki, D. A. (2015). *Handbook of item response theory modeling*. Taylor & Francis Group.
- Roberts, J. S. (1995). *Item response theory approaches to attitude measurement* [Doctoral dissertation, University of South Carolina, USA].
- Roberts, J. S. (2016). Generalized graded unfolding model. W. J. van der Linden (Eds.) *Handbook of item response theory volume one: Models*. (pp. 369-393). Taylor & Francis Group.
- Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (1999). *Estimating parameters in the generalized graded unfolding model: Sensitivity to the prior distribution assumption and the number of quadrature points used*. Paper presented at the Annual Meeting of the National Council on Measurement in Education.
- Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory* (Vol. 5). Sage Publications.
- Sijtsma, K., & Van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, 70(1), 137-158.
- Spector, P. E., Katwyk, P. T., Brannick, M. T., & Chen, P. Y. (1997). When two factors don't reflect two constructs: How item characteristic can produce artifactual factors? *Journal of Management*, 23(5), 659-677.
- Speer, A. B., Robie, C., & Christiansen, N. D. (2016). Effects of item type and estimation method on the accuracy of estimated personality trait scores: Polytomous item response theory models versus summated scoring. *Personality and Individual Differences*, 102, 41–45.
- Stark, S. (2001). *MODFIT: A computer program for model-data fit*. University of Illinois at Urbana-Champaign.
- Stevens, J. (1996). *Applied multivariate statistics for the social science*. Lawrence Erlbaum Associates.
- Studts, C. R. (2008). *Improving screening for externalizing behavior problems in very young children: Applications of item response theory to evaluate instruments in pediatric primary care* [Doctoral dissertation, University of Louisville]. <https://kb.osu.edu/>

- Supple, A. J., & Plunkett, S. W. (2011). Dimensionality and validity of the Rosenberg Self-Esteem Scale for use with Latino adolescents. *Hispanic Journal of Behavioral Sciences*, 33(1), 39-53.
- Tay, L., & Drasgow, F. (2012). Theoretical, statistical, and substantive issues in the assessment of construct dimensionality: Accounting for the item response process. *Organizational Research Methods*, 15(3), 1-22.
- Tendeiro, J., & Castro-Alvarez, S. (2019). GGUM: An R package for fitting the generalized graded unfolding model. *Applied Psychological Measurement*, 43(2), 172-173.
- Thomas, J. M., & Oliver, A. (1999) Rosenberg's self-esteem scale: Two factors or method effects. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 84-98.
- Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of statistical software*, 20(11), 1-19.
- Van der Linden W. J. & Hamleton, R.K. *Handbook of modern item response theory* (1997). Springer-Verlag.
- Van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis*, 11(2), 139-163.
- Van Schuur, W. H., & Kiers, H. A. L. (1994). Why factor analysis often is the incorrect model for analyzing bipolar concepts, and what model to use instead? *Applied Psychological Measurement*, 18(2), 97-110.
- Wang, J., Siegal, H. A., Falck, R. S., & Carlson, R. G. (2001) Factorial structure of Rosenberg's Self-Esteem Scale among crack-cocaine drug users. *Structural Equation Modeling*, 8(2), 275-286.
- Wang, Y., Kim, E. U., Dedrick, R. F., Ferron, J. M., & Tan, T. (2018). A multilevel bifactor approach to construct validation of mixed-format scales. *Educational and Psychological Measurement*, 78(2), 253-271.
- Wismeijer, A. A., Sijtsma, K., van Assen, M. A., & Vingerhoets, A. J. (2008). A comparative study of the dimensionality of the self-concealment scale using principal components analysis and Mokken scale analysis. *Journal of Personality Assessment*, 90(4), 323-334.
- Wouters, E, Booyesen, F. L. R., Ponnet, K., & Baron, Van Loon, F. (2012). Wording effects and the factor structure of the Hospital Anxiety & Depression Scale in HIV/AIDS patients on antiretroviral treatment in South Africa. *PLoS ONE*, 7(4), 1-10.
- Zijlstra, W. P., Van der Ark, L. A., & Sijtsma, K. (2011). Robust Mokken scale analysis by means of the forward search algorithm for outlier detection. *Multivariate behavioral research*, 46(1), 58-89.

## Principles for Minimizing Errors in Examination Papers and Other Educational Assessment Instruments

Irenka Suto<sup>1,\*</sup>, Jo Ireland<sup>1</sup>

<sup>1</sup>Cambridge Assessment, UK

### ARTICLE HISTORY

Received: Dec. 03, 2020

Revised: Mar. 11, 2020

Accepted: Mar. 15, 2021

### Keywords:

Error,

Fairness,

Instrument design,

Examination design.

**Abstract:** Errors in examination papers and other assessment instruments can compromise fairness. For example, a history question containing an incorrect historical date could be impossible for students to answer. Incorrect instructions at the start of an examination could lead students to answer the wrong number of questions. As there is little research on this issue within the educational assessment community, we reviewed the literature on minimizing errors in other industries and domains, including aviation, energy, and medicine. We identified generalizable principles and applied them to our context of educational assessment. We argue that since assessment instrument construction is a complex system comprising numerous interacting components, a holistic approach to system improvement is required. Assessment instrument errors stem primarily from human failure. When human failure occurs, it is not good enough to suggest that ‘to err is simply human’. Instead it is necessary to look deeper, evaluating the latent working conditions that underpin the efficacy of procedures, making the human failure more or less likely. Drawing from the aviation industry’s ergonomic SHELLO model, we articulate and explore three of the most critical working conditions that relate to our context: (i) time pressure, (ii) workload and stress, and (iii) wider organizational culture, including good error data collection. We conclude with recommendations for best practice in minimizing errors in assessment instruments. A ‘good’ error culture should be promoted, which avoids blaming individuals. Errors should be acknowledged readily by all, and system owners should take a scientific approach to understanding and learning from them.

“Science, my lad, is made up of mistakes, but they are mistakes which it is useful to make, because they lead little by little to the truth.”

(Jules Verne, 1864)

## 1. INTRODUCTION

As well as motivating students to pursue their ambitions, fair assessments can build trust and confidence in education systems within society at large (Nisbet & Shaw, 2020). To date, much research on improving assessment instruments has focussed upon the key topics of validity and reliability. An additional but oft-overlooked aspect of fairness is the presence of errors in final

---

\*CONTACT: Irenka Suto ✉ [suto.i@cambridgeassessment.org.uk](mailto:suto.i@cambridgeassessment.org.uk) 📍 Cambridge Assessment, The Triangle Building, Shaftesbury Road, Cambridge, CB2 8EA.

or near-final versions of assessment instruments<sup>†</sup> (Baranowski, 2006; Rhoades & Madaus, 2003; Rodriguez, 2015). For example, a simple typographical error could make an examination question unanswerable. A multiple-choice question could have two correct response options, confusing students, or even no correct response options. Some missing information could make a question harder to answer than intended. Faulty instructions at the start of an assessment could lead students to answer the wrong number of questions.

In many assessment contexts, due to high levels of professionalism, errors of this kind are an infrequent albeit longstanding problem. In England and Wales, for example, the vast majority of high-stakes examinations for secondary school students are error-free; the national regulator reported that in 2019, just 71 errors were identified across more than 6300 question papers, non-exam assessments and materials administered that year (Ofqual, 2019). However, occasionally errors are reported in the UK's national media, (for example, Richardson, 2017; Meredith, 2019), and as in South Korea and New Zealand (New Straits Times, 2015; BBC, 2017) some error incidents have led to public outcries. This is because even rare errors can have wide-ranging and unpredictable impacts on students. Their anxiety levels can be affected, as can time management, and therefore their general performance during the assessment. Ultimately, students' life chances can be damaged. It is clear that whether instruments are summative or formative, paper-based or computer-based, innovative or traditional, and whether they are created by teachers, teams within assessment organizations, national experts, or others, the assessment community should strive to make them free from errors.

In this paper, we argue that each assessment instrument error should be viewed not merely as the result of human error about which little can be done, but as a symptom of a deeper and more complex problem which spans the international assessment community. As Dekker (2002a) points out:

“Although it is a forgiving stance to take, organizations that suggest that ‘to err is simply human’ may normalise error to the point where it is no longer interpreted as a sign of deeper trouble.” (Dekker, 2002a, p. 145).

We take some of the first steps in understanding why errors occasionally occur in assessment instruments, and why the detection of errors can be slow despite the numerous checks included in most construction processes. We draw upon the wealth of research literature on error reduction that exists in complex sectors such as medicine, manufacturing, the nuclear industry, and aviation, extracting those principles that generalise across contexts. In recent decades, greater understanding of how and why errors occur in these domains has been credited with significant improvements in safety as well as quality, saving countless lives. There is a clear opportunity for educational assessment professionals to utilise this considerable body of knowledge too.

In general, there are three main strategies for addressing the problem of errors. First, make fewer errors in the first place. Secondly, detect more of the errors that do arise, and do so rapidly; that is, make fewer errors in detecting errors. Thirdly, improve methods of negating any undesirable consequences of errors. In our context, a critical limitation of the third approach is the impossibility of mitigating the impact of an error at the level of the individual. Whilst one student might be badly confused, upset and/or delayed by a particular error, another might not even notice it. Giving all students in a cohort full marks (or no marks) for a question containing

<sup>†</sup> It is the norm for errors to arise and to be corrected rapidly during the (typically iterative) early development phase of instrument creation, during which teams of professionals work to maximise the quality of items and instructions. This article focuses *not* on these ‘early’ errors, but on the small minority of errors that evade detection during revision and checking procedures, to reach or almost reach students.



an error is therefore too crude a remedy, and even the most sophisticated statistical methods cannot identify how particular individuals have been affected. We therefore focus on the first two of these strategies. Our overarching goal has been to identify some key principles for best practice in minimizing errors in assessment instruments.

## **2. COMPLEX SYSTEMS**

Many systems through which errors arise and are detected (or not) are complex. Oates (2017) draws from Mitleton-Kelly (2003) to explain an important distinction between complex systems and complicated systems in education:

“Complicated systems have many parts and many interactions, but give predictable outcomes. A chronograph is complicated, but gives a highly regulated and consistent output: a measurement of time. By contrast, complex systems possess a large number of interacting components, with outcomes which are not a simple function of the interaction of the parts.”  
(Oates, 2017, p. 9)

Oates (2017) goes on to argue that educational systems in all countries are complex and there is no single aspect of innovation which will secure a perfect system. Instead of cherry-picking initiatives from other contexts, a holistic approach to system improvement is required in which all of the components of the system are identified and included in its initial analysis.

To our knowledge, comprehensive analysis of systems in which assessment instruments are constructed and checked for errors has yet to happen. Instead, systems have evolved via the addition of extra checks, often in direct response to errors reaching students (for example, Harrison, 2011). This is partly because poor performance, when noticed, frequently calls for a rapid response. In the UK, assessment organizations must be seen by everyone, from students and teachers, to the national regulator, to the general public, to be doing something tangible and immediate to address the problem in the system. This is likely to be the case in other assessment contexts too. Also, it is usually easier to focus on one or two components of a system than to attempt its complete review, which risks the potential consequence of a complete overhaul being recommended.

One cumulative effect of multiple ‘add-an-extra-check’ initiatives is the diffusion of responsibility that each checker experiences. A second cumulative effect can be a cumbersome, costly, and overstretched construction process in which deadlines are hard to meet. When the process eventually approaches breaking point, consequent initiatives then reverse the direction of travel by focusing on streamlining and reducing the activities within the process. It is easy to envisage the overall state of affairs as a pendulum of change swinging slowly back and forth in response to external pressures.

In complex systems, this initiative-based approach to innovation and reform is risky. Without rigorous experimental trialling and evaluation (the gold standard of which is widely considered to be the Randomised Controlled Trial (RCT) approach used in medicine), it is impossible to conclude whether or not any particular action results in fewer errors. What is of greater concern, however, is that if action is taken prior to full analysis, then it can result in an inadequate response to the real causes of poor performance. Moreover, through its implementation, premature action can affect the system in unforeseen ways, creating new problems rather than remedying existing ones (Oates, 2017). To give a simple example, suppose an assessment organization relies upon outdated computer software which staff find laborious and unintuitive to use. Instigating yet another check of an already much-checked assessment instrument could lead to administrative overload for whoever is organizing the process and inadequate time to complete all checks, resulting in other errors. Achtenhagen (1994) developed an influential ‘cycle of planned failure’ to describe this kind of problem (Figure 1).



**Figure 1.** Achtenhagen's (1994) cycle of planned failure, as reported in Oates (2017).



Arguably, the key to breaking this cycle of planned failure is a comprehensive analysis of the problem, which includes the identification of less visible components of the system. That is, it is crucial to consider the covert contributors to poor performance and not just the most obvious ones. Complete comprehensiveness is an enormous and elusive research ambition. However, the general approach of gaining a better overview of the system is one that we have sought to apply here to the problem of assessment instrument errors. We have begun to identify and articulate components and the more covert ones in particular, in an attempt to look more widely than has been done in previous efforts to minimize errors.

### 3. THE SWISS CHEESE MODEL

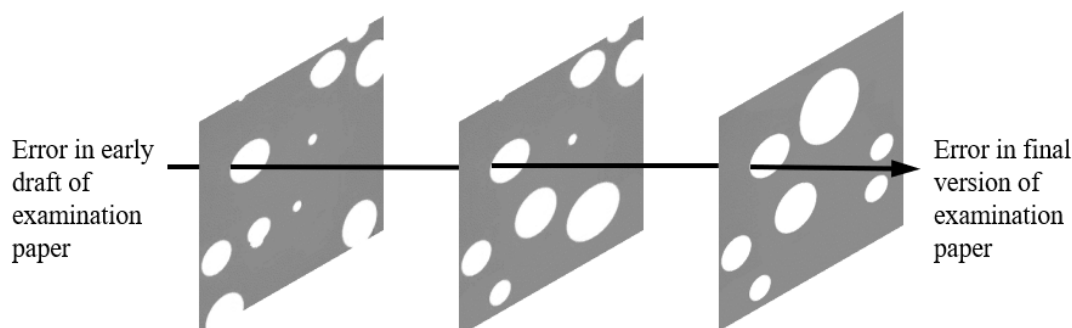
The pre-eminent name in research on the causes and detection of error is James Reason (e.g. 1990, 2013) whose theories have been applied in a range of hazardous industries including aviation, medicine and off-shore engineering. His work has led to considerable reductions in errors and their negative consequences and thereby to marked improvements in industrial safety. Analysing many contrasting disasters in the 1970s, 1980s, and 1990s, Reason identified three shared characteristics:

- (1) Contributory factors which were present within the system prior to the occurrence of the disaster. All complex systems contain these 'resident pathogens'.
- (2) Numerous defences, checks, and safeguards which were already in place within the system. These were designed to prevent known hazards from damaging people or assets.
- (3) An unanticipated concatenation of human unsafe acts and local triggers, which defeated the numerous defences, creating a trajectory of opportunity for accidents to occur.

This analysis led Reason to create his most famous contribution to the field of error research: the Swiss Cheese model (Reason, 1990, discussed at length in Reason, 2008). In this model, which we have adapted to our context in [Figure 2](#), the system defences that an organization or community puts in place are represented as slices of cheese. In an ideal world these defensive layers would be intact. In reality, however, they resemble Emmental cheese in having numerous

‘holes’. In contrast to holes in Swiss cheese however, ‘holes’ in systems are continually in flux, opening, closing, and moving around. The existence of holes in any particular defensive layer is not usually a problem. It is only when holes in successive layers align that a pathway of opportunity for disaster is created.

**Figure 2.** *Swiss Cheese Model (adapted from Reason, 1990).*

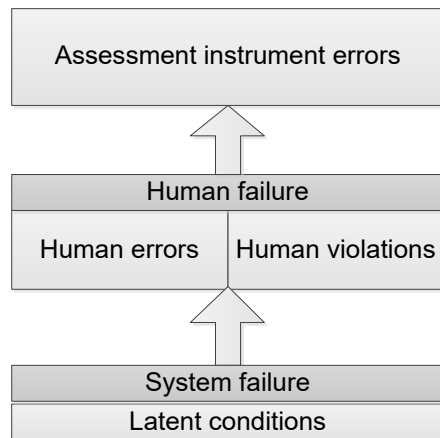


It is crucial to stress that in *any* of the components within an industrial system, there is the potential for practices to occur which engender holes in system defences. It is usually relatively easy to identify the concrete stages or activities within a production process, and then to focus investigations on the human errors (unintended) and procedural violations (intended) that could occur within them. These two kinds of human failure give rise to holes in the defensive ‘Swiss cheese’ layers that open and shut only briefly; their effects are short-lived.

Broadly speaking, within large assessment organizations, the production of examination papers (and other assessment instruments) entails an initial construction phase in which questions are drafted. This is followed by an iterative and often lengthy phase of reviewing and editing, in which questions are checked, re-checked, refined, and combined into examination papers. This phase may include pre-testing the questions with students, or ‘working’ the items or paper as a student proxy. In the next phase, questions and papers are then modified for students with particular needs or for a different mode of administration (for example, an on-screen version of a paper-based examination may be created). Finally, checks are made at a senior level prior to formal sign-off for printing and distribution to candidates. Human errors and violations of procedure could potentially occur during any of these activities, giving rise to errors in examination papers.

Unfortunately, the components of assessment instrument construction and other complex systems (such as aviation, e.g. Wiegmann & Shappell, 2003) are actually far more wide-ranging than this. Reason (2013) argues that system designers, builders, and managers, and procedure writers, inadvertently create ‘latent conditions’ (also called ‘resident pathogens’) which give rise to much larger and longer-lasting holes in the defensive layers. Latent conditions may lie dormant and undiscovered for years until one day they combine with human failures (errors and violations) and local triggers to create an accident trajectory.

We applied Reason’s work to our context to create a simple model of assessment instrument errors (Figure 3). That is, we adopted his theoretical position that system-level failure engenders human failure, which in turn gives rise to manifested errors (Reason, 2013) such as those that appear in assessment instruments.

**Figure 3.** Model of assessment instrument errors.

It follows that the question of what causes errors in assessment instruments can be addressed at two levels: at a psychological level of explanation, and at a system level of explanation. Battmann and Klumb (1993) and more recently Reason (2013) have explored the occurrence of violations, and there is an even richer psychological literature on when and why different types of human errors occur. Common explanatory psychological phenomena include inattentive blindness (Bruner & Postman, 1949; Aimola Davies et al., 2013), inadequate situational awareness (Endsley, 1995; Wickens, 2008), strong habit intrusions (Reason, 2013), and various limitations to working memory (Baddeley, 2010; Reason, 2013). These phenomena have been applied extensively to explain errors in industries such as aviation (Jones & Endsley, 1996), construction (Akinci, 2014), and medicine (Gawande, 2011; Pronovost & Vohr, 2011). In this paper, however, we focus on the latent working conditions that can contribute to system failure and underpin these human failures. These are often known as the root causes of errors. According to Reason (2013), whilst human failures take specific forms which can be hard to predict, latent conditions can be identified before a negative event takes place. A proactive form of system management is therefore needed, which entails regularly monitoring the system's vital signs.

#### 4. THE SHELLO MODEL OF LATENT WORKING CONDITIONS

The aviation industry has made huge improvements to its safety record by identifying and addressing problems with latent working conditions within its systems. It has accepted for some time that human factors play a critical role in every aviation activity, from flight training to airline management (International Civil Aviation Organization, 1993). In a seminal paper, Edward (1972) argued that four types of interacting resources contribute to aviation accidents: software, hardware, environment, and liveware (people). He suggested that the source of every accident can be categorised as liveware, or as a combination of three major relationships: liveware-software, liveware-hardware, and liveware-environment. Edward's model is known as SHEL. It spawned a field of study described by Wiegmann and Shappell (2003) as the ergonomic perspective on human error because it emphasises human-machine-environment interactions. Over time it has been modified by multiple authors.

Chang and Wang (2010), for example, extended SHEL to become SHELLO. They identified the following factors as significant in accidents:

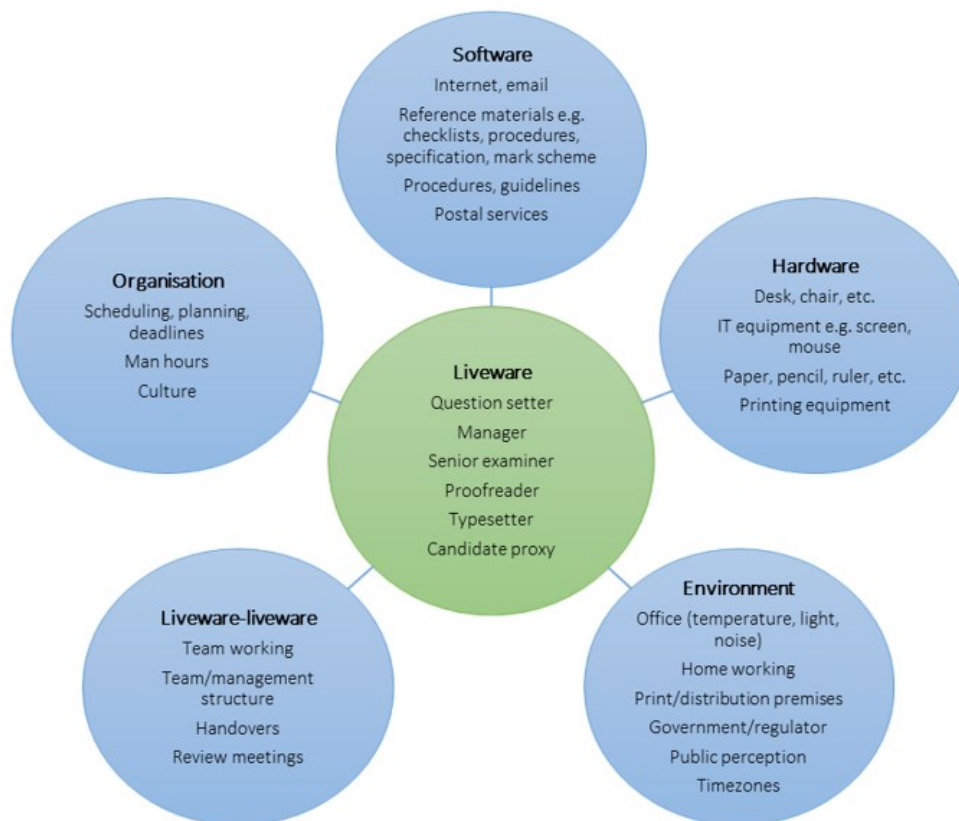
- Software (procedures, manuals, checklists)
- **H**ardware (tools, equipment, physical structure)
- **E**nvironment (physical environment, work patterns, management structures, public perception of industry)

- Liveware (people, managers)
- Liveware-liveware (person-to-person communication)
- Organization (managerial model, decision-making patterns, culture).

Human operators feature in all interactions in this model (liveware to software, liveware to hardware, and so on) and carry risks of committing errors and violations. For those working within the system, the SHELLO model can contribute to an awareness of the context and the need for the factors to dovetail with one another to prevent breakdowns which might result in human errors. To use an example from air traffic control, the cause of an error might be cited as ‘operator fatigue’ which is contained within the liveware category. However, further investigation might show that the organization operated a culture of working long shifts, or that some aspect of the office environment had a part to play. SHELLO has been used successfully to develop numerous risk management strategies, for example, to help airline pilots to reduce runway excursions (Chang et al., 2016).

We used SHELLO as a basis for understanding the latent working conditions affecting educational assessment instrument construction. To do this, we populated its template with relevant factors (Figure 4). Taken as a broad suggestion of the factors which might be involved without being tied to any particular construction process, Figure 4 shows an interactively complex system.

Figure 4. A SHELLO model of the factors affecting assessment instrument construction.



Although the model is simple, it shows the potential impact that a system can have on error and how defences can be breached. Suppose, for example, that while undertaking a new check, a flickering light causes the checker to lose concentration and phone a janitor for help. On hold to the janitor and mindful of time, the checker continues with the check without paying full attention to the task, and makes a slip. This human error results in an error in the examination.

Does the examination error stem ultimately from problems with liveware, the environment, the organization, or all three? It is clear that both resources and working culture play an important role in determining the quality of the examination.

#### 4. TIME PRESSURE

In the SHELLO model, time pressure lies within the ‘organization’ category of latent conditions. Although it would be easy to assume that time pressure always has a negative effect on task accuracy, the issue is more complex than a straightforward speed/accuracy trade-off. Drawing upon regulatory focus theory (Higgins, 1997), Förster et al. (2003) describe two types of goal pursuit among workers: *promotion* focus, and *prevention* focus (Table 1). In proof-reading and similar tasks, colleagues with a promotion focus adopt a risky processing style that is concerned with getting ‘hits’, that is, spotting lots of errors in the text quickly. Colleagues with a prevention focus, on the other hand, adopt a more careful processing style. They are concerned with avoiding making errors in spotting errors in the text. The focus that someone adopts can present at personality or task level, and can be chronic or momentary. For example, Förster et al. (2003) suggest that a promotion focus increases with a colleague’s proximity to goal completion.

**Table 1.** *Types of goal pursuit.*

	Promotion focus	Prevention focus
Concern of colleague	Accomplishments and aspirations Gains and non-gains	Safety and responsibility Losses and non-losses
Behaviour of colleague	Strategic eagerness Risk taking	Strategic vigilance Risk averse
Task speed	High	Low

In one of the experiments described by Förster et al. (2003), two groups of participants were asked to complete a proofreading task as quickly and accurately as possible, identifying errors which had been deliberately created in a passage of text, within a fixed time period. One group was given a promotion focus: they would receive more money for a good speed/accuracy score. The other group was given a prevention focus: they would lose money if they didn’t achieve a high enough score.

The promotion focus was found to enable faster proofreading and the identification of more of the easy-to-spot errors compared with the prevention focus. In contrast, the prevention focus led to higher accuracy in finding more difficult errors than the promotion focus did. Through speed and searching for easy errors, the promotion focus maximised proofreading performance overall, as measured by the total number of errors detected in the fixed time period. Förster et al. (2003) concluded that speed/accuracy trade-offs are a function of both regulatory focus and task difficulty. Whereas easy errors are found quickly with a promotion focus which enhances speed, difficult errors are more accurately found with a prevention focus.

The findings indicate that system designers and managers should think carefully about whether to encourage checkers to adopt a promotion focus or a prevention focus, through remuneration and performance management strategies, for example. They should also use caution when making decisions about how much time is allocated to tasks. Scrutiny of the data on the types of error arising in the assessment instrument construction process may indicate whether a particular focus is likely to yield better outcomes at a particular stage of the process. It may be that both focuses could be used in successive checks of an instrument, to improve the detection of both easy-to-detect and difficult-to-detect errors. It is worth noting that the goal of proof-reading and similar checks is usually to detect *every* error, and a prevention focus seems particularly



appropriate. Given that a prevention focus requires plenty of task time, however, encouraging a prevention focus in combination with rapid working to tight deadlines could be a recipe for disaster.

## **5. WORKLOAD AND STRESS**

Time pressure is usually linked to workload. When both are *too* high, colleagues become stressed. According to Dekker (2002a), tunnelling and regression are both reactions to stressful situations which can result in a loss of situational awareness. Tunnelling describes a fixation on an increasingly narrow portion of one's operating environment (Dekker, 2002a). There are many well-recounted examples, particularly in aviation. For example, a flight crew becomes distracted by an anomaly in the cockpit and fails to notice a loss of altitude. Other examples come from medicine: a doctor becomes so focussed on attempting and re-attempting a difficult surgical procedure that she fails to notice time passing and the patient's condition deteriorating (Syed, 2015). Regression occurs when actors revert to previously learned routines and fail to notice the critical differences of the current situation. Tunnelling and regression could have implications for the assessment instrument construction process if, for example, a question writer has in mind a previous oversight and consequently focuses on that aspect of the task while failing to notice other potential errors.

Dekker (2002a) argued that one way to think about workload and stress is to identify the type of demand-resource mismatch. The problem may not always be time pressure. For example, when working to a tight schedule, the coping resource that a particular colleague requires might be professional skills in workload management. If these skills are insufficiently developed then this could contribute to human errors or violations. For a different colleague working to exactly the same schedule, the problem may be slow computer software. If this resource is inadequate, then the colleague may become stressed and make errors despite excellent workload management skills.

Other related factors shaping the impact of workload can include multiple, competing goals, and not only in terms of the balance across elements such as minimizing costs, maximising accuracy, and adhering to deadlines. Colleagues with different roles within the same team working towards a common, larger goal may not feel pressure from, or responsibility for the same smaller goals set in order to reach the larger one (Dekker, 2002a). There may also be mismatches between team members' knowledge and/or an assumption that others possess the same knowledge, which can result in a lack of coordination.

## **6. THE WIDER ORGANIZATIONAL CULTURE SURROUNDING ERRORS**

Over the past three decades, leaders in industries ranging from transportation and aviation to off-shore energy and nuclear power have identified serious issues with their culture surrounding errors. Their acknowledgements of problems, coupled with concerted efforts to rectify them, have been credited widely for significant reductions in major incidents and for higher safety and quality standards in general. This has not been the case in every workplace, however. The idea of organizational culture can seem so vague and elusive that some senior managers simply pay lip service to it by sending their staff on generic courses, or by checking periodically that they have a formal procedure in place for everything.

In the context of assessment instrument construction, failure to engage with this issue deeply would be a huge oversight. It is worth thinking about what a poor culture surrounding errors looks like in practice, and contrasting it with a good culture. This is what Matthew Syed has done for other industries (Syed, 2015). The principles that he draws together can be applied to assessment instrument construction. The essence of Syed's argument is that a good culture is



one in which we readily acknowledge errors and take a highly systematic and scientific approach to understanding and learning from them.

## 7. GOOD DATA COLLECTION

First, the systematic collection of good data on errors and violations is at the heart of a good organizational culture. It is important to understand what types of errors and violations are occurring, and where in the system they arise. In order to spot recurring or ‘signature’ errors (those with a subtle pattern) the data needs to be detailed and comprehensive. Syed (2015) offers criminal justice as an example of a system in which detailed error data is *not* usually collected, although wrongful convictions have long been established unequivocally (Borchard, 2013). As in many other countries, the jury system in England and Wales is a secretive one. As jurors’ activities (and errors) cannot be scrutinised, there is no opportunity to learn from them.

## 8. COGNITIVE DISSONANCE

Moreover, Syed (2015) argues that rather than admitting to failure and using it as a learning opportunity, members of many police forces and prosecution services experience ‘cognitive dissonance’. This is the inner tension we feel when our beliefs are contradicted by evidence (Festinger, 1957). We do not like to perceive ourselves as irrational or foolish, or to have wasted a lot of time pursuing a cause in vain, as it threatens our self-esteem. Rather than accept that our original judgements were faulty, denial is a more comfortable option psychologically. It is much easier to reframe the evidence, spin it, filter it, or ignore it altogether. If anything we tend to become even more entrenched in our beliefs.

There are many examples of police officers and legal prosecutors refusing to accept DNA exonerations in cases they worked hard on (Innocence Project, 2021). Another example of cognitive dissonance is of people standing by decisions to abandon their families and possessions to join cults, even after prophecies of the world’s end have turned out to be wrong. A typical post-prophecy argument would be that the cult leader and his/her followers are praying so hard and behaving so well that God has shown mercy, which is an even stronger reason to stick with the cult.

Cognitive dissonance could play a part in assessment instrument construction when managers (who may be conscientious workers with considerable expertise) face the discomfort of an error passing undetected through a carefully designed system. They may try to resolve conflicting beliefs by placing the blame elsewhere (e.g., a checker failed in their role) or they may argue that despite everyone’s best intentions, such failures cannot be avoided. According to Dekker (2002b), when failure results in cognitive dissonance, it is usually easiest to place the blame on an individual. He suggests: ‘Faced with a bad, surprising event, people seem more willing to change the individuals in the event, along with their reputations, rather than amend their basic beliefs about the system that made the event possible’ (Dekker, 2002b). Reverting to the ‘bad apple’ theory provides (unfounded) reassurance that the system is essentially safe and errors arise from unpredictable humans working within the system.

## 9. AVOIDING A BLAME CULTURE

According to Reason (2013) when an accident occurs, the key question is not who blundered, but how and why the system defences failed. Enquiries into mishaps frequently reveal errors and violations committed by those at the coalface. At this point it is easy for senior managers to conclude that ‘to err is simply human’ and that all processes worked as they were designed to. As discussed previously, however, the crucial next step is then to investigate the workplace factors (latent conditions) contributing to these errors and violations. These will be factors in the SHELLO model, such as work pressure, inadequate training or briefing, under-staffing,

inappropriate tools and equipment, and so on. These provocative factors are probably the consequence of decisions made by senior management. Reason (2013) points out that such decisions may turn out to be mistaken, but not necessarily so. Almost all high-level decisions simultaneously have positive consequences for some colleagues and negative consequences for others elsewhere in the system. It is rarely possible to please/help everyone all of the time.

As causal factors in the workplace are systemic in nature, blame at the individual level is unhelpful. Some form of ‘no blame’ culture in which colleagues feel able to report every failure, however big or small, is crucial to obtaining comprehensive error data for analysis and system improvement. However, Dekker (2017) argues that many organizations today have a retributive just culture instead. This approach asks: which rule is broken? Who did it? How bad was the breach and what should the consequences be? Who gets to decide this? Where staff are penalised for every error through performance management strategies, remuneration or public shaming, they will be unwilling to own up to slips, mistakes and violations. They will be more likely to hide them, risking problems further down the line. Moreover, at the level of process development, the managers responsible for them will be prone to cognitive dissonance. That is, they will find endless justifications and work-arounds for the decisions that they have made and the procedures that they have implemented. According to Dekker (2017), retributive justice rarely promotes honesty, openness, learning and prevention. Instead, he argues for a culture of restorative justice. This approach asks: Who is affected? What do they need? Whose obligation is it to meet that need? How do you involve the community in this conversation? Edmundson (1999) describes the ‘psychological safety’ that is needed for restorative justice to take hold within teams.

## **10. A SCIENTIFIC APPROACH TO UTILISING ERRORS**

In contrast to the criminal justice system, the aviation industry has a very positive culture surrounding error. Syed (2015) termed this ‘black box’ thinking, after the indestructible box with which every aircraft is equipped. During a flight the box records all instructions sent to the on-board electronic systems, as well as the conversations and sounds in the cockpit. When an accident occurs, the data in the box is analysed and the causes of the accident are identified. Rather than concealing, ignoring or stigmatising failure, aviation culture treats every incident as a data rich learning opportunity. Independent investigators are given *carte blanche* to interrogate all the data. Since any information provided by interested parties is inadmissible in court, their openness and full disclosure is probable. Afterwards, the report is made available to the public and airlines are legally obliged to implement the recommendations. As everyone can access the data, everyone can learn from the errors. Procedures can then be improved, to avoid any repeat of the accident.

Learning from failure is also at the heart of the modern scientific method. The philosopher Karl Popper (1963) argued that science advances through its vigilant response to its own errors. Scientific theories make predictions that can always be tested and this is a huge strength. Unlike in astrology or psychoanalysis, hypotheses are made which can be refuted definitively. When this happens, new ones are developed, the field of enquiry progresses and our body of scientific knowledge grows.

As mentioned previously, the gold standard in scientific method, at least in some circumstances, is arguably the RCT. It has revolutionised pharmacology, for example. Without RCTs, there is a risk that closed loop thinking is perpetuated through skewed interpretations of evidence. That is, those who feel they have benefited from a new treatment may be highly vocal about it whilst those who did not benefit may slip under the evaluative radar. This leads to a false perception of efficacy, potentially perpetuating the use of the treatment on very shaky grounds. Although RCTs are widely used to test the efficacy of medical interventions, they have proven equally

successful in other contexts. Syed (2015) cites examples from large-scale manufacturing to British Cycling and the Olympic Team GB, where the performances of both products and people have been optimised. A systematic ‘trial and error’ approach is often taken until success is achieved.

The success of this approach relies upon rapid feedback on the outcome of each trial and a cultural willingness to try again and again, using the feedback to learn about what works and what does not work. To minimize errors in assessment instrument construction, RCTs could in theory be used repeatedly in research using past instruments to establish the relative efficacies of different checks. They could also be used to investigate the skillsets needed to perform particular checks, and aspects of procedure (e.g. time) needed to develop instruments with minimal errors. Such a systematic approach to error would reduce weaknesses within systems which would otherwise persist due a reliance on unjustifiable assumptions that current procedures are optimal.

Although system improvers frequently seek fast one-off elixirs, the ‘slowly but surely’ approach outlined above actually embodies the theory of marginal gains. This is the idea that lots of small improvements add up to a large improvement, so it is worth making each small improvement. Because the search for marginal gains takes time, it should ideally be part of an organization’s usual activity. Given the infrequency of assessment instrument errors, however, it must be recognised that the resource involved in employing RCTs in this search would be huge. The gains to validity that might be achieved by devoting this resource to other areas might actually be larger, and cost-benefit analyses would undoubtedly be needed prior to embarking on this approach. A more realistic and cost-effective approach to minimizing errors might instead be to focus upon the considerable insights that can be obtained through detailed analyses of routinely collected error data, in error logs, for example. Such analyses might ultimately lead to a smaller number of highly targeted RCTs which focus specifically upon the most persistent problems.

Kahneman (2011) stresses that organizations seeking to improve should routinely look for efficiency improvements, and the operative concept is routine. He argues that expertise develops through a growth mindset and continual learning at the organizational level as well as at an individual level. Similarly, Weick et al. (1999) claim that the power of a safe culture lies in instilling an ongoing ‘collective mindfulness’ of the many entities that can compromise a system’s safeguards. Reason (2013) suggests that if there is a phrase that captures the essence of an unsafe culture, it is unwarranted insouciance. His epitaph for a lot of culture-induced organizational accidents would be: ‘There was always something more pressing to do.’ Of course, this approach costs time and money, and ultimately that must be weighed against of costs of serious errors occurring.

## 11. WORKFORCE ATTITUDES AND INTERPERSONAL SKILLS

A good organizational culture relies heavily upon the attitudes and interpersonal skills engendered in the workforce. Working in the fields of aviation and air traffic control, Kontogiannis and Malakis (2009) identified multiple attitudinal factors which play a part in error detection. These include: vigilance and alertness (including being able to ‘make the familiar strange’); suspicion and curiosity; awareness of vulnerability to errors, and awareness of degradation and disengagement, for example through distraction, fatigue or illness. Being able to cope with frustration from errors was also considered important. The authors also identified relevant team factors. These include: assertiveness, for example, feeling able to question decisions of senior colleagues and those senior staff being open to challenges; the abilities to cross-check and monitor others; the ability to adopt multiple perspectives; and strong communication of intent.

In aviation, these factors are developed in staff through Crew Resource Management (CRM) training. CRM evolved in response to evidence that many aviation accidents did not originate from aircraft technical issues or the crew's lack of knowledge, but from the responses of the crew to the situation in which they found themselves. CRM training aims to develop cognitive and social skills in support of technical training (Civil Aviation Authority, 2014). It is a mandatory component of commercial aircrew training in most countries. Such is the success of CRM that the format has been adapted for use by other industries, such as medicine, nuclear power and the offshore oil industry (Bleetman et al., 2012; Flin et al., 2002). The training content varies between industries, but generally includes teamwork, situation awareness, risk assessment, decision making, communication and workload management.

CRM can play a key role in mitigating cognitive dissonance as a cause of errors. For example, suppose a senior manager asks a checker to carry out an additional check of an assessment instrument. The checker believes that the procedure could increase time pressure further down the chain of checks, but experiences cognitive dissonance because he or she simultaneously believes that (i) the proposed check is wrong and shouldn't be carried out, and (ii) that the manager holds seniority over them and is right. Through CRM training, the checker should feel confident enough to voice concerns to the senior colleague. Equally, having also had CRM training, the senior manager should welcome a query without seeing it as a challenge to authority.

## **12. CONCLUSION**

In this paper we have considered how errors arise in different industries, primarily at a system level of explanation, identifying generalisable principles and applying them to our context of educational assessment instrument construction. The literature we have reviewed is just the tip of the iceberg, given that in some industries, whole books and entire journals have been dedicated to some of the issues explored. It is clear that aviation, energy, and medicine take errors extremely seriously, and this is because errors compromise safety as well as quality; they can be literally a matter of life and death. Although the consequences of errors in educational assessment instruments are rarely so overtly catastrophic, they may nonetheless have life-changing consequences for students.

We have argued that since assessment instrument construction is a complex system comprising numerous interacting components, a holistic approach to system improvement is required. Cherry-picking initiatives from other contexts or introducing yet another examination paper check will not work. Within most assessment construction systems it is relatively easy to identify concrete activities and the human errors and violations that can occur when they are carried out. When human failure occurs, it is not good enough to explain it away by suggesting that all procedures worked as intended but that 'to err is simply human'. It is necessary to look deeper. That is, it is crucial to evaluate the latent working conditions that underpin the efficacy of the procedures, making the human failure more or less likely. This is how organizations in other industries successfully improve their performance.

Latent working conditions which can give rise to human failure and ultimately to errors in question papers are created unwittingly by system designers and procedure writers, and by senior management more generally. These conditions are wide-ranging. They relate to software, hardware, the working environment, the people involved, and organizational culture. Potentially affecting all latent conditions, a good organizational culture is one in which individuals are not blamed for their errors. Instead, errors are acknowledged readily by all but are not trivialised. Senior managers take a highly systematic and scientific approach to understanding and learning from them.

Drawing from these conclusions, we recommend five linked principles for best practice in minimizing errors in assessment instruments. First, a culture of restorative justice should be promoted, in which individuals are not blamed or penalised for errors. This approach asks: Who is affected? What do they need? Whose obligation is it to meet that need? How do you involve the community in this conversation? Secondly, coupled with psychological safety, this will make it possible to collect truly comprehensive data on errors, including data on the latent conditions that engender errors. This will in turn make it possible to identify recurrent ‘signature’ errors and their potential causes.

Thirdly, there is a need to instil in all authors and checkers of instruments an ongoing collective mindfulness of the many entities that can compromise the system’s safeguards. It should be part of routine activity to investigate as many errors as possible - ideally all – to gain an understanding of why things went wrong. Fourthly, there is a need to hypothesise potential solutions to problems and test them scientifically. As a starting point, an RCT approach could be used to determine the relative efficacies of different types of checks, using ‘seeded’ errors in past assessment instruments. Fifthly, this approach of learning from errors needed to be embedded into organizational culture at all levels of staff, so that the necessary resource is made available.

Finally, it is worth noting that the implications of the literature we have reviewed can be extended well beyond our stated context of assessment instrument construction. Awarding organizations and others involved in test construction produce numerous other types of document, including syllabuses, procedural manuals, reports for the regulator, legal contracts, research papers, and so on. All of these documents are prone to errors and the consequences can be serious. Arguably, there is nothing to stop anyone in the educational assessment community from adopting the mindset and approach to improvement advocated here in these related contexts.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

### Authorship contribution statement

**Dr Irenka Suto** is a Principal Research Officer. Prior to joining Cambridge Assessment in 2005, she studied at the University of Cambridge and conducted post-doctoral research into financial decision-making processes. She has a long-standing interest in the many human judgments and decisions entailed in educational assessment, as made by students, teachers, examiners and administrators.

**Jo Ireland** is a Research Officer at Cambridge Assessment. Her research focuses mainly on the comparability and validity of assessments, and has included the development and application of tools to analyse the cognitive demand of examination questions.

### ORCID

Irenka Suto  <https://orcid.org/0000-0001-6871-901X>

Jo Ireland  <https://orcid.org/0000-0003-1237-7860>

## 13. REFERENCES

- Achtenhagen, F. (1994, June). Presentation to Third International Conference of Learning at Work, Milan, Italy.
- Akinci, B. (2014). Situational Awareness in Construction and Facility Management. *Frontiers of Engineering Management*, 1(3), 283-289. <https://doi.org/10.15302/J-FEM-2014037>



- Aimola Davies, A., Waterman, S., White, R. & Davies, M. (2013). When you fail to see what you were told to look for: Inattention blindness and task instructions. *Consciousness and Cognition*, 22(1), 221-230. <https://doi.org/10.1016/j.concog.2012.11.015>
- Baddeley, A. (2010). Working memory. *Current Biology*, 20(4), R136-R140.
- Baranowski, R. (2006). Item editing and editorial review. In S. Downing & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 349-357). Lawrence Erlbaum Associates.
- Battmann, W., & Klumb, P. (1993). Behavioural economics and compliance with safety regulations. *Safety Science*, 16, 35-46. <https://www.sciencedirect.com/science/article/pii/092575359390005X>
- BBC (2017, November 21) NZ minister orders probe into 'impossible' maths exam. <https://www.bbc.co.uk/news/blogs-news-from-elsewhere-42065574>
- Bleetman, A., Sanusi, S., Dale, T., & Brace, S. (2012). Human factors and error prevention in emergency medicine. *Emergency Medicine Journal*, 29, 389-393. <http://emj.bmj.com/content/29/5/389.long>
- Borchard, E. M. (2013). *Convicting the innocent and state indemnity for errors of criminal justice*. The Justice Institute. (Original work published 1932)
- Bruner, J. S., & Postman, L. (1949). On the perception of incongruity: a paradigm. *Journal of Personality*, 18(2), 206-23. <https://doi.org/10.1111/j.1467-6494.1949.tb01241.x>
- Chang, Y.-H., & Wang, Y.-C. (2010). Significant human risk factors in aircraft maintenance technicians. *Safety Science*, 48(1), 54-62. <https://doi.org/10.1016/j.ssci.2009.05.004>
- Chang, Y.-H., Yang, H.-H., & Hsiao, Y.-J. (2016). Human risk factors associated with pilots in runway excursions. *Accident Analysis & Prevention*, 94, 227-237. <https://doi.org/10.1016/j.aap.2016.06.007>
- Civil Aviation Authority (2014). Flight-crew human factors handbook. CAA. <http://publicapps.caa.co.uk/docs/33/CAP%20737%20DEC16.pdf>
- Dekker, S. (2002a). *The Field Guide to Human Error Investigations*. Ashgate.
- Dekker, S. (2002b). Reconstructing human contributions to accidents: the new view on error and performance. *Journal of Safety Research*, 33(3), 371-385. [https://doi.org/10.1016/S0022-4375\(02\)00032-4](https://doi.org/10.1016/S0022-4375(02)00032-4)
- Dekker, S. (2017) *Just Culture: Restoring trust and accountability in your organization*. CRC Press.
- Edmundson, A. (1999). Psychological Safety and Learning Behavior in Work Teams. *Administrative Science Quarterly*, 44(2), 350-383.
- Edward, E. (1972). Man and machine: systems for safety. Proceedings of the British Airline Pilots Association Technical Symposium, London.
- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1), 32-64.
- Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford University Press.
- Flin, R., O'Connor, P., & Mearns, K. (2002). Crew resource management: improving team work in high reliability industries. *Team Performance Management: An International Journal*, 8(3/4), 68-78. <https://doi.org/10.1108/13527590210433366>
- Förster, J., Higgins, E. T., & Bianco, A. T. (2003). Speed/accuracy decisions in task performance: Built-in trade-off or separate strategic concerns? *Organizational Behavior and Human Decision Processes*, 90(1), 148-164. <https://www.sciencedirect.com/science/article/pii/S0749597802005095>
- Gawande, A. (2011). *The checklist manifesto: How to get things right*. Metropolitan Books.
- Harrison, A. (2011, June 9) Students hit by more exam errors. *BBC*. <https://www.bbc.co.uk/news/education-13710868>
- Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist*, 52(12), 1280-1300. <https://pdfs.semanticscholar.org/6b64/5e0418ae70e82cc322dd6fbf0647ae2523e4.pdf>



- Innocence Project (2021). *The innocence project*. <https://www.innocenceproject.org>
- International Civil Aviation Organization (1993). *Investigation of Human Factors in Accidents and Incidents. Human Factor Digest No.7*. <https://skybrary.aero/bookshelf/books/2037.pdf>
- Jones, D. G., & Endsley, M. R. (1996). Sources of situation awareness errors in aviation. *Aviation, Space, and Environmental Medicine*, 67(6), 507-512.
- Kahneman, D. (2011). *Thinking, fast and slow*. Penguin Group.
- Kontogiannis, T., & Malakis, S. (2009). A proactive approach to human error detection and identification in aviation and air traffic control. *Safety Science*, 47, 693-706.
- Meredith, R. (2019, May 17) AS-level Economics exam error under investigation in NI. *BBC*. <https://www.bbc.co.uk/news/uk-northern-ireland-48313904>
- Mitleton-Kelly, E. (2003). Ten principles of complexity and enabling infrastructures. In E. Mitleton-Kelly (Ed.), *Complex systems and evolutionary perspectives on organisations: the application of complexity theory to organisations*. Elsevier.
- New Straits Times (2015) *S. Korea exam chief resigns over errors in high-stakes college test*. <https://www.nst.com.my/news/2015/09/s-korea-exam-chief-resignover-errors-high-stakes-college-test>
- Nisbet, I., & Shaw, S. (2020). *Is Assessment Fair?* Sage.
- Oates, T. (2017). A Cambridge Approach to improving education. *Cambridge Assessment*. <http://www.cambridgeassessment.org.uk/Images/cambridge-approach-to-improving-education.pdf>
- Ofqual (2019). GCSE, AS & A level summer report 2018. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/852440/GQ-Summer-Report-2019-MON1100.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/852440/GQ-Summer-Report-2019-MON1100.pdf)
- Popper, K. (1963). *Conjectures and refutations: The growth of scientific knowledge*. Routledge and Kegan Paul.
- Pronovost, P., & Vohr, E. (2011). *Safe Patients, Smart Hospitals: How One Doctor's Checklist Can Help Us Change Health Care from the Inside Out*. Penguin books.
- Reason, J. (1990). *Human error*. Cambridge University Press.
- Reason, J. (2008). *The human contribution*. Ashgate.
- Reason, J. (2013). *A life in error: from little slips to big disasters*. Ashgate.
- Richardson, H. (2017, May 26) GCSE exam error: Board accidentally rewrites Shakespeare. *BBC*. <https://www.bbc.co.uk/news/education-40059967>
- Rhoades, K., & Madaus, G. (2003). *Errors in standardized tests: A systemic problem*. Boston College.
- Rodriguez, M. (2015). Selected-response item development. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 259-273). Routledge.
- Syed, M. (2015). *Black box thinking. Marginal gains and the secrets of high performance*. John Murray.
- Verne, J. (1996). *Journey to the centre of the earth*. Wordsworth Editions Limited. (Original work published 1864.)
- Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (1999). Organising for high reliability: Processes of collective mindfulness. In R. S. Sutton & B. M. Staw (Eds.). *Research In Organizational Behavior*, 21, 23-81.
- Wickens, C. D. (2008). Situation awareness: Review of Mica Endsley's 1995 articles on situation awareness theory and measurement. *Human Factors*, 50(3), 397-403. <https://journals.sagepub.com/doi/pdf/10.1518/001872008X288420>
- Wiegmann, D. A., & Shappell, S. A. (2003). *A Human Error Approach to Aviation Accident Analysis*. Ashgate.

## Examination of Wording Effect of the TIMSS 2015 Mathematical Self-Confidence Scale Through the Bifactor Models

Esra Oyar<sup>1,\*</sup>, Hakan Yavuz Atar<sup>2</sup>

<sup>1</sup>Gazi University, Department of Educational Science, Ankara, Turkey

### ARTICLE HISTORY

Received: Apr. 12, 2020

Revised: Jan. 14, 2021

Accepted: Mar. 16, 2021

### Keywords:

Wording effect,

Method factor,

Mathematical self-esteem,

TIMSS

**Abstract:** The aim of this study is to examine whether or not the positive and negative items in the Mathematical Self-Confidence Scale employed in TIMSS 2015 lead to wording effect. While examining whether the expression effect is present or not, analyzes were conducted both on the general sample and on a separate sample for female and male students. To this end, data of 5724 students from Turkey who participated in TIMSS 2015 were used. Six different measurement models were created in the analysis of data and tested with Confirmatory Factor Analysis. The study revealed that positive items have a higher mean than the negative ones. In addition, it was concluded that the bifactor models fit the data better compared to the traditional DFA model, in which the model where negative items were taken as a separate factor are those that best fit the data. This situation is verified both in the general sample and the subgroups of females and males. In conclusion, it is recommended that the scale items should be created carefully and whether the positive and negative items result in separate factors should be examined.

## 1. INTRODUCTION

Each measurement instrument is created for a specific purpose, under specific conditions and in a way to apply to specific individuals (Erkuş, 2003). Thus, one of the psychometric properties that are sought for in any measurement instrument is the degree to which it serves its purpose, in other words, its validity. Validity is the process of evidence collection with the aim of supporting the inferences to be drawn from the test scores obtained through measurement instruments (Cronbach, 1984). This process involves determining the degree to which the structure intended to be measured is being measured. However, some situations encountered during the measurement threaten validity and lead to errors in the measurement of the intended structure. One of the situations that threaten validity is the method factor (Ford & Scandura, 2018). Method factor occurs when participants systematically respond to the items differently due to the wording of the items in the scale (DiStefano & Motl, 2009). In case the measurement instrument includes method factors such as item characteristic (social desirability, etc.), item content (positive or negative items, etc.) and measurement content (time or place of

---

\*CONTACT: Esra OYAR ✉ [esra.tas18@gmail.com](mailto:esra.tas18@gmail.com) 📍 Gazi University, Department of Educational Science, Ankara, Turkey

measurement, etc.) (Podsakoff et al., 2003), the researcher cannot measure the intended trait due to difference from the real factor in the structure that is intended to be measured, which threatens the validity (Chen, 2017; Yang et al., 2012). If the test has negative and positive items, it causes a method factor due to the item content. This situation is defined as the wording effect in the literature (Gu et al., 2015). It has been suggested in the literature that measuring various structures in social sciences including personality, attitude and anxiety requires the use of positive and negative items evenly (DeVellis, 2003; Weijters et al., 2013), which is argued to decrease the response bias (Weijters et al., 2013). When scale items include negative statements, participants read them more carefully, thereby eliminating responses that have the same response patterns (Podsakoff et al., 2003). The main assumption when using both types is that the negative items will represent the structure in the same way as their positive counterparts (Marsh, 1996). In other words, when the negative items are reverse coded, both item sets should be psychometrically indistinguishable. However, recent studies have revealed that the coexistence of positive and negative items in a scale results in systematic measurement error and thus leads to biased interpretation of results (Gu et al., 2015; Schriesheim et al., 1991). In addition, researchers state that a two-factor structure is produced when mixed items are used (Greenberger et al., 2003; Ibrahim, 2001), which jeopardizes the structure validity (Schmitt and Stuits, 1985; Woods, 2006), and that positive items have a higher mean compared to the negative items (Weems, Onwuegbuzie and Collins, 2006). In the measurement of the structure, wording effect not only poses a threat against the validity but also can decrease the reliability of both the scale items and the scores (Gu et al., 2015; Weems et al., 2003; Yang et al., 2012). Therefore, if the wording effect is modeled through a proper measurement model, researchers can assess the psychometric properties (validity, reliability, etc.) of the data more precisely based on this effect (Gu et al., 2015). When the literature review is examined, considering that the positive and negative items in scale development and adaptation studies may cause difficulties in construct validity, testing this situation has been deemed worthy of research.

Various methods are employed in modeling the wording effect. The most frequently used methods are Confirmatory Factor Analysis (CFA) models and bifactor models (DiStefano and Motl, 2006; Tomas & Oliver, 1999). Confirmatory Factor Analysis (CFA) is a type of Structural Equation Modeling (SEM) and helps to analyze measurement models that allow to establish relationships between the observed variables or indicators (items) that measure the same latent traits or factors (Brown, 2006). Another measurement model employed in the wording effect is the bifactor model (Wang et al., 2018). Bifactor models were developed by Holzinger and emerged as a type of confirmatory factor analysis (Jennrick & Bentler, 2011). In recent years, bifactor models have been increasingly used as an alternative but more advantageous approach in testing the multi-facet structures and in addressing the subject of dimensionality in psychological research (Chen & Zhang, 2018). This model includes one common factor that represents the shared variant in all scale items and an additional group factor that represents the shared variant in the items in a group (Reise, 2012). The common factor represents the individual differences in the target factor which is common to the items and the researcher deals with. Group factor, on the other hand, refers to the shared variant in item responses that cannot be explained by the common factor (Reise et al., 2010). Common factor and group factor are assumed to be orthogonal.

In studies examining the wording effects, (i) a model incorporating only the relevant factor, (ii) bifactor models incorporating positive and negative items as separate factors in addition to the common factor, and (iii) measurement models incorporating the correlation between the error terms of the positive and negative items are created (Chen et al., 2010; Gu et al., 2015; Horan et al., 2003; Marsh, 1996). Bifactor models in which positive and negative items are included as separate factors are also called correlated method (CM) (Lindwall et al., 2012). Similarly, measurement models including the correlation between the error terms of positive and negative

items are defined as correlated uniqueness (CU) (Lindwall et al., 2012). Both models are measurement models which attempt to identify the wording effect of positive and negative items; however, they have some differences. The CM model incorporates certain latent method factors underlying the scale items of the same method (in other words, item formats expressed as positive or negative) along with a latent factor. On the contrary, the CU models are based on establishing a correlation between the remains of positive and negative items (Lindwall et al., 2012; Wu, 2008). Thus, the CM model can be predicted by other factors or variables, but it is not the case in the CU model. Interpretation of method factors is easier and clearer in the CM model than the CU model (Wu, 2008).

In the light of the foregoing, in order to determine whether or not the test items referred to as negative measure a structure other than the intended one, Weems et al. (2006) conducted a study on 153 university students who studied education and psychology. The study revealed that the mean scores the students obtained from the positive items were higher than that from the negative items. In their study, Yang et al. (2012) examined whether or not the positive and negative items in the Attitude Toward Mathematics Learning Scale in TIMSS 2007 had a wording effect on the Taiwan and America sample. The sample of the study consists of the data of 4111 Taiwanese and 7831 American fourth-grade students. A series of CFA showed that there is a wording effect for both samples. Negative items are claimed to have lower reliability and approximately 25% of the score variance in the negative items are told to be caused by the measurement method, not the latent trait. In conclusion, the researchers stated that whether the items had wording effect should be examined and the negative statements should be worded as simple as possible. In another study, which examines the wording effect based on TIMSS scales, Michaelides (2019) performed some analyses by way of an 18-item motivation scale. The scale included three sub-scales. The measurement models that were created are, respectively, (i) one-dimensional model, (ii) three-dimensional model, (iii) second-degree factor model with three sub-scales, (iv) the model in which three dimensions are correlated and negative method factor is included, (v) a model in which the uniqueness variance of negative items are correlated, and (vi) the model in which negative and positive items are included as factors. When the fitting values of the measurement models are considered, the model in which the correlation between negative items was established yielded the best result.

Studies examining the method factor caused by wording are usually carried out on adults (DiStefano & Motl, 2009; Horan et al., 2003; Tomas & Oliver, 1999). When the verbal skills of younger participants are considered, however, this effect might be greater (Yang et al., 2012). Benson and Hocevar (1985) investigated the wording effect in the attitude scales on fourth- to sixth-grade children in the USA by way of the item sets consisting of 15 items. The first item set included only the positive items while the other one included only the negative items. At the end of the study, it was determined that students did not give the same response to the positive and negative items having the same content and were likely to demonstrate a less positive attitude in negative items. Researchers stated that little children cannot express agreement by giving a negative response to a negative statement or disagreement by giving a positive response to a negative statement. Thus, assessing whether using positive and negative items in combination results in the wording effect for younger participants is of importance (Yang et al., 2012). Based on these studies, it is important to investigate whether a separate latent structure is formed in the inclusion of negative statements in the analysis by reverse coding, especially in young age groups, to reveal the structure correctly. Another point examined in terms of wording effect is whether it differentiates depending on gender (DiStefano & Motl, 2009; Michaelides et al., 2016). Studies also tested measurement invariance by taking gender variable as a subgroup. However, this study did not attempt to determine whether the measurement model accepted based on general sample is similar for both female and male but rather to find out which measurement model fits the data better for both females and males.

## 1.2. Purpose

The aim of this study is to determine whether or not the responses of eighth-grade students to the scale items consisting of both positive and negative items in the Mathematical Self-Confidence Scale conducted in TIMSS 2015 have a wording effect by means of Confirmatory Factor Analysis based on the bifactor models that have been created. To this end, the presence of the wording effect will be investigated not only on the general sample but also on separate samples created both for male and female students by way of creating different measurement models (Models 1-6).

## 1.3. Research Questions

This study includes attempts to address the following research questions:

1. Is there a significant difference between the scores the students got from between the mean scores the students got from the positive and negative items in the Mathematical Self-Confidence Scale?
2. Do the positive and negative items in the Mathematical Self-Confidence Scale result in a wording effect?
  - a. Is there a wording effect in the general sample?
  - b. Is there a wording effect for female students?
  - c. Is there a wording effect for male students?

## 2. METHOD

### 2.1. Research Design

The purpose of this study is to investigate whether there is a method/wording effect on the items in TIMSS 2015 Mathematical Self-Confidence Scale by way of CFA models and bifactor models. This is a descriptive study in that it aims to put forward the current situation (Büyüköztürk et al., 2017).

### 2.2. Study Group

In this study, students who participated in the 2015 TIMSS exam from Turkey constitute the working group. Among these students, data of 5724 8th grade students who responded to all items in the "Confidence in Mathematics" scale were used. 48.5% (2779 people) of these students are female students and 51.5% (2945 people) are male students.

### 2.3. Data Collection Tool

The measurement tool used in this study is the Scale of Self-Confidence in Mathematics, which was developed in a different language and adapted to Turkish (Table A1). Within the scope of the study, the effects of positive and negative items on the construct validity of the scale were examined. In the analyzes, it was tried to determine whether a separate structure was formed in the case of positive or negative matter. For this reason, it is thought that cultural effect from a scale obtained by adaptation study will not make a difference in the response pattern to positive and negative items.

There are a total of 9 items in the Mathematical Self-Confidence Scale administered in TIMSS 2015, which was designed to determine the self-confidence degree of students in the Mathematics class, these items consist of four positive and five negative items. Items and information related to them are available in ANNEX1.

Translation of the items in the scale has been obtained from the TIMSS 2011 final report. Students' responses to these items are evaluated on a 4-point Likert scale of 1) Completely agree, 2) Partially agree, 3) Partially disagree, 4) Completely disagree. In the analysis phase, positive items were reverse scored and the total score was calculated based on 9 items in the scale.



## 2.4. Data Analysis Procedures

In order to seek an answer to the first research question of this study, “Is there a significant difference between the scores the students got from the means of the positive and negative items in the Mathematical Self-Confidence Scale?” paired sample t-test was performed based on the mean scores of students for the positive and negative items (Kirk, 2007). Since the number of the positive and negative items in the scale is different, in order to ensure that both total scores will be in the same range, total scores of students for positive and negative items were divided by the total number of items in the relevant score. Cohen's d was used to calculate the effect size.

$$d = \frac{t}{\sqrt{N}}$$

Following the standard of Cohen (1988), effect size estimates of 0.2, 0.5 and 0.8 were considered as small, medium and large, respectively. In the study, six different measurement models were created in order to address the second research question and tested through confirmatory factor analysis. These models are as follows:

*1. Model:* Single-factor model for the Mathematical Self- Confidence variable.

*2. Model:* Bifactor model composed of both Mathematical Self- Confidence factor and positive and negative items.

*3. Model:* Bifactor model composed of both Mathematical Self- Confidence factor and positive items.

*4. Model:* Bifactor model composed of both Mathematical Self- Confidence factor and negative items.

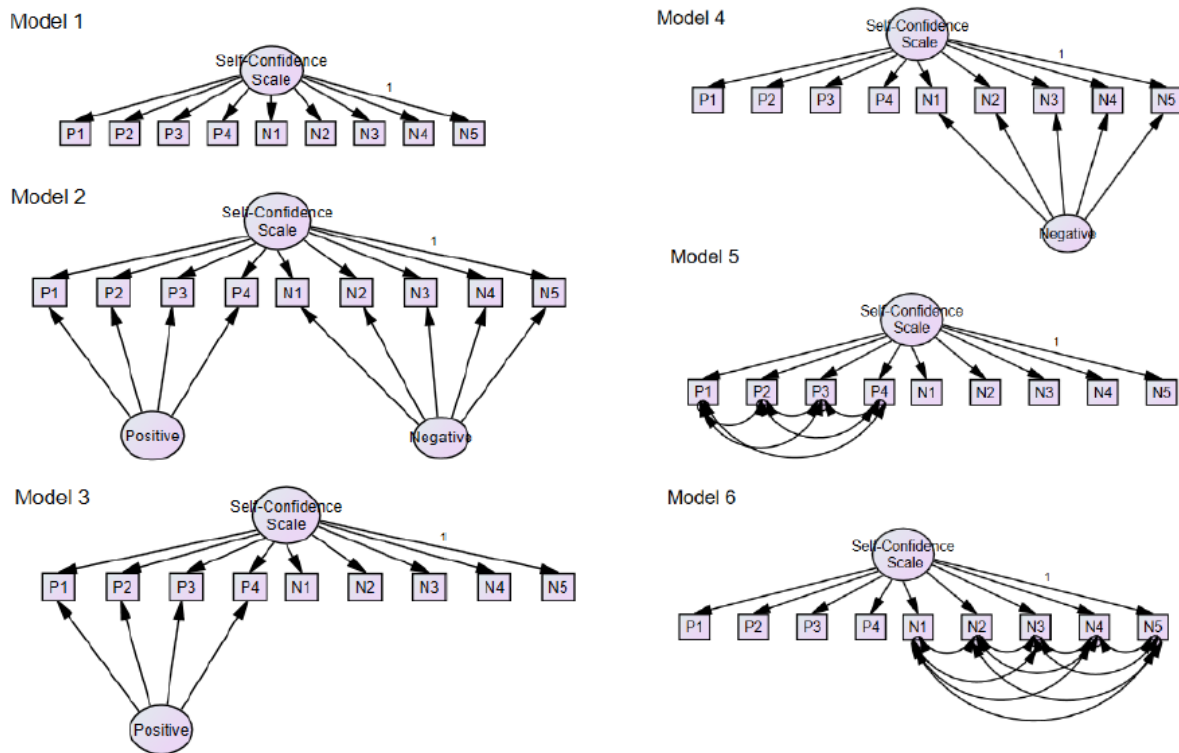
*5. Model:* A Mathematical Self- Confidence factor including correlated uniquenesses among positively worded items

*6. Model:* A Mathematical Self- Confidence factor including correlated uniquenesses among negatively worded items

The figural representations of the models are presented in [Figure 1](#). The purpose of Model 1 is to create a measurement model for a single latent factor (Mathematical Self- Confidence). The measurement model was created assuming that all items in the scale fall under a single latent factor and their model fit indices were examined. In Model 2, positive and negative items are collected under a separate latent factor for each in addition to the Mathematical Self-Confidence latent factor and the bifactor model was created. Model 3 and Model 4 differ from Model 2 in that the bifactor model was created with the assumption that only the positive items and only the negative items fall under a latent factor, respectively, in addition to the Mathematical Self- Confidence latent factor. In Model 5 and Model 6, a correlation was established between latent variances of positive and negative items, respectively, and the model contained a single latent factor (Mathematical Self- Confidence). Goodness of fit indices obtained from all models was examined and the model that fits the data best was accepted. This process was carried out not only on the general sample but also on the sub-samples containing only females or only males, and efforts were exerted to find out which measurement model fits the data in the relevant sample. At this point, the primary aim is to determine which one of the measurement models created displays the best fit in each of the three data sets.

Model 2, Model 3 and Model 4 are CM models, whereas Model 5 and Model 6 are CU models. CM models incorporate positive and negative items as a distinct latent factor in addition to the common latent factor. CU models, on the other hand, create a measurement model by correlating residual variances (uniqueness variances) rather than gathering negative items under a latent factor for each.



**Figure 1.** Model Representations.

### 2.4.1. Assessment criteria

In the evaluation of measurement models, the values of  $\chi^2$ , RMSEA, SRMR, CFI and TLI in the MPlus package program output were examined.

- RMSEA value smaller than 0.08, CFI and TLI values greater than 0.95 and SRMR value smaller than 0.06 indicate that the data and the model represent a perfect fit, whereas RMSEA value smaller than 0.10, CFI and TLI values greater than 0.90 and SRMR value smaller than 0.08 indicate that the data and the model represent an adequate fit (Hu & Bentler, 1999).
- Low RMSEA and SRMR but high CFI and TLI values in a measurement model are interpreted as the model fits the data better than other models.

### 2.4.2. Testing of the assumptions

In data analysis, the first missing data, extreme value and normality assumption checks were carried out. Since the missing data did not exceed 5%, students with missing data were excluded from the study. Information regarding the sample on which the analyses were performed is shown in Table 1.

Following the deletion of the missing data, the sample included data from 5724 students, 2779 (48.5%) females and 2945 (51.5%) males. Examination of z scores for the extreme value revealed that there is no student score out of  $\pm 3$  range, so there is no extreme value in the data. Finally, skewness and kurtosis values were checked for normality assumption. The skewness and kurtosis values for total scores and scores obtained from positive and negative items are in the range of  $\pm 1$ . Thus, considering the sample size and skewness and kurtosis values, it can be said that the data has a normal distribution (Büyükoztürk, 2012).

**Table 1.** Descriptive Statistics for Sampling.

		f	%
Gender	Female	2779	48.5%
	Male	2945	51.5%
Total		5724	100%

### 3. RESULT / FINDINGS

Mean and standard deviation values were calculated not only for the entire sample but also for both subgroups of female students and male students for each item in the scale. Calculated values for items are as shown in Table 2.

**Table 2.** Statistics for Items.

Items	General		Female students		Male students	
	$\mu$	SD	$\mu$	SD	$\mu$	SD
M1	2.96	.943	2.96	.954	2.97	.932
M2*	2.46	1.101	2.47	1.125	2.44	1.077
M3*	2.52	1.156	2.52	1.173	2.53	1.139
M4	2.79	.967	2.77	.953	2.81	.979
M5*	2.60	1.130	2.59	1.150	2.61	1.111
M6	2.37	1.042	2.27	1.024	2.47	1.050
M7	2.62	1.055	2.61	1.056	2.64	1.055
M8*	2.17	1.149	2.16	1.163	2.18	1.136
M9*	2.32	1.143	2.33	1.153	2.31	1.132

\*negative items

Examination of the values in the table reveals that item means obtained from the entire sample and the means of female and male students are close. In the scale, item 8, “Mathematics is harder for me than any other subject” has the lowest mean, while item 1 “I usually do well in mathematics” has the highest. Means obtained from the positive items are higher than the means obtained from the negative items both in the general sample and in the subgroups of females and males.

Paired sample t-test was employed to find out whether there is a significant difference between the mean scores students got from the positive and negative items in the scale, the results of which are shown in Table 3.

**Table 3.** Paired Sample T-Test Results for Positive and Negative Items.

		$\mu$	SD	t	p
Items	Positive	2.69	.850	23.92	.000*
	Negative	2.41	.891		

\* $p < 0.05$

When table values are examined, it is seen that the scores students got from positive ( $\mu = 2.69$ ) and negative items ( $\mu = 2.41$ ) differentiate significantly and this difference is in favor of the positive items ( $t = 23.92, p < 0.01$ ). In other words, students got higher scores from the positive items compared to the negative items. Cohen's d was calculated with the values obtained from the t-test result ( $d = 0.32$ ). It is seen that the value obtained from the analysis results has a

medium size effect. In the study, six different measurement model were created for the second research question. Goodness of fit indices obtained from the analyses are presented in [Table 4](#).

**Table 4.** *Goodness of Fit Index Results for General Sample.*

	df	$\chi^2$	RMSEA	CFI	TLI	SRMR
Model 1	27	4931.74	0.178	0.73	0.63	0.098
Model 2	21	2049.29	0.130	0.89	0.81	0.371
Model 3	25	2261.62	0.125	0.87	0.82	0.372
Model 4	24	1242.16	0.094	0.93	0.90	0.168
Model 5	21	631.97	0.071	0.96	0.94	0.036
Model 6	17	168.13	0.039	0.99	0.98	0.012

It seems that the data do not fit the single-factor structure (Model 1) for this model ( $\chi^2 = 4931.74$ ; RMSEA= 0.178; CFI= 0.73; TLI=0.63; SRMR=0.098). The results of the bifactor model, the model which was created second, fit the data better than the previous model ( $\chi^2 = 2049.29$ ; RMSEA= 0.130; CFI= 0.89; TLI=0.81; SRMR=0.371). However, the obtained values are not in the desired range for perfect fit. For Model 3, examination of the results revealed that the data fit the model better than the other models ( $\chi^2 = 2261.62$ ; RMSEA= 0.125; CFI= 0.87; TLI=0.82; SRMR=0.372). Model 4 has proven to fit the data best compared to previous models. ( $\chi^2 = 1242.16$ ; RMSEA= 0.094; CFI= 0.93; TLI=0.90; SRMR=0.168). Among Model 5 and Model 6, the model in which a correlation was established between the error terms of negative items (Model 6) showed the best fit ( $\chi^2 = 168.13$ ; RMSEA= 0.039; CFI= 0.99; TLI=0.98; SRMR=0.012).

Finally, considering the fit indices, the models that fit the values best were found to be Model 4 and Model 6. In both models, negative items were included in the measurement model. By adding negative items to the model, it can be said that negative items cause a wording effect as a result of obtaining the most suitable model for the data. [Table 5](#) shows standardized factor loading values for each items obtained from the created models.

**Table 5.** *Standardized Factor Loading Values for General Sample.*

Items	Model 1	Model 2			Model 3		Model 4		Model 5	Model 6
		SC	PI	NI	SC	PI	SC	NI	SC	SC
y1	.74	.66	.66		.66	.66	.89		.53	.82
y4	.71	.62	.63		.63	.63	.84		.48	.80
y6	.70	.63	.60		.63	.61	.82		.48	.78
y7	.69	.61	.61		.61	.61	.81		.46	.77
y2	.60	.70		.85	.74		.35	.75	.69	.40
y3	.73	.77		.13	.81		.58	.58	.77	.56
y5	.42	.59		.08	.61		.27	.60	.55	.22
y8	.67	.84		.02	.83		.48	.66	.79	.46
y9	.69	.87		-.03	.84		.52	.64	.80	.50

When the table values were examined, standardized factor loading values for Model 1 were predicted to be between .42 and .74. In Model 2, loading values for common factor were predicted to be between .59 and .87., and for negative and positive factors in Model 2, the factor loading values were predicted to be between .60 and .66 and between -.03 and .85, respectively. In Model 3 the factor loading values were predicted to be between .61 and .84 for general factor,

and between .61 and .66 for positive items. Loading values under common factor were predicted to be between .27 and .89 for Model 4, and factor loading values were predicted to be between .58 and .75 for negative items. In model 5 and 6, loading values for common factor were predicted to be between .46 and .80, .22 and .82, respectively.

As a result, according to the standardized factor load values, the results obtained from all models except Model 2 are at acceptable values. However, considering the fit indices, it can be said that the most suitable model for the data is Model 4 and Model 6. Among the measurement models, results of goodness of fit indices for the group of female and male students are presented in Table 6.

**Table 6.** Goodness of Fit Index Results for Female and Male Students.

Goodness of Fit Index Results for Female Students						
	df	$\chi^2$	RMSEA	CFI	TLI	SRMR
Model 1	27	1710.29	0.150	0.83	0.77	0.070
Model 2	21	1173.63	0.141	0.88	0.80	0.340
Model 3	25	1292.21	0.135	0.87	0.82	0.337
Model 4	24	775.94	0.106	0.92	0.89	0.166
Model 5	21	405.64	0.081	0.96	0.93	0.034
Model 6	17	113.94	0.045	0.99	0.98	0.012

Goodness of Fit Index Results for Male Students						
	df	$\chi^2$	RMSEA	CFI	TLI	SRMR
Model 1	27	3721.12	0.216	0.55	0.41	0.132
Model 2	21	1034.61	0.128	0.88	0.79	0.397
Model 3	25	1135.38	0.123	0.87	0.81	0.403
Model 4	24	558.10	0.087	0.94	0.90	0.173
Model 5	21	297.65	0.067	0.97	0.94	0.042
Model 6	17	64.45	0.031	0.99	0.99	0.012

Examining the table values for female students, the data does not seem to fit the single-factor structure model ( $\chi^2 = 1710.29$ ; RMSEA= 0.150; CFI= 0.83; TLI=0.77; SRMR=0.070). In the second model, the model fits the data better ( $\chi^2 = 1173.63$ ; RMSEA= 0.141; CFI= 0.88; TLI=0.80; SRMR=0.320). However, obtained values are not in the desired range for perfect fit. As the third model, examination of the results revealed that the data fit the model better than the other models ( $\chi^2 = 1292.21$ ; RMSEA= 0.135; CFI= 0.87; TLI=0.82; SRMR=0.316). In the next model, only negative items are included in the model as a factor and it is determined that it is the model that best fits the data compared to the previous models ( $\chi^2 = 775.94$ ; RMSEA= 0.106; CFI= 0.92; TLI=0.89; SRMR=0.161). Finally, between Model 5 and Model 6, the model in which a correlation was established among the error terms of negative items (Model 6) showed the best fit ( $\chi^2 = 113.94$ ; RMSEA= 0.045; CFI= 0.99; TLI=0.98; SRMR=0.012). Finally, considering the fit indices, the models that fit the values best were found to be Model 4 and Model 6. In both models, negative items were included in the measurement model. In this case, negative items in the scale items cause a wording effect in the subgroup consisting of female students.

Examining the table values for male students, the data does not seem to fit the single-factor structure model ( $\chi^2 = 3721.11$ ; RMSEA= 0.216; CFI= 0.55; TLI=0.41; SRMR=0.132). In the second model, the model fits the data better ( $\chi^2 = 1034.61$ ; RMSEA= 0.128; CFI= 0.88; TLI=0.79; SRMR=0.372). Fit indices are not in the acceptable range for both models. As the

results of the third model revealed ,the data fit the model better than the other models ( $\chi^2 = 1135.39$ ; RMSEA= 0.123; CFI= 0.87; TLI=0.81; SRMR=0.379). In the next model, it is determined that it is the model that best fits the data compared to the previous models ( $\chi^2 = 558.10$ ; RMSEA= 0.087; CFI= 0.94; TLI=0.90; SRMR=0.163). Finally, between Model 5 and Model 6, the model in which a correlation was established among the error terms of negative items (Model 6) showed the best fit ( $\chi^2 = 64.45$ ; RMSEA= 0.031; CFI= 0.99; TLI=0.99; SRMR=0.012). As a result, it is seen that Model 4 and Model 6 are the measurement models that show best fit, similar to the result obtained for the general sample and the subgroup of female students. In both models, negative items were included in the measurement model. In this case, negative items in the scale items cause a wording effect on the subgroup of male students. Table 7 shows standardized factor loading values for each items obtained from the measurement models created for female and male students.

**Table 7.** Standardized Factor Loading Values for Female and Male Students.

Standardized Factor Loading Values for Female Students										
Items	Model 1	Model 2			Model 3		Model 4		Model 5	Model 6
		SC	PI	NI	SC	PI	SC	NI	SC	SC
y1	0.78	0.66	0.66		0.67	0.67	0.89		0.63	0.84
y4	0.74	0.67	0.58		0.67	0.58	0.84		0.59	0.81
y6	0.73	0.66	0.56		0.66	0.55	0.81		0.58	0.78
y7	0.71	0.64	0.59		0.64	0.58	0.81		0.56	0.77
y2	0.65	0.75		0.83	0.75		0.44	0.71	0.71	0.51
y3	0.77	0.78		0.09	0.81		0.69	0.46	0.78	0.67
y5	0.48	0.60		0.03	0.61		0.35	0.53	0.56	0.34
y8	0.70	0.85		-0.02	0.84		0.55	0.63	0.80	0.55
y9	0.73	0.88		-0.07	0.85		0.60	0.59	0.81	0.59

Standardized Factor Loading Values for Male Students										
Items	Model 1	Model 2			Model 3		Model 4		Model 5	Model 6
		SC	PI	NI	SC	PI	SC	NI	SC	SC
y1	0.75	0.66	0.66		0.66	0.66	0.88		0.42	0.80
y4	0.73	0.60	0.66		0.60	0.66	0.84		0.37	0.80
y6	0.74	0.61	0.64		0.61	0.64	0.83		0.39	0.79
y7	0.71	0.59	0.64		0.59	0.63	0.80		0.37	0.76
y2	0.47	0.61		0.88	0.73		0.26	0.77	0.68	0.29
y3	0.62	0.77		0.18	0.82		0.47	0.66	0.77	0.45
y5	0.30	0.57		0.14	0.60		0.12	0.63	0.54	0.11
y8	0.57	0.82		0.08	0.82		0.40	0.69	0.77	0.38
y9	0.59	0.86		0.03	0.83		0.44	0.68	0.78	0.41

When the table values were examined for female students, standardized factor loading values for Model 1 were predicted to be between .48 and .78. In Model 2, loading values for common factor were predicted to be between .60 and .88. In Model 2, for positive and negative factors, the factor loading values were predicted to be between .56 and .66 and between -.07 and .83, respectively. In Model 3, loading values under common factor were predicted to be between .61 and .85, and between .55 and .67 for positive items. Loading values under common factor were predicted to be between .35 and .89 for Model 4, and factor loading values were predicted

to be between .46 and .71 for negative items. In Model 5 and Model 6, where correlations between errors were included in the model, factor loading values were predicted to be between .56 and .81 and between .34 and .84, respectively.

When the table values were examined for male students, standardized factor loading values for Model 1 were predicted to be between .30 and .75. In Model 2, loading values for common factor were predicted to be between .57 and .86. When positive and negative items were taken as factor, the factor loading values were predicted to be between .64 and .66 and between .03 and .88, respectively. Loading values under common factor were predicted to be between .59 and .83 for Model 3, and between .63 and .66 for positive items. Finally, loading values under common factor were predicted to be between .12 and .88 for Model 4, and factor loading values were predicted to be between .63 and .77 for negative items. In Model 5 and Model 6, where correlations between error terms were included in the model, standardized factor loading values were predicted to be between .37 and .78 and between .11 and .88, respectively.

As a result, for both samples, according to the standardized factor loading values, the results obtained from all models except Model 2 are within acceptable range. However, when evaluated together with the fit indices, it can be said that the most suitable model for the data is Model 4 and Model 6.

#### **4. DISCUSSION and CONCLUSION**

The aim of this study is to examine whether or not the positive and negative items in the Mathematical Self-Confidence Scale employed in TIMSS 2015 cause a wording effect. For this purpose, in addition to the general sample, subgroups of male and female students were examined separately. Based on the study, it was determined that there was a significant difference between the scores students got from the positive items and the scores they got from the negative items. The mean of the students from the negative items is lower than the mean they got from the positive items. Second, it was determined that the measurement models that best fit the data were the models incorporating the method factor for negative items (Model 4 and Model 6). Although negative items are considered as a separate factor in both models, Model 6 gives better results than Model 4. This may be due to the fact that CU models that allow residuals to be correlated consider not only the variance associated with the wording effect, but also unknown factors (Wu et al., 2017). However, Model 4 can be accepted as the measurement model for the relevant scale since it is easy to interpret (Wu, 2008). In conclusion, negative items for both the general sample and the groups of female and male students in this study cause a method factor in the respondents. The method factor generally represents the “nuance” variance that is not desired in the observed output related to the way the information is collected, rather than the variance intended to be measured (Maul, 2013).

In this study, it is seen that the mean of positive items is higher than the average of negative items because students do not agree more with negative items than positive items. In other words, while the students did not give negative responses such as "I partially disagree" or "I completely disagree" to the negative items; they give positive responses to positive items such as "I partially agree" or "I completely agree". One reason students prefer to respond less to negative items may be "social desire". Social desirability refers to the tendency of the participants to give socially desired answers instead of choosing answers that reflect their true emotions (Grimm, 2010). For example, "I usually do well in mathematics" is the item with the highest average ( $\mu = 2.96$ ) and most of the students answered this item as "I partially agree" or "Strongly agree". However, the item with the lowest average in the scale is "Mathematics is harder for me than any other subject" ( $\mu = 2.17$ ). The students agreed with this item at a moderate level compared to the previous sample item. There are studies in the literature on the data obtained from TIMSS conducted in different years. Similarly, Marsh (1986) found that younger



students and students with poor reading skills could not respond appropriately to the negative items in the rating scales. As a result, it can be said that expressing negative items requires special attention, especially for students in the younger age group, and scale items should be formed with simpler expressions rather than a long and complex structure.

There are similar studies on TIMSS scales, in which positive and negative items cause a wording effect (Hooper et al., 2013; Wang et al., 2018). In their analysis on the Mathematical Self-Confidence Scale administered in TIMSS 2011, Hooper et al. (2013) put forth that there are differences in terms of psychometric properties between positive and negative items. Confirmatory Factor Analysis was adopted in this study for analysis. In their study, they stated that the model fit indices recorded a remarkable increase when correlations were established between the error terms of negative items in both fourth-grade data and eighth-grade data, which can be argued to cause a wording effect for the negative items in the scale. In another study carried out on the same scale, Wang et al., (2018) investigated the presence of wording effect through multi-level models in which students were divided into classes. As a result of this study, it was determined that there are both intra-level and inter-level wording effects in scale items. The results of both studies are similar to this study. In this study, bifactor models and the Mathematical Self-Confidence Scale administered in TIMSS 2015 were examined and it was determined that negative items caused a wording effect. Recent studies show that bifactor models are frequently used in determining the wording effect (Hyland et al., 2014; Wang et al., 2015).

Another finding obtained from the study is that the same measurement model was used both for the general sample and for the groups created only for female students or only for male students. In each of the three samples, the best result was obtained when the correlations between error terms of negative items were included in the model. Similar findings have been found in the literature (DiStefano & Motl, 2009; McLarty et al., 1989). In their study, DiStefano and Motl (2009) examined whether the wording effect differs by gender based on the Rosenberg Self-Esteem Scale (RSES) items. The study showed that there is a method factor for the items worded negatively in the RSES scale for both men and women, but this effect does not differ by gender.

As a result, this study examined whether the positive/negative items in the scale items cause the method factor, and whether the structure contains a method factor for female students and male students as well as for the general sample. This study can also be carried out with the data of English-speaking or non-English-speaking students or students in different countries speaking different languages. Similarly, it can be determined whether the scale items cause a wording effect based on different age groups. In addition, in the scale development process, negative items can be included by considering the group to which the scale will be administered. Similarly, if a scale is to be adapted, it can be examined whether the positive/negative items cause a wording effect and analyses can be made based on the appropriate measurement model.

### **Declaration of Conflicting Interests and Ethics**

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

### **Authorship contribution statement**

**Esra Oyar:** Investigation, Methodology, Resources, Visualization, Software, Formal Analysis and Writing, Supervision. **Hakan Yavuz Atar:** Methodology, Visualization, Supervision and Validation.

## ORCID

Esra OYAR  <https://orcid.org/0000-0002-4337-7815>

Hakan Yavuz ATAR  <https://orcid.org/0000-0001-5372-1926>

## 5. REFERENCES

- Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary school children. *Journal of Educational Measurement*, 22(3), 231-240. <https://doi.org/10.1111/j.1745-3984.1985.tb01061.x>
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Guilford Publications.
- Büyüköztürk, Ş. (2012). *Sosyal bilimler için veri analizi el kitabı: İstatistik, araştırma deseni, SPSS uygulamaları ve yorum* (16. Baskı). Pegem Akademi. [Handbook of data analysis for social sciences: Statistics, research design, SPSS practice and interpretation (16. Edition). Pegem Academy].
- Büyüköztürk, Ş., Çakmak, E. K., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2017). *Bilimsel araştırma yöntemleri*. Pegem Yayıncılık.
- Chen, Y. (2017). On the impact of negatively keyed items on the assessment of the unidimensionality of psychological tests and measures. [Doctoral dissertation, The University of British Columbia]. ProQuest Dissertations and Theses.
- Chen, Y. H., Rendina-Gobioff, G., & Dedrick, R. F. (2010). Factorial invariance of a Chinese self-esteem scale for third and sixth grade students: evaluating method effects associated with positively and negatively worded items. *The International Journal of Educational and Psychological Assessment*, 6 (1), 21-35.
- Chen, F. F., & Zhang, Z. (2018). Bifactor models in psychometric test development. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 325–345). John Wiley, Sons Ltd.
- Cronbach, L. J. (1984). *Essentials of psychological testing (4<sup>th</sup> edition)*. Harper & Row.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2<sup>nd</sup> edition)*. Lawrence Erlbaum.
- DeVellis, R. F. (2003). *Scale development: Theory and applications (2<sup>nd</sup> edition)*. Sage
- DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling*, 13, 440-464. [https://doi.org/10.1207/s15328007sem1303\\_6](https://doi.org/10.1207/s15328007sem1303_6)
- DiStefano, C. & Motl, R. W. (2009). Self-esteem and method effects associated with negatively worded items: Investigating factorial invariance by sex. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(1), 134-146. <https://doi.org/10.1080/10705510802565403>
- Erkuş A. (2003). *Psikometri üzerine yazılar. (1. baskı)*. Türk Psikologlar Derneği Yayınları.
- Ford, L. R., & Scandura, T. A. (2018). A typology of threats to construct validity in item generation. *American Journal of Management*, 18(2). <https://doi.org/10.33423/ajm.v18i2.298>
- Greenberger, E., Chen, C., Dmitrieva, J., & Farruggia, S.P. (2003). Item-wording and the dimensionality of the rosenberg self-esteem scale: Do they matter?. *Personality and Individual Differences*, 35(2003), 1241-1254. [https://doi.org/10.1016/S0191-8869\(02\)00331-8](https://doi.org/10.1016/S0191-8869(02)00331-8)
- Grimm, P. (2010). *Social desirability bias*. Wiley International Encyclopedia of Marketing. Hoboken, Wiley.
- Gu, H., Wen, Z., & Fan, X. (2015). The impact of wording effect on reliability and validity of the core self-evaluation scale (CSES): A bi-factor perspective. *Personality and Individual Differences*, 83, 142-147. <https://doi.org/10.1016/j.paid.2015.04.006>

- Harvey, R. J., Billings, R. S., & Nilan, K. J. (1985). Confirmatory factor analysis of the job diagnostic survey: Good news and bad news. *Journal of Applied Psychology, 70*, 461-468. <https://doi.org/10.1037/0021-9010.70.3.461>
- Hooper, M., Arora, A., Martin, M. O., & Mullis, I. V. S., (2013, June). *Examining the behavior of "reverse directional" items in the TIMSS 2011 context questionnaire scales*. Paper Presented at the 5th IEA International Research Conference. National Institute of Education, Nanyang Technological University, Singapore.
- Horan, P. M. , DiStefano, C. & Motl, R. W. (2003) Wording effects in self-esteem scales: methodological artifact or response style?. *Structural Equation Modeling, 10*(3), 435-455. [https://doi.org/10.1207/S15328007SEM1003\\_6](https://doi.org/10.1207/S15328007SEM1003_6)
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. <https://doi.org/10.1080/10705519909540118>
- Hyland, P., Boduszek, D., Dhingra, K., Shevlin, M., & Egan, A. (2014). A bifactor approach to modelling the Rosenberg Self Esteem Scale. *Personality and Individual Differences, 66*, 188-192. <https://doi.org/10.1016/j.paid.2014.03.034>
- Ibrahim, A.M. (2001). Differential responding to positive and negative items: The case of a negative item in a questionnaire for course and faculty evaluation. *Psychological Reports, 88*, 497–500. <https://doi.org/10.2466/pr0.2001.88.2.497>
- Kirk, R. (2007). *Statistics: an introduction*. Nelson Education.
- Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., Raudsepp, L., Liukkonen, J., & Thøgersen-Ntoumani, C. (2012). Method effects: The problem with negatively versus positively keyed items. *Journal of personality assessment, 94*(2), 196-204. <https://doi.org/10.1080/00223891.2011.645936>
- Marsh, H. W. (1986). The bias of negatively worded items in rating scales for young children: A cognitive-developmental phenomenon. *Developmental Psychology, 22*, 37-49.
- Marsh, H. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifacts?. *Journal of Personality and Social Psychology, 70*, 810-819. <https://doi.org/10.1037/0022-3514.70.4.810>
- Maul, A. (2013). Method effects and the meaning of measurement. *Frontiers in Psychology, 4*, 169. <https://doi.org/10.3389/fpsyg.2013.00169>
- McLarty, J. R., Noble, A. C., & Huntley, R. M. (1989). Effects of item wording on sex bias. *Journal of Educational Measurement, 26*(3), 285-293. <https://doi.org/10.1111/j.1745-3984.1989.tb00334.x>
- Michaelides, M. P. (2019). Negative keying effects in the factor structure of TIMSS 2011 motivation scales and associations with reading achievement. *Applied Measurement in Education, 32*(4), 365-378. <https://doi.org/10.1080/08957347.2019.1660349>
- Michaelides, M. P., Zenger, M., Koutsogiorgi, C., Brähler, E., Stöbel-Richter, Y., & Berth, H. (2016). Personality correlates and gender invariance of wording effects in the German version of the rosenberg self-esteem scale. *Personality and Individual Differences, 97*, 13-18. <https://doi.org/10.1016/j.paid.2016.03.011>
- Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*, 879-903. <https://doi.org/10.1037/0021-9010.88.5.879>
- Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47* (5), 667–696. <https://doi.org/10.1080/00273171.2012.715555>
- Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of personality assessment, 92*(6), 544-559. <https://doi.org/10.1080/00223891.2010.496477>

- Schmitt, N., & Stuits, D.M. (1985). Factors defined by negatively keyed items: The result of careless respondents?. *Applied Psychological Measurement*, 9, 367-373. <https://doi.org/10.1177/014662168500900405>
- Schriesheim, C. A., Eisenbach, R. J., & Hill, K. D. (1991). The effect of negation and polar opposite item reversals on questionnaire reliability and validity: An experimental investigation. *Educational and Psychological Measurement*, 51(1), 67-78. <https://doi.org/10.1177/0013164491511005>
- Tomas, J. M. & Oliver, A. (1999). Rosenberg's self-esteem scale: Two factors or method effects. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 84-98. <https://doi.org/10.1080/10705519909540120>
- Wang, W. C., Chen, H. F., & Jin, K. Y. (2015). Item response theory models for wording effects in mixed-format scales. *Educational and Psychological Measurement*, 75(1), 157-178. <https://doi.org/10.1177/0013164414528209>
- Wang, Y., Kim, E. S., Dedrick, R. F., Ferron, J. M., & Tan, T. (2018). A multilevel bifactor approach to construct validation of mixed-format scales. *Educational and psychological measurement*, 78(2), 253-271. <https://doi.org/10.1177/0013164417690858>
- Weems, G.H., Onwuegbuzie, A.J., & Collins, K.M.T. (2006). The role of reading comprehension in responses to positively and negatively worded items on rating scales. *Evaluation & Research in Education*, 19(1), 3-20. <https://doi.org/10.1080/09500790608668322>
- Weems, G. H., Onwuegbuzie, A. J., & Lustig, D. (2003). Profiles of respondents who respond inconsistently to positively-and negatively-worded items on rating scales. *Evaluation & Research in Education*, 17(1), 45-60. <https://doi.org/10.1080/14664200308668290>
- Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods*, 18, 320–334. <https://doi.org/10.1037/a0032121>
- Woods, C.M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, 28(3), 189–194. <https://doi.org/10.1007/s10862-005-9004-7>
- Wu, C. H. (2008). an examination of the wording effect in the rosenberg self-esteem scale among culturally chinese people. *The Journal of Social Psychology*, 148 (5), 535-552. <https://doi.org/10.3200/SOCP.148.5.535-552>
- Wu, Y., Zuo, B., Wen, F., & Yan, L. (2017). Rosenberg self-esteem scale: Method effects, factorial structure and scale invariance across migrant child and urban child populations in China. *Journal of personality assessment*, 99(1), 83-93. <https://doi.org/10.1080/00223891.2016.1217420>
- Yang, Y., Chen, Y. H., Lo, W. J., & Turner, J. E. (2012). Cross-cultural evaluation of item wording effects on an attitudinal scale. *Journal of Psychoeducational Assessment*, 30(5), 509-519. <https://doi.org/10.1177/0734282911435461>

## 6. APPENDIX

Table A1. Items in the Mathematics Self- Confidence Scale.

Codes	Items - English	Items - Turkish
BSBM19A	I usually do well in mathematics	Matematikte genellikle iyiyimdir.
BSBM19B	Mathematics is more difficult for me than for many of my classmates*	Matematik birçok sınıf arkadaşına göre bana daha zor gelir.*
BSBM19C	Mathematics is not one of my strengths*	Matematik başarılı olduğum alanlardan biri değildir. *
BSBM19D	I learn things quickly in mathematics	Matematik konularını hızlı öğrenirim.
BSBM19E	Mathematics makes me nervous*	Matematik beni gerginleştirir/endişelendirir.*
BSBM19F	I am good at working out difficult mathematics problems	Zor matematik problemleri çözmekte iyiyimdir.
BSBM19G	My teacher tells me I am good at mathematics	Öğretmenim matematikte iyi olduğumu söyler.
BSBM19H	Mathematics is harder for me than any other subject*	Matematik benim için diğer alanlardan daha zordur.*
BSBM19I	Mathematics makes me confused*	Matematik benim kafamı karıştırır.*

\*Reverse scored items.



## Teachers' Knowledge and Perception about Dyslexia: Developing and Validating a Scale

Duygu Tosun <sup>1,\*</sup>, Serkan Arikan <sup>1</sup>, Nalan Babur <sup>1</sup>

<sup>1</sup>Bogazici University, Faculty of Education, Istanbul, Turkey

### ARTICLE HISTORY

Received: Feb. 04, 2020

Revised: Feb. 28, 2021

Accepted: Mar. 16, 2021

### Keywords:

Scale development,  
Measurement invariance,  
Teacher knowledge of  
dyslexia,  
Teacher perception of  
dyslexia,

**Abstract:** Teachers have an important role in the achievement progress of students with dyslexia. Therefore, measuring teachers' knowledge and perception of dyslexia is important. Given that an instrument that measures both teachers' knowledge and perception of dyslexia is not available, this study aims to develop a scale to measure primary school teachers' knowledge and perception of dyslexia. Two hundred and one primary school teachers participated in the study, and exploratory factor analysis was conducted to identify the dimensions of the scale and to select scale items. Configural, metric and scalar invariance across gender groups was supported. This study also examines whether teachers' knowledge and perception of dyslexia differ with regard to their backgrounds. The results showed that there was no significant relationship between primary school teachers' teaching experience and their knowledge of dyslexia. Also, their knowledge of dyslexia did not differ with regard to other variables of the study. On the other hand, there was a positive, but weak relationship between teaching experience and teachers' negative perceptions of dyslexia. Primary school teachers who took a course about dyslexia in college had lower negative perceptions of dyslexia than teachers who did not do so. Teachers' perceptions did not differ with regard to taking an in-service seminar, reading a book or an article or teaching a student with dyslexia. The current study is expected to contribute to dyslexia research in terms of providing a scale to measure teachers' knowledge and perception of dyslexia.

## 1. INTRODUCTION

Learning to read is the primary goal for the first years of schooling. Students acquire reading skills through a systematic literacy education which mostly depends on language-based activities offered by teachers. Teachers are critical figures and play a significant role in teaching reading acquisition. General education or special education teachers who are specifically trained for effective reading instruction might be among the first to detect learning difficulties in students. Furthermore, teachers have a much more important role for students with dyslexia. Dyslexia is a language-based learning difficulty that affects word reading, spelling, and writing (Proctor et al., 2017; Vellutino et al., 2004).

It is reported that 80% of students who need special education suffer from dyslexia (National Center for Statistics, 2008). Demir (2005) reported that, according to parent surveys, 33% of the students in first grade were at risk for dyslexia in Turkey. On the other hand, first grade

---

\*CONTACT: Duygu Tosun ✉ [blgn\\_duygu@hotmail.com](mailto:blgn_duygu@hotmail.com) 📍 Esenyurt Anadolu Lisesi, İstanbul, Turkey



teachers indicated that 25% of first grade students displayed increased difficulties while learning to read and write (Demir, 2005). Research has shown that with the help of a teacher who provides appropriate reading instruction, students with dyslexia may have better academic success (e.g., Bos et al., 2001; Hornstra et al., 2010; Moats, 2009; Moats & Foorman, 2003; Rubin, 2002; Snow et al., 1998). It is also reported that the reading achievement of dyslexic students, in particular, is affected by their teachers' knowledge and capabilities (e.g., Gwernan-Jones & Burden, 2010; Hellendoorn & Ruijsenaars, 2000; Lane et al., 2009; Mills, 2006; Rubin, 2002). These studies proved that literacy acquisition should be done through effective and specialized approaches by a well-trained teacher (Brady & Moats, 1997; Rubin, 2002). In order to assist students to improve their reading skills and access content curriculum, all teachers should be aware of the effective instructional strategies on literacy (Boling & Evans, 2008; Gwernan-Jones & Burden, 2010). Teachers should have a high level of reading instruction knowledge for effectively teaching students because their choice of instructional and intervention programming is affected and guided by their knowledge (Foorman & Moats, 2004; Snow et al., 1998; Spear-Swerling & Brucker, 2004). In other words, more knowledgeable teachers are better equipped to facilitate reading achievement in students relative to those with less knowledge (Snow et al., 1998; Spear-Swerling & Brucker, 2004). Overall, more knowledgeable teachers are more likely to identify students with dyslexia compared to less knowledgeable ones (Gwernan-Jones & Burden, 2010; Spear-Swerling, 2009; Taylor et al., 2002).

Besides teachers' knowledge of dyslexia, how they perceive dyslexia has an important effect on students with dyslexia. It is known that in addition to knowledge, teachers' perception of dyslexia also affects the capability of dealing with dyslexia. A teacher who has a negative perception of dyslexia would be expected to rate the achievement level of dyslexic students as low (Hornstra et al., 2010). This negative perception causes teachers to decrease their expectations from dyslexic students. On the contrary, teachers who have a correct understanding of dyslexia are more likely to help students overcome challenges posed by their disability (Hornstra et al., 2010).

Teachers play a significant role in identifying and including students with dyslexia, so having accurate knowledge of dyslexia is critical. Therefore, it is important to explore what teachers really know about dyslexia as well as their perceptions of it. In order to do so, it is necessary to evaluate them with a valid and reliable scale.

### **1.1. Measuring Teachers' Knowledge and Perception of Dyslexia**

Teachers' knowledge and perception of dyslexia have attracted researchers' attention, and several studies have been conducted to measure teachers' knowledge and perception of dyslexia. For example, Ferrer, Bengoa, and Joshi (2016) investigated in-service and pre-service teachers' knowledge and beliefs of developmental dyslexia. They developed the Knowledge and Beliefs about Developmental Dyslexia Scale with 36 items. Every item in the scale is a statement about dyslexia and teachers are asked to evaluate the statements as true, false, or no idea. The scale measures teachers' knowledge and misconceptions about developmental dyslexia in three areas: General information about the nature, causes and outcome of developmental dyslexia, symptoms of developmental dyslexia and the treatment of developmental dyslexia. Their study indicated that teachers' knowledge was not correlated with their age and gender. A statistically significant correlation was found between pre-service teachers' scale scores and training about dyslexia in their university studies. In-service teachers' scale scores were significantly correlated with their years of teaching experience, postgraduate training about dyslexia, and prior exposure to a child with dyslexia. In-service teachers' knowledge of dyslexia was positively correlated to their self-confidence in teaching children with dyslexia. Washburn, Mulcahy, Musante and Joshi (2017) used a survey that included items

about fluency, word study, vocabulary and comprehension. Besides demographic information, teachers were also asked to answer two open-ended questions measuring characteristics of reading disability and characteristics of dyslexia. The results showed that certification area, certification grade level and exposure to literacy-related content did not predict teachers' knowledge of reading disabilities.

Research shows that a teacher's beliefs and perceptions may affect their classroom behavior and shape their teaching style (Nijakowska et al., 2018). Some teachers may not openly express their perceptions about students with dyslexia. Such teachers may be emotionally loaded, which may impact their instructional practices negatively and lead to resistance to change. Nijakowska and colleagues (2018) report that there seems to be a two-way interaction between teacher perceptions and educational practices. Even though teachers need to have a positive perception and sufficient knowledge regarding students with dyslexia, literature shows that many general education and special education teachers are not adequately prepared to teach children with dyslexia (e.g., Aktan, 2020; Balcı, 2019; Bos et al., 1999; Esen & Çiftçi, 2000; Fırat & Koçak, 2018; Mather et al., 2001; Moats, 2009; Şahin et al., 2020; Washburn et al., 2011). Teachers often may not be aware of their negative perception that may affect their teaching and attitudes towards children with dyslexia. When designing a professional training program, it is crucial to understand teachers' level of knowledge about dyslexia and their perception of students with dyslexia. Knowing teachers' perception of dyslexia may help researchers develop and design adequate professional training and teaching models.

In Turkey, although there are studies regarding dyslexia, these studies mainly focus on measuring the teachers' knowledge about dyslexia (Akçay, 2014; Altun et al., 2011; Altuntaş, 2010; Doğan, 2013; Yurdakal, 2014). Altuntaş (2010) and Doğan (2013) developed questionnaires and knowledge tests about dyslexia and used them as data-gathering instruments in their studies. Altun et al. (2011) conducted a qualitative study that used semi-structured interview techniques in the data collection process. However, these studies only measured teachers' knowledge of dyslexia. Research studies investigating teachers' knowledge and perceptions toward children with dyslexia are rare in Turkey (e.g., Başar & Göncü, 2018; Gever, 2017; Şahin et al., 2020). We, therefore, decided to develop a scale that would help us obtain information about teachers' knowledge and perception related to dyslexia.

In sum, many studies have shown that primary school teachers are not well equipped for supporting and educating students with dyslexia. Results of these studies consistently displayed that many primary school teachers lacked the accurate knowledge about dyslexia and research-based skills for teaching students with dyslexia (e.g., Aktan, 2020; Balcı, 2019; Esen & Çiftçi, 2000; Fırat & Koçak, 2018; Şahin et al., 2020; Washburn et al., 2011).

## **1.2. Correlates of Teachers' Knowledge and Perception of Dyslexia**

Studies emphasized that accurate knowledge and positive perception of dyslexia can help teachers to assist, teach and support students with dyslexia (Hornstra et al., 2010). For this reason, researchers investigated both teachers' knowledge and perception of dyslexia as well as the factors related to knowledge and perception. Ferrer et al. (2016) reported that in-service teachers' knowledge of dyslexia was related to the factors such as post-training of dyslexia, years of teaching experience, prior exposure to a dyslexic student, and high self-esteem. Washburn and colleagues (2017) conducted an exploratory study with 271 pre-service and in-service teachers in order to investigate novice teachers' knowledge about the characteristics of learning disabilities and dyslexia. Their findings showed that teachers had a clear understanding of learning disabilities when asked about reading disabilities, whereas they had misconceptions of dyslexia when asked about dyslexia. Their knowledge about learning disabilities and dyslexia was not dependent on certification type, certification grade level, or exposure to reading content. The results indicated that teachers listed more language and literacy-related

characteristics with the term learning disability than with the term dyslexia, which showed that teachers were confused about the true definition of dyslexia.

When we examined dyslexia studies conducted in Turkey, for example, Altuntaş (2010) study showed that teachers' knowledge of dyslexia was not related to having a dyslexic student and the type of school they work. Teachers generally had insufficient knowledge about dyslexia and did not feel well-prepared to teach dyslexic students. Altun et al. (2011) found that every teacher faced reading disabilities in their classrooms. Teachers perceived themselves as insufficient in the area of reading disabilities and did not feel capable of teaching students who struggled with them. Doğan (2013) showed that the reading disability knowledge level of Turkish language teachers who teach secondary school level was higher than that of primary school teachers. Turkish language teachers were also more successful in identifying students with reading disabilities relative to primary school teachers. Another important finding of the study was that novice teachers were much more knowledgeable about reading disabilities than experienced teachers. Akçay (2014) designed a study to determine elementary school teachers' awareness of dyslexic students from grade one to grade four. The findings revealed that elementary school teachers' awareness level of dyslexia didn't change according to the gender, teaching experience, type of certification, type of faculty, the grade of students they teach, their beliefs about their qualifications, taking an in-service training, and the classroom size. On the other hand, Yurdakal (2014) reported that primary school teachers' knowledge level of dyslexia was adequate. Last but not least, one of the most recent studies conducted by Şahin et al. (2020) examined primary school teachers' knowledge and attitudes toward dyslexia. The researchers reported that even though most teachers had positive attitudes toward students with dyslexia, the lack of knowledge and not having effective teaching skills showed the need for education and training related to dyslexia among educators. In sum, the studies mentioned here show that this topic requires urgent attention among educators and professionals in Turkey. Therefore, researchers should continue to explore this area in order to enhance understanding, knowledge, and a positive attitude toward dyslexia.

### 1.3. Present Study

In order to measure primary school teachers' knowledge and perception levels regarding dyslexia, the present study aimed to develop a reliable and valid scale using data from Turkey. Through this scale, the study investigated measurement invariance across groups to test the comparability of the subgroups. How teachers' knowledge and perception of dyslexia differ based on their background was also examined.

## 2. METHOD

### 2.1. Participants

The participants of the study were 201 primary school teachers who volunteered to participate. The study included 145 female (72.1 %) and 56 male teachers (27.9 %). Teaching experiences of teachers ranged from 1 to 23 years. The mean of the teaching experience was 11.01, the median was 10.00, and the standard deviation was 5.67. 19.4% of the teachers stated that they had never heard the term dyslexia. 87.1% of the teachers reported not having taken a course on dyslexia during their university education. Most of the teachers (93.5%) had not yet taken an in-service training of dyslexia. The vast majority of them (75.6%) did not read a book or an article on dyslexia. The majority of the teachers (70.1%) did not teach a student with dyslexia, and most of them (82.6%) thought that they had inadequate academic knowledge to teach a student with dyslexia.

## 2.2. Instrument

### 2.2.1. Teachers’ knowledge and perception scale

The aim of the study was to develop a scale to measure primary school teachers’ knowledge and perception of dyslexia. The scale was hypothesized to measure two dimensions: teachers’ knowledge of dyslexia and teachers’ perception of dyslexia. Based on a detailed literature review, investigation of current dyslexia questionnaires (Akçay, 2014; Yurdakal, 2014), and experts’ suggestions, a pool of items was developed by the researchers. Fifty-six items were developed initially to measure teachers’ knowledge of dyslexia and perception of dyslexia. **Table 1** provides a table of specification of the scale. The scale included 5-point Likert scale items. In the scale, teachers were asked to give 1 point to strongly disagree and 5 to strongly agree.

**Table 1.** *Table of Specification.*

Dimensions	Item Numbers
Knowledge of Dyslexia	1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 32, 33, 36, 37, 42, 43, 47, 49, 52, 53, 55
Perceptions of Dyslexia	10, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 34, 35, 38, 39, 40, 41, 44, 45, 46, 48, 50, 51, 54, 56

Items related to knowledge of dyslexia are statements that focus on the goals that a dyslexic student can achieve and cannot achieve. For example, the items ‘A student with dyslexia experiences difficulties in remembering the seasons and months in order’ and ‘A student with dyslexia needs to read the same paragraph again and again’ are items related to knowledge of dyslexia. Items measuring teachers’ perception of dyslexia are either pedagogical statements or statements about the general perception of dyslexia. For example, ‘Dyslexia is a disease’ and ‘A student with dyslexia should not receive an education with other students’ are exemplar items of perception dimension of the scale.

The questionnaire's demographic part had items related to gender, years of teaching experience, education level, and the type of department they graduated from. Additionally, this part of the scale aimed to get more data about teachers regarding dyslexia and included seven yes-no questions related to dyslexia. Yes or no questions ranged from: “Did you take a course on dyslexia during your university education?” to “Have you ever taken an in-service seminar on dyslexia?”

### 2.3. Data Analysis

The scale was first administered to 30 teachers in order to control the clarity and the language of the statements. All of the teachers were from public schools. The statements were revised according to the feedback of these 30 teachers and a researcher group’s suggestions. For instance, the question, “Do you think that you have sufficient academic knowledge to teach a student with dyslexia?” in the demographic part was included in the final form based on this feedback.

After completing the revisions, the scale was administered to the sample. In order to decide the dimensions and related items, exploratory factor analysis using the principal axis factor extraction technique with direct oblimin rotation was conducted. Problematic items that had 0.400 or less item loadings to a primary factor were discarded. Also, if an item was loaded to two factors simultaneously (factor loading difference of an item to a primary factor and other factor is less than 0.100) that item was also eliminated (Field, 2013).

The reliability of the data was evaluated using Cronbach's alpha coefficient. A reliable scale should have 0.70 or above Cronbach's alpha coefficient. Cronbach's alpha value above 0.70 is acceptable, 0.80 is good, and 0.90 and above is excellent. Higher values mean the data has higher internal consistency (George & Mallery, 2001)

To collect further evidence regarding the scale's structure, measurement invariance analysis for gender groups was conducted. As the differences between gender groups is a topic of interest of many researchers, providing evidence regarding measurement invariance for gender groups is required for valid comparisons. Having measurement invariance across gender groups implies that the scale scores of males and females can be comparable. To test measurement invariance, the fit values obtained in configural, metric and scalar models are compared. In the configural model, whether the same factor structure exists across the gender is tested. In the metric model, factor loadings of the BTPS were constrained to be equal across the gender groups. In the scalar model, item thresholds are constrained to be equal for males and females in addition to the factor loadings (Milfont & Fischer, 2010; Vandenberg & Lance, 2000). Measurement invariance is assessed by comparing  $\Delta CFI$  and  $\Delta RMSEA$  values with cutoff criteria ( $\Delta CFI \leq .01$ ,  $\Delta RMSEA \leq .015$ ) suggested by Chen (2007) and Cheung and Rensvold (2002).

After deciding the items related to each dimension, teachers' knowledge and perception scores were calculated. These scores were used to conduct correlational analysis and group comparison analysis to achieve the study's second goal. For the correlational analysis, the significance, direction, and magnitude of the relationship is evaluated. For the group comparisons, independent samples *t*-test was conducted and effect size (*d*) was estimated. According to Cohen (1988), *d* value around 0.20 represents a small difference, 0.50 means medium difference, and 0.80 implies large differences between the groups.

### 3. RESULT / FINDINGS

#### 3.1. Factor Structure of the Scale

The exploratory factor analysis was conducted and items that did not belong to any factor were eliminated. Kaiser-Meyer-Olkin's measure of sampling adequacy value of .840 indicated that the underlying factors might cause the proportion of variance in the items. Bartlett's test of sphericity ( $p < .05$ ) showed that the correlation matrix was different from an identity matrix. Therefore, the data was appropriate for conducting the exploratory factor analysis. As a result of the exploratory factor analysis procedure, two meaningful factors emerged. These two-factors explained 51% of the total variance. Table 2 shows factor loadings obtained as a result of exploratory factor analysis. Factor one included the items 16, 15, 9, 11, 8, 13, 18, 12, 3 and 17. All of the ten items were related to primary school teachers' knowledge of dyslexia as hypothesized. Therefore, the first dimension was named as *knowledge of dyslexia*. Factor two had the items 28, 24, 19, 20, 27, and 21. These 6 items were related to primary teachers' negative perception of dyslexia. The second dimension was called as *perception of dyslexia*.



**Table 2.** Rotated Factor Matrix of Exploratory Factor Analysis.

Item Number	Factor	
	1	2
q16	.747	
q15	.736	
q9	.721	
q11	.707	
q8	.703	
q13	.655	
q18	.630	
q12	.613	
q3	.575	
q17	.522	
q28		.753
q24		.675
q19		.652
q20		.649
q27		.585
q21		.549

### 3.2. Reliability of the Scale

In order to examine the reliability of the scale, Cronbach’s alpha coefficient was calculated for each dimension (see Table 3). Knowledge and perception dimension’s alpha values indicated good internal consistency. Additionally, Cronbach’s alpha value for all items was reported.

**Table 3.** Cronbach’s Alpha Coefficients.

	Dimensions		
	Knowledge	Perception	All Items
Cronbach’s Alpha	.89	.81	.78
Number of items	10	6	16

### 3.3. Measurement Invariance across Gender Groups

Configural, metric and scalar invariance of the scale across gender groups was evaluated (see Table 4). Configural invariance results indicated that fit indexes were within acceptable level (TLI = .904, CFI = .918, RMSEA = .100). This means that the factor structure of the scale was similar for males and females. Metric invariance results showed that the change in the metric model's fit values supported the invariance ( $\Delta CFI = .003$ ,  $\Delta RMSEA = -.005$ ). Metric invariance means that the factor loadings were equivalent across gender groups. Scalar invariance results showed that the fit values' change supported the invariance ( $\Delta CFI = -.007$ ,  $\Delta RMSEA = -.009$ ). Scalar invariance means that item thresholds were invariant and the mean score of males and females were comparable.



**Table 4.** Measurement Invariance Analysis Results of the Scale.

	$\chi^2$	df	$\chi^2/df$	TLI	CFI	RMSEA (90% CI)	$\Delta$ CFI	$\Delta$ RMSEA
Configural	412.12	206	2.02	.904	.918	.100 (.086-.114)	-	-
Metric	418.30	220	1.90	.914	.921	.095 (.081-.108)	.003	-.005
Scalar	465.15	266	1.75	.921	.928	.086 (.073 -.099)	.007	-.009

Note.  $\chi^2$  = Chi-square, df = degrees of freedom, TLI = Tucker Lewis index, CFI = comparative fit index, RMSEA = root mean square error of approximation; CI = confidence interval,  $\Delta$ CFI = change in values of CFI,  $\Delta$ RMSEA = change in values of RMSEA.

### 3.4. Descriptive Statistics of Scale Scores

The descriptive statistics of scale scores were reported in Table 5. The minimum plausible score was 10 and the maximum score was 50 for the knowledge factor. For the perception factor, the plausible minimum score was 6 and the maximum score was 30. Skewness and kurtosis values and histogram of the distributions indicated that knowledge scores had normal distribution and perception scores had right-skewed distribution.

**Table 5.** Descriptive Statistics of Dimensions,

	Knowledge	Perception
Mean	36.98	12.38
Median	37.00	12.00
Std. Deviation	7.37	5.35
Minimum	13.00	6.00
Maximum	50.00	30.00
Skewness	-0.02	0.76
Kurtosis	-0.35	0.12

### 3.5. Knowledge of Dyslexia and Related Demographic Variables

A high score on knowledge factor indicated a teacher has more knowledge about dyslexia. The results of Pearson Product Moment correlation analysis showed that there was no significant relationship between teachers' teaching experience and their knowledge of dyslexia ( $r = .01, p > .05$ ). Teachers' knowledge of dyslexia did not differ with regard to taking a course ( $t(196) = -.06, p > .05$ ), taking an in-service seminar of dyslexia ( $t(196) = .59, p > .05$ ), reading a book or an article of dyslexia ( $t(196) = -1.35, p > .05$ ), and teaching a student with dyslexia ( $t(196) = -1.10, p > .05$ ).

### 3.6. Perception of Dyslexia and Related Demographic Variables

High scores on this factor indicate teachers have negative perceptions regarding dyslexia. The correlational analysis results showed a weak positive relationship between primary school teachers' experience and their perception of dyslexia ( $r = .20, p < .01$ ). This means that experienced teachers have more negative perceptions regarding dyslexia. The results of the  $t$ -test indicated that there was a significant difference between teachers' perception of dyslexia concerning taking a course on dyslexia ( $t(193) = 3.06, p < .05$ ) and the effect size is large;  $d = -.82$ . Primary school teachers who took a course about dyslexia during university education had lower negative perception ( $M = 9.22, SE = .73$ ) compared to primary school teachers who did not take a course about dyslexia during university education ( $M = 12.78, SE = .41$ ). On the other hand, there was no significant difference between teachers' perception of dyslexia with regard to taking an in-service seminar of dyslexia ( $t(193) = -.81, p > .05$ ), with regard to reading a

book or an article on dyslexia ( $t(193) = 1.05, p > .05$ ) and with regard to teaching a student with dyslexia ( $t(193) = .57, p > .05$ ).

#### **4. DISCUSSION and CONCLUSION**

This study aimed to develop and validate a scale on primary school teachers' knowledge and perception regarding students with dyslexia. Evidence regarding the measurement invariance across gender groups was provided. This study also examined the factors related to teachers' knowledge and perception of dyslexia. The demographic questions provided an overview of teachers' knowledge and perception regarding dyslexia.

##### **4.1. Scale Development**

The primary aim of the study was to develop and validate a scale on primary school teachers' knowledge and perception of dyslexia. Compared to other studies, such as Gwernan-Jones & Burden's (2010) study, the main focus of the present study was to design and develop its own questionnaire for primary school teachers. It was shown in the current study that the scale measures two dimensions which are knowledge and perception of dyslexia. Teachers' Knowledge and Perception of Dyslexia Scale was shown to be a reliable scale with good internal consistency. Evidence related to the validity of the scale was also provided. This scale fills the gap in measuring teachers' knowledge and perception of dyslexia in Turkey and could be used in other studies to measure teachers' knowledge and perception of dyslexia. Measurement invariance results imply that the scores obtained using this scale could be used to compare gender groups.

##### **4.2. Factors Related to Dyslexia**

In the study, factors related to teacher knowledge and perception regarding students with dyslexia were also investigated. The results showed that there was not a significant relationship between teachers' knowledge of dyslexia and their teaching experience. In other words, teachers' knowledge of dyslexia did not increase based on the years they spent teaching. This finding of the study is similar to Akçay (2014). In her study, Akçay (2014) reported that primary school teachers' awareness levels did not change according to their teaching experience. On the contrary, Doğan (2013) revealed that novice teachers were much more knowledgeable about dyslexia than experienced teachers. Ferrer et al. (2016) reported that long years of teaching provided teachers with knowledge of dyslexia. In other words, according to Ferrer et al. (2016), experienced teachers are much more knowledgeable about dyslexia. Overall, in Turkey, there is a need for in-service training to improve teacher knowledge of dyslexia.

The current study found a weak positive relationship between primary school teachers' perception of dyslexia and their teaching experience. Similarly, Yurdakal (2014) reported that teachers' perception of educational activities regarding dyslexia differs according to their teaching experience and novice teachers have much more positive perceptions. It is shown in the current study that experienced teachers are more likely to perceive dyslexia more negatively. These results of the study may be due to the fact that a large percentage of the teachers (77.1%) who participated in the study did not take a course about dyslexia. Studies have revealed that teachers who were trained on dyslexia are more likely to have a positive perception of dyslexia (Hornstra et al., 2010). Additionally, primary school teachers who took a course about dyslexia during their university education had lower negative perceptions compared to primary school teachers who did not do so. In that regard, the current study has similar findings with Hornstra et al. (2010). These findings suggest that there is a need to educate experienced teachers who have not taken a course related to dyslexia.

Another finding of the study is related to teaching a student with dyslexia. The results showed that there was not a significant difference between teachers' knowledge of dyslexia and

teachers' perception of dyslexia between those who taught a student with dyslexia and those who did not. This result is consistent with the results of the study conducted by Altuntaş (2010) reporting that teaching a student with dyslexia did not contribute to teachers' knowledge. On the other hand, these findings are inconsistent with the findings of Ferrer et al. (2016). They reported that teachers' knowledge of dyslexia was related to being exposed to a student with dyslexia. The experience a teacher had and the support provided the teacher when teaching a student with dyslexia might affect the knowledge.

### 4.3. Teachers and Dyslexia

Demographic questions of the study also provided important information regarding to teachers and dyslexia. Findings of the study indicated that 19% of the primary school teachers, which is not a negligible percentage, did not hear the term dyslexia. This finding is consistent with Bingöl (2003), who reported that teachers were not aware of the term dyslexia. When primary school teachers' knowledge of dyslexia was investigated, interestingly enough, teachers reported that they had accurate knowledge of dyslexia. On the other hand, the amount of teachers (19%) who have misconceptions of dyslexia and do not have accurate knowledge of dyslexia should be taken into consideration. This finding indicates that not all of the primary school teachers are aware of the term dyslexia and they lack of the necessary knowledge to distinguish and support a student with dyslexia. Similarly, Başar and Göncü (2018) reported that primary school teachers have a conceptual misunderstanding about learning disabilities. Based on the findings of the study, many primary school teachers lacked research-based knowledge or had incorrect information about learning disabilities.

It is evident that primary school teachers play vital roles in the lives of students, especially students with dyslexia. Therefore, having an accurate knowledge of dyslexia is critically important. In this respect, the study has similar findings with Washburn and colleagues (2011) reporting that while some of the teachers have valid knowledge of dyslexia, some teachers have misconceptions about it. Some teachers' lack of knowledge about dyslexia was evident when they were asked to describe dyslexia.

Another interesting finding of the study showed that most teachers (83%) did not think that they had sufficient academic knowledge to teach a student with dyslexia. This finding is consistent with other studies reporting that the vast majority of the teachers lacked the necessary training about dyslexia and did not have sufficient skills when teaching students with dyslexia (Altun et al., 2011; Altuntaş, 2010; Bell et al., 2011; Moreau, 2014; Polat et al., 2012). The teacher training programs might be responsible for such a response here. Most of the teachers did not feel well prepared to teach dyslexic students and did not have adequate and accurate knowledge of dyslexia because most of them did not take a course on dyslexia during their university education (87.1%). According to Ferrer et al. (2016) the fact that teachers lack accurate knowledge of dyslexia is directly related to university coursework, university textbooks, and professional development courses.

In the present study, only a small percentage of the teachers took an in-service seminar on dyslexia (6.5%). Also, the findings revealed that there was not a significant difference in teachers' knowledge of dyslexia and teachers' perception of dyslexia even after having taken an in-service seminar. Therefore, these seminars on dyslexia are not reaching teachers and are not effective. The results are consistent with Akçay (2014) who argues that elementary teachers' awareness levels did not differ after taking in-service seminars. Teachers reported that they needed additional training on dyslexia and that they lacked the support they need to teach students with dyslexia (Polat et al., 2012).

Overall, the results of the present study revealed that primary school teachers in Turkey need professional support regarding dyslexia. As in-service teachers are likely to have students with special needs, including students with dyslexia in their classrooms, every teaching education

program should include courses on dyslexia. (Bos et al., 2001; Hornstra et al., 2010). Studies also reported that professional development and teacher qualification has an effect on teachers' perception of dyslexia (Bos et al., 2001; Hornstra et al., 2010; Mather et al., 2001). If teachers receive training of dyslexia, they have more positive perception of inclusive education. Furthermore, it is reported that teachers who received formal or informal training of dyslexia have more positive perceptions of individualized teaching than those teachers who did not receive training on dyslexia (Hornstra et al., 2010).

Based on the results of the studies here, it is clearly seen that teachers should be provided with specific, valid and research-based education on dyslexia. It is also shown that they are not adequately equipped with the skills to educate students with dyslexia (e.g., Altun et al., 2011; Altuntaş, 2010; Bell et al., 2011; Moreau, 2014; Polat et al., 2012; Şahin et al., 2020). Last but not least, the need for designing powerful, accurate and engaging workshops or seminars is very crucial regarding teacher training on dyslexia. Unfortunately, professional development seminars on dyslexia are, many times, poorly designed and not serving to the needs of the teachers. Therefore, professional development training programs should be given consistently and frequently. Such training programs should (a) be well-designed; (b) include powerful instructional strategies and activities on teaching dyslexia; (c) have up-to-date, evidence-based information about dyslexia.

#### **4.4. Limitations**

The present study had an important limitation based on sampling technique. Convenience sampling technique was used therefore the results of the study cannot be generalized to entire primary school teachers. There is a need to extend the sample and test the factor structure in future studies. Another limitation was that the current study focused on elementary school teachers. It would add valuable information to the literature to extend the sample by preschool teachers or middle school teachers. Testing discriminant validity with other scales might add value to the study; therefore, future research might be conducted to test the relationships between the current scale and other scales.

#### **Acknowledgments**

This paper was produced from the first author's master dissertation entitled Development of a scale on primary school teachers' knowledge and perception of dyslexia submitted to Boğaziçi University.

#### **Declaration of Conflicting Interests and Ethics**

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s). There is not any ethical violation in the study.

#### **Authorship contribution statement**

**Duygu Tosun:** Investigation, Resources, Visualization, Software, Formal Analysis, and Writing the original draft. **Serkan Arıkan:** Methodology, Supervision and Validation. **Nalan Babür:** Investigation, Framing, Supervision and Validation.

#### **ORCID**

Duygu Tosun  <https://orcid.org/0000-0001-5174-3910>

Serkan Arıkan  <https://orcid.org/0000-0001-9610-5496>

Nalan Babür  <https://orcid.org/0000-0002-7052-0488>

## 5. REFERENCES

- Akçay, D. (2014). *İlkokul 1-4. sınıf öğretmenlerinin disleksi ile ilgili farkındalık düzeylerinin incelenmesi* [Investigation of elementary school teacher's awareness level of dyslexia]. [Unpublished master's thesis]. Marmara University.
- Aktan, O. (2020). Determination of educational needs of teachers regarding the education of inclusive students with learning disability. *International Journal of Contemporary Educational Research*, 7(1), 149-164. <https://doi.org/10.33200/ijcer.638362>
- Altun, T., Ekiz, D., & Odabaşı, M. (2011). Sınıf öğretmenlerinin sınıflarında karşılaştıkları okuma güçlüklerine ilişkin nitel bir araştırma [A qualitative study on reading difficulties faced by primary teachers in their classrooms]. *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi*, 17, 80-101. <https://dergipark.org.tr/tr/pub/zgefd/issue/47948/606654>
- Altuntaş, F. (2010). *Sınıf öğretmenlerinin disleksiye ilişkin bilgileri ve dislektik öğrencilere yönelik çalışmaları* [Classroom teachers' knowledge about dyslexia and their trainings for dyslectic students]. [Unpublished master's thesis]. Hacettepe University.
- Balcı, E. (2019). Disleksi hakkında öğretmen görüşleri ve karşılaştıkları sorunlar [Teachers' opinions about dyslexia and the challenges they face]. *Ege Eğitim Dergisi*, 20(1), 162-179. <https://doi.org/10.12984/egeefd.453922>
- Başar, M. & Göncü, A. (2018). Sınıf öğretmenlerinin öğrenme güçlüğüyle ilgili kavram yanlışlarının giderilmesi ve öğretmen görüşlerinin değerlendirilmesi [Clearing misconceptions of primary school teachers about learning disabilities and evaluation of teacher opinions]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 33(1), 185-206. <http://www.efdergi.hacettepe.edu.tr/upload/files/2483-published.pdf>
- Bell, S., McPhillips, T., & Doveston, M. (2011). How do teachers in Ireland and England conceptualise dyslexia? *Journal of Research in Reading*, 34(2), 171-192. <https://doi.org/10.1111/j.1467-9817.2009.01419.x>
- Bingöl, A. (2003). Ankara'da ilkökul 2. ve 4. sınıf öğrencilerinde gelişimsel disleksi oranı [The prevalence of developmental dyslexia among the 2. and 4. grade students in Ankara]. *Ankara Üniversitesi Tıp Fakültesi Mecmuası*, 56(2), 67-82. [https://doi.org/10.1501/Tipfak\\_0000000053](https://doi.org/10.1501/Tipfak_0000000053)
- Boling, C. J., & Evans, W. H. (2008). Reading success in the secondary classroom. *Preventing School Failure*, 52(2), 59-66. <https://doi.org/10.3200/PSFL.52.2.59-66>
- Bos, C., Mather, N., Dickson, S., Podhajski, B., & Chard, D. (2001). Perceptions and knowledge of preservice and in-service educators about early reading instruction. *Annals of Dyslexia*, 51, 97-120.
- Bos, C., Mather, N., Friedman Narr, R., & Babur, N. (1999). Interactive, collaborative professional development in early literacy instruction: Supporting the balancing act. *Learning Disabilities Research and Practice*, 14(4), 227-238.
- Brady, S., & Moats, L. C. (1997). *Informed instruction for reading success- foundations for teacher preparation* [Paper presentation]. International Dyslexia Association, Baltimore. [https://www.researchgate.net/publication/234653061\\_Informed\\_Instruction\\_for\\_Reading\\_Success\\_Foundations\\_for\\_Teacher\\_Preparation](https://www.researchgate.net/publication/234653061_Informed_Instruction_for_Reading_Success_Foundations_for_Teacher_Preparation)
- Chen, F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modelling*, 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, 233-255. [https://doi.org/10.1207/S15328007SEM0902\\_5](https://doi.org/10.1207/S15328007SEM0902_5)
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.



- Demir, B. (2005). *Okulöncesi ve ilköğretim birinci sınıfa devam eden öğrencilerde özel öğrenme güçlüğüünün belirlenmesi* [Assesment of learning disability in kindergarten and first grade primary school students]. [Unpublished master's thesis]. Marmara University.
- Doğan, B. (2013). Türkçe ve sınıf öğretmenlerinin okuma güçlüğüne ilişkin bilgileri ve okuma güçlüğü olan öğrencileri belirleyebilme düzeyleri [Determining Turkish Language and elementary classroom teachers' knowledge on dyslexia and their awareness of diagnosing students with dyslexia]. *Okuma Yazma Eğitimi Araştırmaları*, 1(1), 20-33. <https://dergipark.org.tr/tr/pub/oyea/issue/20479/218123>
- Esen, A., & Çiftçi, İ. (2000). Sınıf öğretmenlerinin öğrenme yetersizliği ile ilgili bilgilerinin belirlenmesi. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, 8(8), 85-90. <https://dergipark.org.tr/tr/download/article-file/114887>
- Ferrer, M. S., Bengoa, C. E., & Joshi, R. M. (2016). Knowledge and beliefs of developmental dyslexia in pre-service and in-service Spanish-speaking teachers. *Annals of dyslexia*, 66, 91–110. <https://doi.org/10.1007/s11881-015-0111-1>
- Fırat, T., & Koçak, D. (2018). Sınıf öğretmenlerinin öğrenme güçlüğüünün tanımına ilişkin görüşleri [Investigating the opinions of class teachers' on the concept of learning difficulty]. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, 18(2), 915-931. <https://doi.org/10.17240/aibuefd.2018..-431461>
- Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.
- Foorman, B. R., & Moats, L. C. (2004). Conditions for sustaining research-based practices in early reading instruction. *Remedial and Special Education*, 25(1), 51- 60.
- George, D., & Mallery, P. (2001). *SPSS for Windows Step by Step: A Simple Guide and Reference*. Allyn & Bacon.
- Gever, A. (2017). *İlkokul ve ortaokul müdürlerinin disleksiye ilişkin bilgi düzeylerinin belirlenmesi* [Determination of levels of knowledge of discretion of primary and secondary school directors]. [Unpublished master's thesis]. Pamukkale University.
- Gwernan-Jones, R., & Burden, R. L. (2010). Are they just lazy? Student teachers' attitudes of dyslexia. *Dyslexia*, 16(1), 66–86. <https://doi.org/10.1002/dys.393>
- Hellendoorn, J., & Ruijssenaars, W. (2000). Personal experiences and adjustment of Dutch adults with dyslexia. *Journal of Remedial and Special Education*, 21(4), 227-239. <https://doi.org/10.1177%2F074193250002100405>
- Hornstra, L., Denessen, E., Bakker, J., van den Bergh, L., & Voeten, M. (2010). Teacher attitudes toward dyslexia. Effects on teacher expectations and the academic achievement of students with dyslexia. *Journal of Learning Disabilities*, 43, 515–529. <https://doi.org/10.1177%2F0022219409355479>
- Lane, H. B., Hudson, R. F., Leite, W. L., Kosanovich, M. L., Strout, M. T., & Fenty, N. (2009). Teacher knowledge about reading fluency and indicators of students' fluency growth in reading first schools. *Reading and Writing Quarterly*, 25, 57-86. <https://doi.org/10.1080/10573560802491232>
- Mather, N., Bos, C., & Babur, N. (2001). Perceptions and knowledge of preservice and inservice teachers about early literacy instruction. *Journal of Learning Disabilities*, 34(5), 472-482. <https://doi.org/10.1177%2F002221940103400508>
- Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research*, 3(1), 112-131. <https://doi.org/10.21500/20112084.857>
- Mills, C. (2006). *Preservice teacher education and the development of socially just dispositions. A review of the literature* [Paper presentation]. Annual Conference of the Australian Association for Research in Education, Adelaide. <https://www.aare.edu.au/data/publications/2006/mil06221.pdf>



- Moats, L. (2009). Knowledge foundations for teaching reading and spelling. *Reading and Writing: An interdisciplinary Journal*, 22, 379-399. <https://psycnet.apa.org/doi/10.1007/s11145-009-9162-1>
- Moats, L. C., & Foorman, B. R. (2003). Measuring teachers' content knowledge of language and reading. *Annals of Dyslexia*, 53, 23- 45. <https://doi.org/10.1007/s11881-003-0003-7>
- Moreau, L. K. (2014). Who's really struggling? Middle school teachers' perceptions of struggling readers. *Research in Middle Level Education Online*, 37(10), 1-17. <https://doi.org/10.1080/19404476.2014.11462113>
- National Center for Statistics (2008). *The condition of education 2008. Indicator 5: Language Minority school aged children*. Washington. <https://nces.ed.gov/pubs2008/2008031.pdf>
- Nijakowska, J., Tsagari, D., & Spanoudis, G. (2018). English as a foreign language training needs and perceived preparedness to include dyslexia learners: The case of Greece, Cyprus, and Poland. *Dyslexia*, 24, 357-379. <https://doi.org/10.1002/dys.1598>
- Polat, E., Adiguzel, T., & Akgun, O. E. (2012). Adaptive web-assisted learning system for students with specific learning disabilities A needs analysis study. *Educational Sciences Theory and Practice*, 12(4), 3243-3258. <https://files.eric.ed.gov/fulltext/EJ1003015.pdf>
- Proctor, C. M., Mather, N., Stephens-Pisecco, T., & Jaffe, L. E. (2017). Assessment of dyslexia. *Communique*, 46(3), 1-10.
- Rubin, D. (2002). *Diagnosis and correction in reading instruction*. Allyn and Bacon.
- Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. National Academy Press.
- Spear-Swerling, L. (2009). A literacy tutoring experience for prospective special educators and struggling second graders. *Journal of Learning Disabilities*, 42, 431-443. <https://doi.org/10.1177%2F0022219409338738>
- Spear-Swerling, L., & Brucker, P. O. (2004). Preparing novice teachers to develop basic reading and spelling skills in children. *Annals of Dyslexia*, 54(2), 332-364. <https://doi.org/10.1007/s11881-004-0016-x>
- Şahin, R., Güven, S. & Alatlı, B. (2020). Sınıf öğretmenlerinin disleksiye yönelik bilgi ve tutumlarının incelenmesi [Investigation of the knowledge and attitudes of primary school teachers towards dyslexia]. *Turkish Studies-Education*, 15(4), 2355-2372. <http://dx.doi.org/10.47423/TurkishStudies.42099>
- Taylor, B. M., Pressley, M., & Pearson, P. D. (2002). Research-supported characteristics of teachers and schools that promote reading achievement. In B. M. Taylor, & P. D. Pearson (Eds.). (2002). *Teaching reading: Effective schools, accomplished teachers*. (pp. 361-373). Lawrence Erlbaum. [https://www.researchgate.net/publication/312839932\\_Research-supported\\_characteristics\\_of\\_teachers\\_and\\_schools\\_that\\_promote\\_reading\\_achievement](https://www.researchgate.net/publication/312839932_Research-supported_characteristics_of_teachers_and_schools_that_promote_reading_achievement)
- 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, 4-70. <https://doi.org/10.1177%2F109442810031002>
- Vellutino, F. R., Fletcher, J. M., Snowling, M. J., & Scanlon, D. M. (2004). Specific reading disability (dyslexia). What have we learned in the past four decades. *Journal of Child Psychology and Psychiatry*, 45(1), 2-40. <https://doi.org/10.1046/j.0021-9630.2003.00305.x>
- Washburn, E. K., Joshi, R. M., & Binks-Cantrell, E. S. (2011). Teacher knowledge of basic language concepts and dyslexia. *Dyslexia*, 17, 165–183. <https://doi.org/10.1002/dys.426>
- Washburn, E. K., Mulcahy, C. A., Musante, G., & Joshi, R. M. (2017). Novice teachers' knowledge of reading related disabilities and dyslexia. *Learning Disabilities: A Contemporary Journal*, 15(2), 169-191. <https://files.eric.ed.gov/fulltext/EJ1160653.pdf>

Yurdakal, İ. H. (2014). *İlkokullarda okuma güçlüğünde yaşanan sorunlar ile eğitim uygulamalarına ilişkin öğretmen ve öğrenci görüşleri* [Teachers' and students' views regarding to problems encountered at primary schools related to reading disorders and educational activities for dyslexic students]. [Unpublished master's thesis]. Pamukkale University.

## Developing a Two-Tier Proportional Reasoning Skill Test: Validity and Reliability Studies

Kubra Acikgul <sup>1,\*</sup>

<sup>1</sup>İnönü University, Faculty of Education, Department of Mathematics Education, Malatya, Turkey

### ARTICLE HISTORY

Received: Dec. 08, 2020

Revised: Mar. 02, 2021

Accepted: Mar. 16, 2021

### Keywords:

Proportional reasoning,  
Two-tier test,  
Middle school.

**Abstract:** The main aim of this study is to develop a useful, valid, and reliable two-tier proportional reasoning skill test for middle school 7th and 8th-grade students. The research was carried out using the sequential explanatory mixed method. The study group of this research comprised of 391 ( $n_{7th-grade} = 223$ ,  $n_{8th-grade} = 168$ ) students. With validity and reliability studies, the content, face, construct, discriminant validity, and reliability coefficient of the test were examined. As a result, the two-tier proportional reasoning skill test with 12 items under 3 factors (qualitative prediction and comparison, missing value, numerical comparison) valid and reliable for adequate values specified in the literature.

## 1. INTRODUCTION

The concept of proportional reasoning has wide usability in mathematics education. According to the National Council of Teachers of Mathematics (NCTM) (2000), proportional reasoning is a unifying concept since most of the many important concepts of elementary school mathematics are related to it. The fact that proportional reasoning is an important factor in establishing relations between concepts makes it to be accepted as one of the basic thoughts that form the core of the mathematics curriculum (Lesh et al., 1988). For example, according to the Common Core State Standards for School Mathematics, proportional reasoning is one of the basic math skills in the USA. Similarly, proportional reasoning also has an important place in the Middle School Mathematics Curriculum in Turkey (Ministry of National Education [MoNE], 2018). Further to that, proportional reasoning skills are the basis for gaining higher mathematical reasoning beyond elementary school mathematics (Lesh et al., 1988). Many subjects of geometry, analysis, and algebra require students to have proportional reasoning skills (Allain, 2000). Thus, it is called the capstone of elementary concepts (e.g., arithmetic, measurement) and the cornerstone of higher-level mathematics (Lesh et al., 1988; Post et al., 1988). Despite the mentioned importance, interest, and emphasis in the curriculums and literature, proportional reasoning skill is mentioned complex, difficult to teach, and cognitively challenging for students (Alfieri et al., 2015; Lamon, 2007). So, it is considered important to examine students' proportional reasoning skills to determine the difficulties experienced by

\*CONTACT: Kübra AÇIKGÜL ✉ [kubra.acikgul@inonu.edu.tr](mailto:kubra.acikgul@inonu.edu.tr) 📍 İnönü University, Faculty of Education, Department of Mathematics Education, Malatya, Turkey

students and teachers. Hilton et al. (2013) stated that developing a diagnostic tool to measure the proportional reasoning skills of the students could be valuable for teachers to determine the teaching activities that are suitable for the specific needs of the students in proportional reasoning education.

This study aimed to develop a useful, valid, and reliable two-tier Proportional Reasoning Skill Test (PRST) for middle school 7th and 8th-grade students. In this study, the content validity, face validity, construct validity, discriminant validity, and Cronbach alpha, and composite reliability coefficients of the PRST were examined. The ability of students to distinguish non-proportional problems from proportional problems shows that they have proportional reasoning (Lim, 2009). So, in the current study, to determine the discriminant validity of the PRST, the relationship between PRST scores and Non-Proportional Reasoning Skill Test (N-PRST) scores was examined. Therefore, this study also aimed to develop a useful, valid, and reliable two-tier N-PRST.

### **1.1. Proportional Reasoning**

Proportional reasoning is fundamental to understand many situations in mathematics and science education (e.g., density, speed) and everyday life problems (Cramer & Post, 1993). Therefore, the importance of proportional reasoning is mentioned and defined in many studies. Tourniaire and Pulos (1985, p. 181) define proportion as “a statement of the equality of two ratios i.e.  $a/b=c/d$ ”. According to Behr et al. (1988, p. 92), proportional reasoning is “a form of mathematical reasoning that involves a sense of covariation and multiple comparisons, and the ability to store and process several pieces of information”. Lamon (2007, p. 647) refers to proportional reasoning as “detecting, expressing, analyzing, explaining, and providing evidence in support of assertions about proportional relationships”. Also, many researchers (e.g., Behr et al., 1992; Cramer et al., 1993; Lesh et al., 1988) explain proportional reasoning as an understanding of the comparisons between quantities embedded in proportional situations. Proportional reasoning is also mentioned as an ability to distinguish between proportional and non-proportional situations (Cramer et al., 1993; Lesh et al., 1988; Lim, 2009). Considering these definitions, in this study, proportional reasoning is defined as an ability that includes multiplicative comparisons between quantities and requires distinguishing between proportional and non-proportional situations.

According to Weinberg (2002), most students have difficulty with proportional reasoning as that they do not understand the proportional situation or the solution strategy to be used. Various researches reveal that (e.g., Arıcan, 2019; Cramer et al., 1993; Dinç-Artut & Pelen, 2015; Mersin, 2018; Pelen & Dinç-Artut, 2015; Singh, 2000; Van Dooren et al., 2010), students could not distinguish proportional and non-proportional situations, and they use additive reasoning instead of multiplicative reasoning while solving proportional problems. Moreover, students use many faulty solution strategies such as not being able to determine when to use proportional reasoning (Van De Walle et al., 2013) and ignoring some of the data given in the problems (Özgün-Koca & Altay, 2009).

Students also have difficulty solving different types of proportional reasoning problems. For example, Lawton (1993) determined that the problem type is a factor that influences proportional reasoning. Dinç-Artut and Pelen (2015) found that the problem types affect the strategies used by students. According to Soyak and Işıksal (2017), to overcome the students' difficulties about proportional situations, students should be exposed to different proportional reasoning problem types. Therefore, it was decided to include different types of problems in the PRST.

Cramer and Post (1993) classify proportional reasoning problems in three categories; missing-value problems, numerical comparison problems, and qualitative prediction and comparison

problems. In missing value problems, three values of four numerical values are given and the other value is asked. In numerical comparison problems, two rates are given and the rates are compared. Qualitative prediction and comparison problems require comparisons that are not dependent on specific numerical values. The test developed in this study includes these three problem types presented by Cramer and Post (1993).

Non-proportional reasoning problems are also used in determining proportional reasoning in this study. Van Dooren et al. (2005) classify non-proportional problems as additive, constant, and linear. The linear problems are in the form of  $f(x) = ax + b$  ( $b \neq 0$ ). The additive problems are expressed as a constant difference between two variables. In the constant problems, there is no relation between the given variables. In this study, Van Dooren et al. (2005) classification was used in developing the N-PRST.

## 1.2. Significance of the Study

Many important topics of the elementary school curriculum are linked to proportional reasoning skills (NCTM, 2000). Especially middle school (5-8th grades) is considered as a critical period to create meanings about proportion reasoning (NCTM, 2000; Van Dooren et al., 2010). Therefore, proportional reasoning is one of the essential mathematical skills that middle school students should have. According to Ayan and Isiksal-Bostan (2019), middle school years are the best period for forming new understandings about proportions and developing proportional reasoning, so it is important to examine the middle school students' proportional reasoning skills. Thus, in this study developing a PRST for middle school students is considered crucial.

Open-ended and multiple-choice problems are problem types measuring students' learnings and understandings in education and research fields (Ozuru et al., 2013). Similarly, when proportional reasoning tests for middle school students are examined, it is seen that most tests consist of open-ended or multiple-choice problems. Also, the number of tests consisting of open-ended questions is significantly greater than the number of multiple-choice tests. For example, Lawton (1993) developed an 8-item written test for 6th-grade students consisting of conventional ratio and proportion problems. Allain (2000) developed a valid and reliable proportional reasoning instrument for girls (6-8th-grades) studying at a middle school in North Carolina. The instrument consists of 10 open-ended problems (part-part-whole, associated sets, comparison, missing value, mixture, graphing, and scale problems). Duatepe et al. (2005) prepared a proportional reasoning test consisting of 10 open-ended items (missing value, quantitative comparison, qualitative comparison, non-proportional type relation, and inverse relation) by using the problems in the literature. Akkus and Duatepe-Paksu (2006) developed a measurement tool and rubrics for measuring and evaluating proportional reasoning skills. The test consists of 15 open-ended problems (missing value, quantitative comparison, qualitative comparison, non-proportional type relation, and inverse relation) applied to the 7th-grade and 8th-grade. Pelen and Dinç-Artut (2015) developed a proportional reasoning test consisting of 24 open-ended missing value word problems. Although most of the proportional reasoning tests consist of open-ended problems, there are a few multiple-choice tests to measure middle school students' proportional reasoning skills. According to Bright et al. (2003), different methods are likely to reveal different information about students' proportional reasoning. Therefore, it is also important to use multiple assessment methods, such as multiple-choice and constructed-response. Their test consisted of 4 multiple-choice problems and 1 constructed problem for 8th-grade and 9th-grade students. Arıcan (2019) developed a proportional reasoning test for middle school students consisting of 22 multiple-choice problems.

As mentioned above, to determine the proportional reasoning skills of middle school students open-ended questions are often used, however, researchers have mentioned some disadvantages of using open-ended problems. According to Hilton et al. (2013), open-ended problems are powerful methods for determining students' understanding, but their use may not be practical



in cases when the number of students is high. Reja et al. (2003) stated that with open-ended problems practitioners have the opportunity to discover the answers that students give spontaneously, but open-ended problems have disadvantages such as needing extensive coding when compared with closed-ended problems. Similarly, Hyman and Sierra (2016) emphasized that, in closed-ended problems such as multiple-choice questions, data is coded and analyzed quickly. According to Burton et al. (1991), multiple-choice problems have applicability in measuring higher-order targets such as understanding, application, and analysis. Despite these features, Hyman and Sierra (2016) pointed out that in multiple-choice problems, in-depth information cannot be obtained as students read only a few of the options and choose the option that best represents their views or behaviors. Similarly, Burton et al. (1991) has indicated that multiple-choice tests do not allow students to determine certain learning outcomes such as produce original ideas, organized personal thoughts.

Considering the aforementioned advantages and disadvantages, two-stage diagnostic tests, were developed in the 1980, have the positive aspects of the multiple-choice tests and minimize their disadvantages (e.g., Haslam & Treagust, 1987; Peterson et al., 1986). In two-tier tests, students have two tasks. The first tier includes multiple-choice problems and asks a student to make a choice, and the second tier asks the student justifications for choices made in the first tier (Haja & Clarke, 2011; Tsui & Treagust, 2010). These tests require more time, more effort, and higher-order skills such as reading, thinking, interpreting, understanding skills (Afnia & Istiyono, 2020; Haja & Clarke, 2011). Also, they give students a chance to justify their choice, thus it shifts the focus to the mathematical reasoning process rather than only answering (Haja & Clarke, 2011).

According to Tamir (1990), there are two important reasons /advantages of justifying multiple-choice problems. First, the students who are asked to justify their choices in the multiple-choice section must explain the reasons for their choice by considering all the options. Secondly, the student, who is aware that his/her choice will be justified, tries to learn the topics in-depth and will be ready to write a complete and adequate justification. Thus, a two-tier test is an effective and sensitive way to evaluate meaningful learning (Tamir, 1989, 1990). Despite the mentioned advantages, Haja and Clarke (2011) pointed out that two-tier tests are not widely used in mathematics education. Indeed, when the proportional reasoning tests for middle school students are examined, two-tier tests are found in just a few studies (Haja & Clarke, 2011; Hilton et al., 2013; Mersin, 2018).

Haja and Clarke (2011) aimed to evaluate middle school students' justification skills with two-tier proportional reasoning tasks. The tasks were "select answer" tasks and "marked answer" tasks. The select answer tasks had two types: 1. The student selects the answer, 2. The student selects the answer and writes a justification. Similarly, marked answer tasks had two types: 1. The student selects justification for the marked answer, 2. The student writes a justification for the marked answer. As a result of the study, it is stated that the two-tier tasks give more information about the students' alternative conceptions and reveal the students' reasoning.

Hilton et al. (2013) draw attention to the importance of justifying students' answers to understand the students' proportional reasoning and developed a two-tier multiple-choice diagnostic test for middle school students. The test consists of 12 items related to non-proportional (constant/additive), one or two-dimensional scale, missing value, familiar rate, rate, translation of representations, relative thinking, inverse proportion. Mersin (2018) mentioned that proportional reasoning measuring tools used in Turkey do not allow students to justify their answers and so she translated the two-tier proportional reasoning diagnostic test developed by Hilton et al. (2013) into Turkish. Since the two-tier PRST in the literature is a limited number, and there isn't any two-tier PRST developed in Turkish, it is considered important to develop a two-tier PRST that can allow to justify their answers.



According to Haja and Clarke (2011), in two-tier problems, if students are asked to construct their justification, given tasks become more intellectually demanding and require students to have more sophisticated expression skills. For this reason, the first tier of the test developed in this study is multiple-choice, and the second tier is prepared in an open-ended format in which students can explain their justifications verbally using their expressions. It is believed that determining the proportional reasoning skills of students with a two-tier PRST can benefit researchers, curriculum planners, and teachers. By using PRST, they can determine and understand the proportional reasoning skill levels of students more accurately. Also, they can determine whether students can differentiate between proportional and non-proportional situations.

## 2. METHOD

### 2.1. Design

In this research, sequential exploratory mixed-method research was used to develop PRST. The sequential exploratory mixed method is a two-phase model. In the first phase, the researcher studies the subject qualitatively; and in the second phase, s/he continues his/her study quantitatively (Creswell & Plano Clark, 2011). In the qualitative phase of the study, a problem pool consisting of two-tier problems was prepared and experts' opinions were taken for face validity. The content validity was evaluated both qualitative (experts' opinions) and quantitative (Davis's (1992) method) methods. Also, in the quantitative stage, construct validity, discriminant validity, and reliability were tested.

### 2.2. Study Group

The study group of this research comprised of 391 middle school students studying in two public schools in the south of Turkey. The convenience sampling method was used to determine the study group. Students were studying in schools where the researcher could easily reach in terms of time and place (Cohen et al., 2013). The aim of the study was explained to the students and volunteer students who participated in the study. The PRST comprised ratio and proportion subjects of the 7th-grade mathematics curriculum (MoNE, 2018). So, the research was conducted with 7th ( $n_{\text{female}}=126$ ,  $n_{\text{male}}=97$ ), and 8th ( $n_{\text{female}}=96$ ,  $n_{\text{male}}=72$ ) grade students.

### 2.3. Procedure

The following stages were followed in the test development process.

#### 2.3.1. Determining the purpose of the test

First, the purpose of the test was determined. The purpose of the test is to determine the proportional reasoning skills of 7th and 8th-grade students as valid and reliable.

#### 2.3.2. Determining the scope and properties of the test and item writing

For determining the scope of the test, qualitative methods (document review, literature review, expert opinions) were used. Firstly, the Mathematics Curriculum (MoNE, 2018) and mathematics textbooks of middle school were examined. It was determined that the subject of ratio-proportion was taught more in 7th-grade. For this reason, the test items were prepared within the context of 7th-grade learning outcomes. The learning outcomes related to the ratio-proportion at the 7th-grade in the Mathematics Curriculum (MoNE, 2018) are presented below.

Learning Outcome 1: If one of the quantities is 1, it determines the value of the other.

Learning Outcome 2: Given one of the two quantities whose ratio is given to each other, it finds the other.

Learning Outcome 3: Decides whether the two quantities are proportional by examining real-life situations.

Learning Outcome 4: Expresses the relationship between two direct proportional quantities.

Learning Outcome 5: Determines and interprets the proportionality constant of two direct proportional quantities.

Learning Outcome 6: Decides whether two quantities are inverse proportional by examining real-life situations.

Learning Outcome 7: Solves problems related to direct and inverse proportion.

Then the literature was reviewed and it was seen that there were different types of problems in determining proportional reasoning skills. Cramer and Post (1993) categorized proportional reasoning problems into three categories: missing-value problems, numerical comparison problems, and qualitative prediction and comparison problems. Considering the Cramer and Post's (1993) category and learning outcomes related to the ratio-proportion at the 7th-grade, the problem pool consisted of 15 problems, 5 of which was qualitative prediction and comparison, 5 of which was missing value, and 5 of which was numerical comparison. Problems included real-life contexts. The problems were prepared to have two-tier answer options. The first tier consisted of a multiple-choice answer, with four choices. The second tier was the open-ended answer part, which includes justifying the answer given in the multiple-choice section. The PRST is presented in the Appendix.

To determine the discriminant validity of the PRST, its correlation with N-PRST was examined. In this context, a two-tier N-PRST was developed in this study. The N-PRST problem pool consisted of 6 problems including 2 problems on additive situations, 2 problems on linear situations, and 2 problems on constant situations (Van Dooren et al., 2005). Problems included real-life contexts and had two-tier answer options: first-tier multiple-choice answers, second-tier open-ended answers.

**Additive:** Both Harmankaya Family and Orçan Family have two children. The sum of the Harmankaya Family's ages is 50, while the sum of the Orçan Family's ages is 60. Accordingly, what will be the sum of the Orçan Family's ages when the sum of the Harmankaya Family's ages will be 100?

- A) 2 times the sum of the Harmankaya Family's ages.
- B) 10+ of the sum of the Harmankaya Family's current ages.
- C) 2 times the sum of the Orçan Family's current ages.
- D) 50+ of the sum of the Orçan Family's current ages.

Justification:

**Linear:** Yusuf, who has TL40, starts to save a fixed amount of money every week to buy the computer game he wants. If Yusuf has a total of TL120 at the end of 4 weeks, how much will he have after 8 weeks?

- A) 8 times of his initial money
- B) 2 times of the money he has at the end of 4 weeks
- C) TL240 more than his initial money
- D) TL80 more than the money he has at the end of 4 weeks

Justification:

**Constant:** Baker Mehmet, who has an oven with a maximum capacity of 60 loaves of bread at a time, bakes bread as the number of the orders he receives each time. He started to receive orders as soon as he opens his bakery. First Ahmet orders 20 loaves of bread for his döner shop, then Yusuf wants 40 loaves of bread for his restaurant. Baker Mehmet bakes the first 20 loaves of bread in 12 minutes. In how many minutes does Uncle Mehmet bake the remaining 40 loaves of bread?

- A) 24 minutes
- B) 12 minutes

- C) 36 minutes  
D) The information provided is inadequate.

Justification:

### **2.3.3. Content and face validity**

The draft PRST and N-PRST were submitted to three experts (one assessment and evaluation expert and two mathematics education experts) to check the face and content validity using an expert opinion form. First, the experts were informed about the content of the PRST (qualitative prediction and comparison, missing value, numerical comparison problems) and N-PRST (additive, constant, and linear problems). The experts assessed the test items in terms of clarity, suitability to the purpose, suitability to the level of 7th-grade and 8th-grade students. Their opinions were evaluated using Davis's (1992) method. According to Davis's (1992) method, each item was evaluated choosing one of these: "a) Appropriate", "b) Item should be slightly revised", "c) Item should be reviewed", "d) Item is not appropriate". The Content Validity Index (CVI) was calculated for each item using the formula  $a+b/n$  ( $a$ = number of experts who ticked the "Appropriate",  $b$ = number of experts who ticked the "Item should be slightly revised",  $n$  = number of experts). All items had CVI values of .80 and above, so all were included in the draft test (Davis, 1992). The opinions of 2 master students who were mathematics teachers were also taken about in terms of clarity, suitability to the purpose, suitability to the level of 7th-grade and 8th-grade students. Next, the draft tests were applied to 7th-grade ( $n= 19$ ) students as a pre-application to determine the comprehensibility and language suitability. Considering the opinions of the experts and students, the necessary arrangements were made.

### **2.3.4. Pilot study and scoring the answers**

In the pilot study, draft tests were applied to 391 middle school students and then, the student answers were scored.

**2.3.4.1. Scoring the Answers of PRST.** In the multiple-choice answer tier, the correct answer is 1 point and the wrong answer is 0 point. For scoring the open-ended tier, the rubrics developed by Akkus and Duatepe-Paksu (2006) were edited and used. Since there are different types of problems in the PRST, two rubrics (a rubric for the items of qualitative prediction and comparison, a rubric for the items of missing value and numeric comparison) are used. For open-ended answers, the lowest score is 0 points and the highest score is 3 points. Consequently, in the two-tier PRST, the lowest score is 0 points and the highest score is 4 points. The rubrics items and points are explained in [Table 1](#).

**Table 1.** *Rubrics items.*

Problem Type	Point	Items
Missing value/Numerical Comparison Problems	0	No answer. There is no clue that proportional reasoning exists. The proportion is established between the wrong variables. There is an additive comparison of data. There is a random use of numbers and operations.
	1	Only the correct answer is given, there is not any mathematical operation. There are some clues that proportional reasoning exists.
	2	There is proportional reasoning among the expected variables, but there is a calculation error or the correct answer isn't provided.
	3	There is proportional reasoning to solve the problem correctly, and the correct answer is given.
Qualitative Prediction and Comparison Problems	0	No answer There is no clue that proportional reasoning exists.
	1	There are some clues that proportional reasoning exists.
	2	There is proportional reasoning to solve the problem correctly. The description is done using the root of the problem.
	3	There is proportional reasoning to solve the problem correctly. The correct answer is given with original sentences, and the explanations are enriched with methods such as forming shapes, drawing, giving examples.

**2.3.4.2. Scoring the Answers of N-PRST.** N-PRST consists of two-tier answer options. In multiple-choice answer options, the correct answer is 1 point and the wrong answer is 0 point. The rubric is used to score open-ended answers. Rubric items and points are explained in [Table 2](#).

**Table 2.** *Rubrics items.*

Problem Type	Point	Items
Non-Proportional Reasoning Skills Problems	0	No answer There is proportional reasoning among the variables. Non-proportional situations are indistinguishable. Multiplication strategy is applied to a constant, additive, or linear situation.
	1	There are clues that non-proportional situations are distinguished from proportional situations. There is an additive comparison of data.
	2	Non-proportional situations are distinguished from proportional situations but the explanation is insufficient or made by using the problem root.
	3	Non-proportional situations are distinguished from proportional situations. The correct answer is given by using original sentences, and the explanations are enriched with methods such as forming shapes, drawing, or giving examples.

The open-ended answers of PRTS and N-PRST were scored by the researcher using the rubrics mentioned above. To ensure the interrater reliability, randomly selected 50 students' answers were scored by a second-rater. Spearman rho correlation coefficient was used to determine the relationship between the scores of the two raters. As a result of the analysis, it was seen that the

Spearman rho correlation of items coefficients ranged between .984 and .999. Also, the Wilcoxon Sign Test was used to determine whether there was a significant difference between the points given by the raters. Wilcoxon Sign Test results showed that there was no significant difference between the raters' scorings ( $p > .05$ ).

### 2.3.5. Construct validity, discriminant validity and reliability studies

**2.3.5.1. Proportional Reasoning Skill Test.** In this study, to determine the construct validity of the PRST, Confirmatory Factor Analysis (CFA) was performed. Also, for the construct validity, corrected item-total correlations were calculated and the discrimination of the items was investigated by examining the differences between the lower 27% and upper 27% groups. To determine the reliability of the test, Cronbach alpha and composite reliability coefficients were calculated. Before the data analyses, the data set for 391 cases checked to ensure the normal distribution assumption. For all items, kurtosis and skewness values were calculated. For the items (except for items 8, 13, and 15), kurtosis and skewness values were found to be between  $\pm 2$ . These values indicated that the items had a normal distribution (Cameron, 2004). However, it was determined that item 8 (skewness= 3.061, kurtosis= 10.889) about numerical comparison and item 13 (skewness= 2.806, kurtosis=9.233) and item 15 (skewness= 2.583, kurtosis= 5.829) about missing value didn't have the normal distribution. For this reason, the data related to these items were excluded from the data set and not included in the analysis. For 12 items, z scores were between  $\pm 3.29$  showed the test had univariate normality. According to Mahalanobis distance values, there were no multivariate outlier values and the data set had multivariate normality ( $p < .001$  for the  $\chi^2$ ) (Tabachnick & Fidell, 2013). The correlation matrix for all items was examined and coefficients were found between .30 and .90 for all cases. These values showed that there were not singularity and multicollinearity problems. While anti-image correlation coefficients for each item ( $r = .863$  to  $.929$ ) were adequate for sampling adequacy of individual items (Field, 2009; Tabachnick & Fidell, 2013), results of KMO statistics (KMO=.897) and Bartlett Sphericity Test ( $\chi^2 = 1290.527$ ,  $df = 66$ ,  $p = .000 < .05$ ) proved the sampling adequacy of data set. Also, when comparing the scores of the lower 27% and upper 27% groups ( $n = 106$ ), the normality of each item ( $n = 12$ ) was examined in terms of the group variable. As a result, for all items, it was determined that the skewness and kurtosis values were outside the  $\pm 2$  range. Therefore, differences between groups were investigated by using the Mann-Whitney U test which is one of the nonparametric tests.

To test the discriminant validity of PRST, the relationship between the scores obtained from the PRST and the scores obtained from the N-PRST was examined. For N-PRST scores, skewness was calculated as = 1.334 and kurtosis as= 1.403 while skewness was calculated as = .545, and kurtosis as =-. 675 for the PRST scores. The skewness and kurtosis values showed that data sets were close to the normal distribution (Cameron, 2004). Accordingly, the Pearson Correlation Coefficient was calculated to examine the relationship between the scores. After determining the construct validity of the two-tier test with CFA, the item statistics of the first multiple-choice tier were calculated. Item analysis for the multiple-choice tier was done with the Test Analysis Program (TAP). Item discrimination index, item difficulty index, and item-total correlation were calculated for the construct validity of the multiple-choice tier.

**2.3.5.2. Non-Proportional Reasoning Skill Test.** In this study, to determine the construct validity of the test, the CFA was performed. Also, for the construct validity, corrected item-total correlations were calculated and the discrimination of the items was studied by examining the differences between the lower 27% and upper 27% groups. To determine the reliability of the test, Cronbach alpha internal consistency coefficients were calculated. All items ( $n = 6$ ) have a normal distribution (kurtosis and skewness values found to be between  $\pm 2$  (Cameron, 2004)). z scores (between  $\pm 3.29$ ) and Mahalanobis distance values ( $p < .001$  for the  $\chi^2$ ) showed that the test had univariate and multivariate normality (Tabachnick & Fidell, 2013).

Correlation coefficients for all items were between .30 and .90. Thus, singularity and multicollinearity problems weren't determined. Results of KMO statistics (KMO=.705) and Bartlett Sphericity Test ( $\chi^2= 287.427$ ;  $df= 15$ ;  $p= .00 <.05$ ) showed the sampling adequacy of data set, and anti-image correlation coefficients for each item ( $r= .677$  to  $.757$ ) were adequate for sampling adequacy of individual items (Field, 2009; Tabachnick & Fidell, 2013).

### 3. RESULT / FINDINGS

#### 3.1. Results of Non-Proportional Reasoning Skill Test

To determine the construct validity of the N-PRST, CFA was performed. The data set ( $n=391$ ) was transferred to the LISREL program and a covariance matrix was prepared. It was determined that t values were between 5.71 and 9.49 and significant ( $p<.01$ ) for all values. Then the goodness of fit values and modification suggestions were examined. According to these suggestions, error variances of item1-item2 and item2-item5 were linked. The goodness of fit values for pre-modification and post-modification are shown in Table 3.

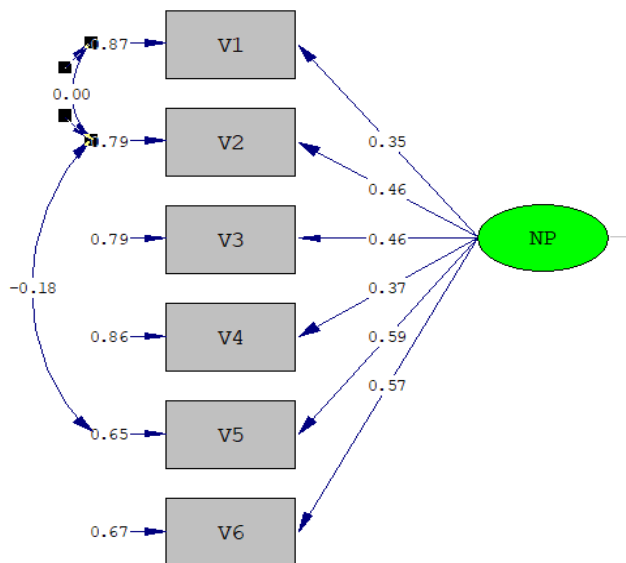
**Table 3.** The goodness of fit values.

The goodness of fit values	Pre-modification	Post-modification
<i>p</i>	.00*	.00*
$\chi^2/df$	52.15/9=5.79	26.06/7=3.72
RMSEA	.111	.08
SRMR	.065	.047
GFI	.96	.98
AGFI	.90	.93
CFI	.85	.93
NFI	.83	.91

\* $p<.05$

As seen in Table 3, after modification goodness of fit values were  $\chi^2/df = 3.72$  (26.06/7), RMSEA = .08, SRMR = .047, GFI = .98, AGFI = .93, CFI = .93, NFI = .91. Figure 1 shows the path diagram for the final model. According to Figure 1, the standardized factor loadings vary between .35 and .59.

**Figure 1.** Path diagram of the model for non-proportional reasoning two-tier test.





**Table 4.** Mann-Whitney *U* test results for the comparison of the upper 27% and lower 27% groups and corrected item-total correlations.

Items	Group	Mean Rank	Sum of Ranks	<i>U</i>	<i>p</i>	Corrected Item-Total Cor.
Item 1	Lower	81.50	10758.00	1980.00	.00*	.563
	Upper	166.82	17683.00			
Item 2	Lower	81.33	10735.00	1957.00	.00*	.555
	Upper	167.04	17706.00			
Item 3	Lower	80.38	10609.50	1831.50	.00*	.543
	Upper	168.22	17831.50			
Item 4	Lower	101.42	13387.50	4609.50	.00*	.381
	Upper	142.01	15053.50			
Item 5	Lower	86.97	11480.00	2702.00	.00*	.503
	Upper	160.01	16961.00			
Item 6	Lower	93.50	12342.00	3564.00	.00*	.434
	Upper	151.88	16099.00			

\**p*<.05

As seen in Table 4, it was determined that there were statistically significant differences between the upper 27% and lower 27% groups for all items, and corrected item-total correlations ranged from .381 and .563. Also, as a result of the reliability analysis, the Cronbach alpha reliability and composite reliability coefficients of the test were .643 and .845.

### 3.2. Results of Two-Tier Proportional Reasoning Skill Test

#### 3.2.1. Construct validity

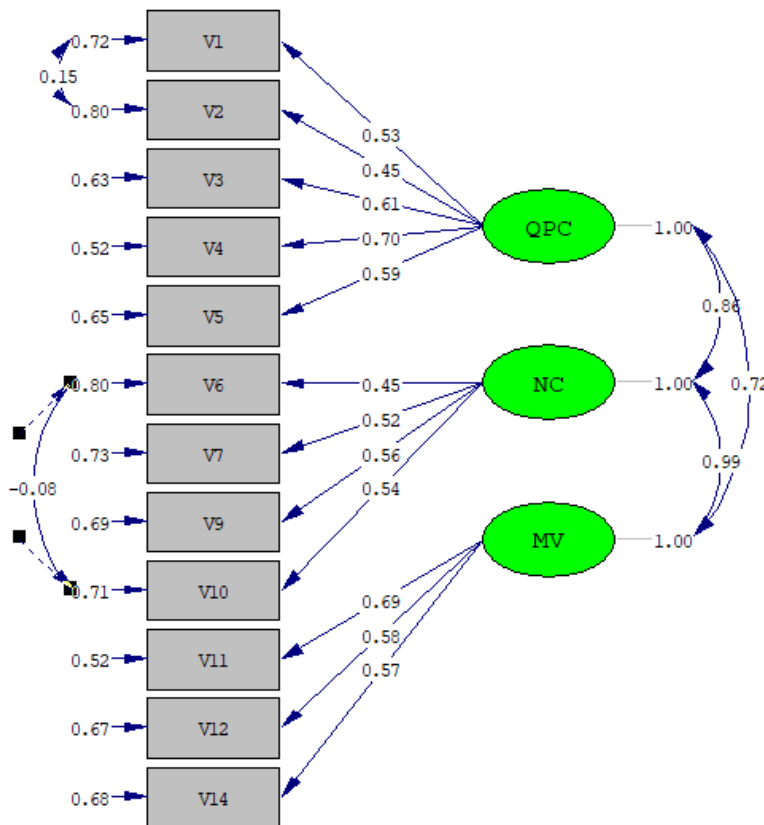
In this study, the CFA was performed to test the 3-factor structure of the PRST. Firstly the data set obtained from 391 students was transferred to the Lisrel program, and a covariance matrix was prepared. For the model, *t* values were between 8.07 and 13.92 and statistically significant (*p*<.01). The CFA analysis computed a significant *p*-value ( $\chi^2 = 96.23$ , *p*= .00013 <.05) for the model. So the goodness of fit values and modification suggestions were examined. According to these suggestions, error variances of item1-item2 and item6-item10 were linked. The goodness of fit values for pre-modification and post-modification are shown in Table 5 and Figure 2 shows the path diagram for the model.

**Table 5.** The goodness of fit values.

The goodness of fit values	Pre-modification	Post-modification
<i>p</i>	.00013*	.00366*
$\chi^2/df$	96.23/51=1.87	79.67/49=1.63
RMSEA	.048	.040
SRMR	.045	.040
GFI	.96	.97
AGFI	.94	.95
CFI	.98	.98
NFI	.96	.96

\**p*<.05

Figure 2. Path diagram of the model for proportional reasoning two-tier test.



As seen in Figure 2, the standardized factor loadings for qualitative prediction and comparison factor varied between .45 and .70, for numerical comparison factor between .45 and .56, and for missing value factor the standardized loadings between .57 and .69. So other goodness of fit values were also examined. As seen in Table 5, goodness of fit values in CFA after modification were  $\chi^2 / df = 1.63$ , RMSEA = .040, SRMR = .040, GFI = .97, AGFI = .95, CFI = .98, NFI = .96.

Mann-Whitney U test results for the comparison of the upper 27% and lower 27% groups and corrected item-total correlation coefficients are presented in Table 6. As seen in Table 6, the corrected-item total correlations ranged from .446 and .592. Also, it was determined that there were statistically significant differences between the upper 27% and lower 27% groups for all items.

**Table 6.** Mann-Whitney *U* test results for the comparison of the upper 27% and lower 27% groups and corrected item-total correlations.

Items	Group	Mean Rank	Sum of Ranks	<i>U</i>	<i>p</i>	Corrected Item-Total Correlation
Item 1	Lower	65.58	7082.50	1196.500	.00*	.504
	Upper	152.62	16788.50			
Item 2	Lower	68.48	7395.50	1509.500	.00*	.446
	Upper	149.78	16475.50			
Item 3	Lower	64.25	6939.50	1053.500	.00*	.569
	Upper	153.92	16931.50			
Item 4	Lower	60.65	6550.00	664.000	.00*	.568
	Upper	157.46	17321.00			
Item 5	Lower	65.91	7118.00	1232.000	.00*	.518
	Upper	152.30	16753.00			
Item 6	Lower	76.15	8224.50	2338.500	.00*	.455
	Upper	142.24	15646.50			
Item 7	Lower	76.34	8244.50	2358.500	.00*	.518
	Upper	142.06	15626.50			
Item 9	Lower	61.71	6664.50	778.500	.00*	.528
	Upper	156.42	17206.50			
Item 10	Lower	79.52	8588.00	2702.000	.00*	.543
	Upper	138.94	15283.00			
Item 11	Lower	73.57	7945.50	2059.500	.00*	.493
	Upper	144.78	15925.50			
Item 12	Lower	62.19	6717.00	831.000	.00*	.592
	Upper	155.95	17154.00			
Item 14	Lower	74.28	8022.00	2136.000	.00*	.526
	Upper	144.08	15849.00			

\* $p < .05$ 

### 3.2.2. Discriminant validity

To test the discriminant validity of PRST, the relationship between the PRST scores and N-PRST scores was examined. As a result of the Pearson Correlation Test, it was determined that the relationship coefficient was  $r = .683$  and statistically significant ( $p = .00 < 0.05$ ).

### 3.2.3. Reliability analysis

According to reliability analysis, Cronbach alpha values were  $\alpha = 0.748$  for the qualitative prediction and comparison factor,  $\alpha = 0.631$  for numerical comparison factor, and  $\alpha = 0.651$  for missing value factor. For the total, the Cronbach alpha reliability coefficient was calculated as  $\alpha = 0.849$ . Also, the composite reliability coefficient of the test was 0.656.

### 3.2.4. Test Statistics

The test statistics of PRST are shown in Table 7.

**Table 7.** Test statistics of PRST.

Factor	Mean	Standard Deviation
QPC	3.03	1.83
NV	1.47	1.35
MV	1.58	1.62
Total	2.15	1.40

As seen in Table 7, it could be said the students' PRST level was medium.

### 3.3. Construct Validity and Reliability of Multiple-Choice Test

Item analysis and test statistics results of the multiple-choice test are shown in [Table 8](#) and [Table 9](#).

**Table 8.** *Item analysis results.*

Item No	Item Discrimination Index ( <i>r</i> )	Item Difficulty Index ( <i>p</i> )	Point Biserial
01	.64	.44	.56
02	.64	.35	.57
03	.71	.37	.62
04	.77	.45	.63
05	.55	.33	.55
06	.37	.15	.53
07	.48	.19	.57
09	.62	.25	.53
10	.62	.30	.62
11	.45	.16	.62
12	.35	.13	.58
14	.31	.13	.50

**Table 9.** *Test statistics.*

Statistics	Value
Mean	2.951
Standard Deviation	2.951
Mean Item Difficulty	.278
Mean Discrimination Index	.545
KR-20	.810
Mean Point Biserial	.573

The item discrimination indices ranged between .31 and .77, the item difficulty indices ranged between .13 and .45, and the point biserial coefficients ranged between .50 and .63. According to test statistics, test discrimination was very good (Ebel, 1965; Wells & Wollack, 2003), but the test was difficult (Crocker & Algina, 2008). Also, the KR-20 value showed the multiple-choice test had good reliability (Kline, 2011; Rudner & Schafer, 2002; Wells & Wollack, 2003).

### 4. DISCUSSION and CONCLUSION

This research aims to develop a useful, valid, and reliable two-tier PRST for middle school 7th and 8th-grade students. The test includes problems that measure qualitative prediction and comparison, missing value, and numerical comparison (Cramer & Post, 1993). The study was designed with the sequential exploratory mixed-method approach, in which qualitative and quantitative research methods were used. Firstly, a problem pool consisting of 15 problems (5 of which was qualitative prediction and comparison, 5 of which was missing value, and 5 of which was numerical comparison) was prepared. The first tier consisted of a multiple-choice problem, with four choices. The second tier was the open-ended answer part, which included explaining and justifying the answer given multiple-choice tier. The two rubrics (a rubric for the qualitative prediction and comparison problems, and a rubric for the missing value and numeric comparison problems) were used for scoring the open-ended tier. The rubrics items

were adapted from the study of Akkus and Duatepe-Paksu (2006). In the two-tier test, the lowest score was 0 and the highest score was 4 points for each problem.

In this study, the face, content, construct, discriminant validity of the PRST were tested. The content and face validity of the test were provided with assessment and evaluation ( $n=1$ ) and mathematics education ( $n=2$ ) experts' opinions (Gable, 1986). For the construct validity studies of the test, firstly CFA was performed. According to Mueller and Hancock (2001), the main advantage of CFA is that it enables researchers to bridge between theory and observation. CFA facilitates testing the relationship between the latent constructs (QPC, NC, MV) and observed variables (Suhr, 2006).

In this study, CFA was carried out to test a 3-factor structure with the data set consisting of 12 problems which ensured the univariate and multivariate normal distribution assumptions. In the factor analysis, the standardized factor loadings values of .40 and above are accepted as meaningful load values (Gable, 1986; Hatcher, 1994). Hair et al. (2014) expressed that factor loadings values of .30 and above are accepted practically significant for sample sizes of 350 or greater. Based on CFA results, it could be said that the standardized factor loadings were meaningful. According to commonly agreed goodness of fit values criteria (e.g., Brown, 2006; Hair et al., 2014; Hu & Bentler, 1999; Tabachnick & Fidell, 2013),  $\chi^2 / df < 2$ , RMSEA, SRMR  $< .05$ , GFI, AGFI, CFI, NFI  $> .90$  values are acceptable values,  $\chi^2 / df < 5$ , RMSEA, SRMR  $< .08$ , GFI, AGFI, CFI, NFI  $> .95$  values are excellent values. Accordingly, when the values calculated as a result of the analysis are taken into consideration, it could be said that the  $\chi^2 / df$ , RMSEA, SRMR, GFI, AGFI, CFI, NFI are excellent values. Also, Mann-Whitney U test results for the comparison of the upper 27% and lower 27% groups scores and corrected item-total correlations showed that the problems tended to measure the same skill and discrimination indexes were high (Büyüköztürk, 2010). According to these results, it could be said that the two-tier PRST has construct validity. Also, item analysis results show, multiple-choice tier has very good discrimination (Ebel, 1965; Wells & Wollack, 2003) and good reliability (Kline, 2011; Rudner & Schafer, 2002; Wells & Wollack, 2003). But according to mean item difficulty, the multiple-choice test is difficult (Crocker & Algina, 2008).

To determine the discrimination of the test, the Pearson Correlation Coefficient was calculated between the PRST scores and the N-PRST score. The N-PRST was developed in this study to determine the PRST's discriminant validity. The two-tier N-PRST consisted 6 problems: additive situations ( $n=2$ ), linear situations ( $n=2$ ), and constant situations ( $n=2$ ). The validity and reliability studies showed that N-PRST was useful, valid (according to results of CFA, Mann-Whitney U test results for the comparison of the upper 27% and lower 27% groups and corrected item-total correlations), and reliable. Cohen (1988) interpreted the strength of relationship as; .10-.29 "small", .30-.49 "medium" and .50-1.0 "large". As a result of the analysis, it was determined that there was a statistically significant large relationship between proportional and non-proportional test scores. The positive relationship could be considered as evidence that students can distinguish between proportional and non-proportional situations, and the discrimination of the PRST is high. According to these results, it could be said that the discriminant validity of the test was ensured (Fornell & Larcker, 1981).

Kline (2011) states that the reliability coefficient is excellent if it is around .90, very good around .80, sufficient around .70, and insufficient under .50. Rudner and Schafer (2002) stated that the 0.50 or 0.60 reliability coefficient for the tests performed in the classroom can be seen as sufficient. Ebel and Frisbie (1991) stressed that when it comes to ratings for a group of people, such as the classroom, the reliability coefficient should be 0.65. The Cronbach Alpha coefficient of the test is very good and the composite reliability coefficient is acceptable.

As a conclusion, it can be said that the two-tier PRST is useful, valid, and reliable to measure middle school students' proportional reasoning skills. The psychometric properties of the test

developed in this study are tested using the data obtained from the middle school 7th and 8th-grade students. Researchers can study further the psychometric properties of the PRST for students who graduated from middle school. It is expected that this valid and reliable PRST will encourage other researchers to study the relationships of proportional reasoning skills with different variables (e.g., academic achievement or attitude on ratio-proportion). This research was conducted with students selected by convenience sampling. Although the test is determined as valid and reliable, to increase the generalizability of the test, it is recommended to examine the psychometric properties of the test with a group determined by random sampling.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

### Authorship contribution statement

**Kübra Açıkgül:** Investigation, Resources, Introduction, Methodology, Software, Analysis, Findings, Discussion, Supervision, and Validation, Writing original draft.

### ORCID

Kubra ACIKGUL  <https://orcid.org/0000-0003-2656-8916>

## 5. REFERENCES

- Afnia, P.N., & Istiyono, E. (2020, February). The development of two-tier multiple choice instruments to measure higher order thinking skills bloomian. In *3rd International Conference on Learning Innovation and Quality Education (ICLIQE 2019)* (pp. 1038-1045). Atlantis Press. <https://doi.org/10.2991/assehr.k.200129.128>
- Akkus, O., & Duatepe-Paksu, A. (2006). Construction of a proportional reasoning test and its rubrics. *Eurasian Journal of Educational Research*, 25, 1-10.
- Alfieri, L., Higashi, R., Shoop, R., & Schunn, C. D. (2015). Case studies of a robot-based game to shape interests and hone proportional reasoning skills. *International Journal of STEM Education*, 2(4), 1-13. <https://doi.org/10.1186/s40594-015-0017-9>
- Allain, A. (2000). *Development of an instrument to measure proportional reasoning among fast-track middle school students*. [Master's thesis]. University of North Carolina State.
- Arıcan, M. (2019). A diagnostic assessment to middle school students' proportional reasoning. *Turkish Journal of Education*, 8(4), 237-257. <https://doi.org/10.19128/turje.522839>
- Ayan, R., & Isiksal-Bostan, M. (2019). Middle school students' proportional reasoning in real life contexts in the domain of geometry and measurement. *International Journal of Mathematical Education in Science and Technology*, 50(1), 65-81. <https://doi.org/10.1080/0020739X.2018.1468042>
- Behr, M., Lesh, R., & Post, T. (1988). Proportional reasoning, In M. Behr and J. Hiebert (Eds.), *Number concepts and operations in the middle grades*. Lawrence Erlbaum Associates.
- Behr, M., Harel, G., Post, T., & Lesh, R. (1992). Rational number, ratio and proportion. In D. Grouws (Eds.), *Handbook on research of teaching and learning* (pp. 296-333). McMillan.
- Brown, T. A. (2006). Confirmatory factor analysis for applied research. In David A. Kenny (Eds.), *Methodology in the Social Sciences*. The Guilford Press.
- Bright, G. W., Joyner, J. M., & Wallis, C. (2003). Assessing proportional thinking. *Mathematics Teaching in the Middle School*, 9(3), 166-172. <https://www.jstor.org/stable/41181882>
- Burton, S. J., Sudweeks, R. E., Merrill, P. F., & Wood, B. (1991). *How to prepare better multiple-choice test items: Guidelines for university faculty*. Brigham Young University



- Testing Services and The Department of Instructional Science. <https://testing.byu.edu/handbooks/betteritems.pdf>
- Büyüköztürk, S. (2010). *Sosyal bilimler için veri analizi el kitabı* [Data analysis handbook for social sciences]. Pegem Akademi.
- Cameron, A. (2004). Kurtosis. In M. Lewis-Beck, A. Bryman and T. Liao (Eds.). *Encyclopedia of social science research methods*. (pp. 544-545). SAGE Publications, Inc.
- Common Core State Standards Initiative (US). Common core state standards for mathematics. <http://www.corestandards.org/Math/>
- Cramer, K., & Post, T. (1993). Proportional reasoning. *The Mathematics Teacher*, 86(5), 404-407. <https://www.jstor.org/stable/27968390>
- Cramer, K., Post, T., & Currier, S. (1993). Learning and teaching ratio and proportion: research implications. In D. Owens (Eds.), *Research ideas for the classroom* (pp. 159-178). Macmillan Publishing Company.
- Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.
- Cohen, L., Manion, L., & Morrison, K. (2013). *Research methods in education*. Routledge.
- Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research*. Sage Publications.
- Davis, L. L. (1992). Instrument review: getting the most from a panel of experts. *Applied Nursing Research*, 5, 194-197. [https://doi.org/10.1016/S0897-1897\(05\)80008-4](https://doi.org/10.1016/S0897-1897(05)80008-4)
- Dinç-Artut, P., & Pelen, M. S. (2015). 6th grade students' solution strategies on proportional reasoning problems. *Procedia-Social and Behavioral Sciences*, 197, 113-119. <https://doi.org/10.1016/j.sbspro.2015.07.066>
- Duatepe, A., Akkus-Cıkla, O., & Kayhan, M. (2005). An investigation on students' solution strategies for different proportional reasoning items. *Hacettepe Journal of Education Faculty*, 28, 73-81. <https://dergipark.org.tr/en/pub/hunefd/issue/7808/102422>
- Ebel, R. L. (1965). *Measuring educational achievement*. Prentice Hall.
- Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*. Prentice Hall.
- Field, A. (2009). *Discovering statistics using SPSS*. Sage Publication.
- Fornell, C., & Larcker, D. F. (1981). Structural equation models with unobservable variables and measurement error: Algebra and statistics. *Journal of Marketing Research*, 18(3), 328-388. <https://doi.org/10.1177/002224378101800313>
- Gable, R. K. (1986). *Instrument development in the affective domain*. Kluwer-Nijhoff Publishing.
- Hair, J. F., Jr., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2014). *Multivariate data analysis*. Pearson New International Edition.
- Haja, S., & Clarke, D. (2011). Middle school students' responses to two-tier tasks. *Mathematics Education Research Journal*, 23(1), 67-76. <https://doi.org/10.1007/s13394-011-0004-5>
- Haslam, F., & Treagust, D. F. (1987). Diagnosing secondary students' misconceptions of photosynthesis and respiration in plants using a two-tier multiple choice instrument. *Journal of Biological Education*, 21(3), 203-211. <https://doi.org/10.1080/00219266.1987.9654897>
- Hatcher, L. (1994). *A step-by-step approach to using the SAS® system for Factor Analysis and Structural Equation Modeling*. SAS Institute, Inc.
- Hilton, A., Hilton, G., Dole, S., & Goos, M. (2013). Development and application of a two-tier diagnostic instrument to assess middle - years students' proportional reasoning. *Mathematics Education Research Journal*, 25(4), 523-545. <https://doi.org/10.1007/s13394-013-0083-6>

- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55. <https://doi.org/10.1080/10705519909540118>
- Hyman, M. R., & Sierra, J. J. (2016). Open-versus close-ended survey problems. *Business Outlook*, 14(2), 1-5.
- Kline, R. B. (2011). *Principles and practice of structural equation modeling*. Guilford Press.
- Lamon, S. J. (2007). Rational numbers and proportional reasoning: Toward a theoretical framework for research. In F. K. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 629–668). Information Age Publishing.
- Lawton, C. A. (1993). Contextual factors affecting errors in proportional reasoning. *Journal for Research in Mathematics Education*, 24(5), 460-466. <https://doi.org/10.2307/749154>
- Lesh, R., Post, T., & Behr, M. (1988). Proportional reasoning. In J. Hiebert & M. Behr (Eds.), *Number concepts and operations in the middle grades* (pp. 93-118). Lawrence Erlbaum & National Council of Teachers of Mathematics.
- Lim, K. (2009). Burning the candle at just one end: Using nonproportional examples helps students determine when proportional strategies apply. *Mathematics Teaching in the Middle School*, 14(8), 492–500. <https://doi.org/10.5951/MTMS.14.8.0492>
- Mersin, N. (2018). An evaluation of proportional reasoning of middle school 5th, 6th and 7th grade students according to a two-tier diagnostic test. *Cumhuriyet International Journal of Education*, 7(4), 319–348. <https://doi.org/10.30703/cije.4266271>
- Ministry of National Education [MoNE], (2018). Matematik dersi öğretim programı. (İlkokul ve Ortaokul 1, 2, 3, 4, 5, 6, 7 ve 8. Sınıflar) [Mathematics curriculum. (Primary and Middle Schools 1, 2, 3, 4, 5, 6, 7 and 8th Grades) [Middle school mathematics curricula for grades 5, 6, 7, and 8]]. <https://mufredat.meb.gov.tr/ProgramDetay.aspx?PID=329>
- Mueller, R. O., & Hancock, G. R. (2001). Factor analysis and latent structure: Confirmatory factor analysis. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social and Behavioral Sciences* (pp. 5239-5244). Pergamon.
- National Council of Teachers of Mathematics (NCTM) (2000). *Principles and Standards for School Mathematics*. National Council of Teachers of Mathematics.
- Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended problems. *Canadian Journal of Experimental Psychology*, 67(3), 215-227. <https://doi.org/10.1037/a0032918>
- Özgün-Koca, S. A., & Altay, M. K. (2009). An investigation of proportional reasoning skills of middle school students. *Investigations in Mathematics Learning*, 2(1), 26-48. <https://doi.org/10.1080/24727466.2009.11790289>
- Pelen, M. S., & Dinç-Artut, P. (2015). 7th grade students' problem solving success rates on proportional reasoning problems. *The Eurasia Proceedings of Educational and Social Sciences*, 2, 96-100. <https://doi.org/10.21890/ijres.71245>
- Peterson, R. F., Treagust, D. F., & Garnett, P. (1986). Identification of secondary students' misconceptions of covalent bonding and structure concepts using a diagnostic test instrument. *Research in Science Education*, 16, 40-48. <https://doi.org/10.1007/BF02356816>
- Post, T. R., Behr, M. J., & Lesh, R. (1988). Proportionality and the development of pre-algebra understandings. In A. Coxford & A. Shulte (Eds.), *The ideas of algebra, K-12* (pp. 78-90). National Council of Teachers of Mathematics.
- Reja, U., Manfreda, K. L., Hlebec, V., & Vehovar, V. (2003). Open-ended vs. close-ended questions in web questionnaires. In A. Ferligoj & A. Mrvar (Eds.), *Developments in applied statistics* (pp. 159–177). FDV.
- Rudner, L. M., & Shafer, W. D. (2002). *What teachers need to know about assessment*. National Education Association.

- Singh, P. (2000). Understanding the concepts of proportion and ratio constructed by two grade six students. *Educational Studies in Mathematics*, 43, 271-292. <https://doi.org/10.1023/A:1011976904850>
- Soyak, O., & Isiksal, M. (2017, February). Middle school students' difficulties in proportional reasoning. Paper Presented at the *CERME 10*, Dublin, Ireland.
- Suhr, D. (2006). Exploratory or confirmatory factor analysis. *SAS Users Group International Conference* (pp. 1- 17). SAS Institute, Inc.
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics*. Pearson.
- Tamir, P. (1989). Some issues related to the use of justifications to multiple-choice answers. *Journal of Biological Education*, 23(4), 285-292. <https://doi.org/10.1080/00219266.1989.9655083>
- Tamir, P. (1990). Justifying the selection of answers in multiple choice items. *International Journal of Science Education*, 12(5), 563-573. <https://doi.org/10.1080/0950069900120508>
- Tourniaire, F., & Pulos, S. (1985). Proportional reasoning: A review of the literature. *Educational Studies in Mathematics*, 16(2), 181-204. <https://doi.org/10.1007/BF02400937>
- Tsui, C. Y., & Treagust, D. (2010). Evaluating secondary students' scientific reasoning in genetics using a two-tier diagnostic instrument. *International Journal of Science Education*, 32(8), 1073-1098. <https://doi.org/10.1080/09500690902951429>
- Van De Walle, J. A., Karp, K. S., & Bay-Williams, J. M. (2013). *Elementary and middle school mathematics: teaching developmentally*. Pearson Education, Inc.
- Van Dooren, W., De Bock, D., Hessels, A., Janssens, D., & Verschaffel, L. (2005). Not everything is proportional: Effects of age and problem type on propensities for over generalization. *Cognition and Instruction*, 23(1), 57-86. [https://doi.org/10.1207/s1532690xci2301\\_3](https://doi.org/10.1207/s1532690xci2301_3)
- Van Dooren, W., De Bock, D., & Verschaffel, L. (2010). From addition to multiplication and back: The development of students' additive and multiplicative reasoning skills. *Cognition and Instruction*, 28, 360–381. <https://doi.org/10.1080/07370008.2010.488306>
- Weinberg, S. L. (2002). Proportional reasoning: One problem, many solutions! In B. Litwiler (Eds.), *Making sense of fractions, ratios, and proportions: 2002 year book* (pp. 138-144). National Council of Teachers of Mathematics.
- Wells, C. S., & Wollack, J. A. (2003). *An instructor's guide to understanding test reliability*. Testing & Evaluation Services. University of Wisconsin. <http://testing.wisc.edu/Reliability.pdf>.

## Detecting Differential Item Functioning: Item Response Theory Methods Versus the Mantel-Haenszel Procedure

Emily Diaz <sup>1,\*</sup>, Gordon Brooks <sup>2</sup>, George Johanson <sup>3</sup>

<sup>1</sup>Westat, Senior Research Associate, Education Studies

<sup>2</sup>Ohio University, Department of Educational Studies

<sup>3</sup>Ohio University, Department of Educational Studies

### ARTICLE HISTORY

Received: Apr. 30, 2020

Revised: Mar. 11, 2021

Accepted: Apr. 04, 2021

### Keywords:

Differential item functioning,  
Monte Carlo simulation,  
Type I error rate,  
Mantel-Haenszel

**Abstract:** This Monte Carlo study assessed Type I error in differential item functioning analyses using Lord's chi-square (LC), Likelihood Ratio Test (LRT), and Mantel-Haenszel (MH) procedure. Two research interests were investigated: item response theory (IRT) model specification in LC and the LRT and continuity correction in the MH procedure. This study enhances the literature by investigating LC and the LRT using correct and incorrect model-data fit and comparing those results to the MH procedure. There were three fixed factors (number of test items, IRT parameter estimation method, and item parameter equating) and four varied factors (IRT model used to generate data and fit the data, sample size, and impact). The findings suggested the MH procedure without the continuity correction is best based on Type I error rate.

## 1. INTRODUCTION

In the field of psychometrics, item bias and test fairness are important issues that must be addressed (Kane, 2013). Item bias, differential item functioning (DIF), and impact are related but not synonymous (Zumbo, 1999). Item impact occurs when groups simply differ in performance on an item; when impact persists after controlling for overall skill on the construct being measured DIF is present; bias is pernicious DIF. Thus, DIF is the key to identifying possibly biased items.

Statistical tests of DIF are prone to both false positives (Type I errors) and false negatives (Type II errors). Roussos and Stout (1996) presented three reasons to research Type I error rates of DIF methods. First, removing a non-DIF item, or making a Type I error, unnecessarily wastes resources. Second, false positives explain why some testing organizations can neither understand nor ascertain the source of DIF in certain items. Finally, highly discriminating items can be mistakenly flagged for DIF (Li et al., 2012). Items with high discrimination indices contain higher information indices and are better able to discern differences between examinees

---

\*CONTACT: Emily DIAZ ✉ [emilydiaz@westat.com](mailto:emilydiaz@westat.com) 📄 Senior Research Associate, Education Studies

with higher and lower levels of the underlying latent trait. Hence, false positives for these items are especially problematic and should not be needlessly removed.

### 1.1. Description of DIF Methods

According to Camilli and Shepard (1994) there are three theoretical reasons to prefer item response theory (IRT) methods over classical test theory (CTT) methods for DIF detection: item parameter estimates derived from IRT are less confounded and influenced with sample specific characteristics; IRT provides more accurate statistical properties of items than CTT to ascertain where the item functions differently (i.e., difficulty, discrimination, or pseudo-guessing); finally, the item characteristic curve (ICC) for each group can be graphed enhancing the understanding of items displaying DIF. According to Thissen et al. (1983) another advantage of IRT over CTT is that the fit between the data and the IRT model can be assessed statistically.

Lord's chi-square (LC) compares the performance of two groups on an item by examining item parameter differences depending on the specified IRT model (Lord, 1980). The group that is hypothesized to be favored, or have a higher probability of getting the item right, is the reference group (Camilli & Shepard, 1994; de Ayala, 2009). The group that is hypothesized to be disadvantaged, or have a lower probability of getting the item right, is the focal group (Camilli & Shepard, 1994; de Ayala, 2009). For LC, the item parameters are estimated separately for each group and are not directly comparable. Therefore, they need to be equated before meaningful comparisons can be made (Rupp & Zumbo, 2006; Stocking & Lord, 1983). LC follows a  $\chi^2$  distribution with degrees of freedom equal to the number of estimated parameters based on the IRT model implemented. Theoretically, LC is analogous to testing the equality of ICCs between the reference and focal groups. When the probability difference of getting an item right between the reference and focal groups is systematically the same across all ability levels, the item displays uniform DIF. Graphically, item characteristic curves for the groups are parallel (Camilli & Shepard, 1994). Non-uniform DIF occurs when the item favors one group over another for certain ability levels but reverses for other ability levels. Graphically, the item characteristic curves are not parallel. A benefit of using LC is that it can detect both uniform and non-uniform DIF.

The likelihood ratio test (LRT) assesses whether allowing the parameters for the studied item to vary across groups significantly improves the fit of the model. If so, then the studied item displays DIF. Judgments concerning fit are based on a comparison of the compact and augmented models. In the augmented model, an IRT model is fit such that all the item parameters are the same for the two groups except for the one item being studied, which varies across groups. In the compact model, the same IRT model specified in the augmented model is fit to the data such that all item parameters including the studied item are constrained to be the same in both groups (Thissen et al., 1988). The LRT test statistic is computed by  $G^2 = -2LL_c - (-2LL_A)$  where  $-2LL_c$  and  $-2LL_A$  denote the negative two log-likelihood ( $-2LL$ ) of the compact and augmented models, respectively. The test statistic is compared to a  $\chi^2$  distribution with degrees of freedom equal to the number of estimated item parameters. An advantage of the LRT over LC is that item parameters are estimated together for both groups and do not need to be equated. However, a disadvantage is that the procedure takes a long time to implement because  $n + 1$  models must be assessed for an  $n$  item test (Thissen et al., 1988). From a theoretical perspective, due to the asymptotic nature of the test statistic, the LRT and LC should yield the same conclusions provided the sample size is large (Millsap & Everson, 1993; Thissen et al., 1993). This study adds to the literature by assessing this claim.

The Mantel-Haenszel (MH) procedure examines the relationship between item performance and group membership after taking into account total test performance (Dorans & Holland, 1993). This method examines whether item responses are independent of group membership



after controlling for observed score. The MH test statistic is compared to a  $\chi^2$  distribution with one degree of freedom and tests if the odds of members of the focal group getting the item right are the same as the odds of the reference group (Dorans & Holland, 1993). The MH statistic has been widely accepted because it is relatively easy to understand and implement, provides a  $\chi^2$  statistical significance test, and uses the odds-ratio as an effect size measure (Holland & Thayer, 1988; Millsap & Everson, 1993). Furthermore, an IRT model does not need to be fit to the data and the procedure does not require large sample sizes (Raju et al., 1993). One disadvantage of the MH is that it was designed to primarily detect uniform DIF (de Ayala, 2009; Millsap & Everson, 1993). However, in some cases MH can detect non-uniform DIF (Marañón et al., 1997; Mazor et al., 1994). Narayanan and Swaminathan (1996) note, that the MH procedure is ineffective in detecting non-uniform DIF that is also not ordinal.

## 1.2. Purposes of the Study

It is important to study DIF because certain measurement techniques require DIF analyses as a prerequisite (Shepard et al., 1985). For example, equating and test adaption are measurement approaches; that allow researchers to compare group estimates (i.e., item and/or person parameters) across separate test administrations, test forms, or groups (Cook & Eignor, 1991). When equating or adapting, truly biased items should not be present because these items are not measuring the concept similarly across groups. Hence, these items are uninformative and in fact can harm results (Kim & Cohen, 1992; Shepard et al., 1985).

Another important reason for studying DIF is that it addresses the validity of test score use because without it a test score is meaningless. In the United States the 1999 *Standards* (American Educational Research Association et al., 1999) called attention to test validity, which assesses whether a test is accurately measuring what it was designed to measure. According to the National Research Council (2007) in order to evaluate the trustworthiness and accuracy of score-based decisions testing companies must provide two types of evidence: the degree to which stated outcomes and purposes are achieved (i.e., intended effects) and the presence, or lack thereof, of adverse impact across groups of examinees. Furthermore, one particular type of evidence for validity is construct validity or the degree to which a test score is an accurate measure of the underlying latent variable it purports to measure (Creswell, 2009). According to Messick (1995) the value implications, interpretations, and meanings resulting from a test scores are a consequential aspect of construct validity. That is, when test scores are used in applied settings such as performance assessment, certification exam, licensure, course placement, college admittance, subject mastery and so forth there needs to be evidence of construct validity (Kane, 2009; Messick, 1995). In particular, DIF analyses statistically assess a potential threat to construct validity at the item level (Camilli, 2006; Kane, 2013).

When assessing DIF, there is a disparity between textbook presentations of IRT DIF methods and their frequency of use not only in practice but also in the Monte Carlo (MC) literature. IRT methods have a theoretical superiority to detect DIF (Camilli & Shepard, 1994; Thissen et al., 1993), yet they may not be as widely implemented in the simulation or MC literature on DIF as the MH and logistic regression procedures (Narayanan & Swaminathan, 1996). Raju (1990) commented that

regardless of a particular investigator's decision for a given study, there is certainly a need for monte carlo [sic] and empirical studies to assess the degree of robustness and uniformity of item bias results obtained with the likelihood ratio,  $\chi^2$ , and area procedures (p. 206).

This sentiment was again echoed by Raju et al. (1993) who stated that

because this study was based on an empirical data set, it was not possible to know how many items were truly biased. There is obviously a need for a comprehensive Monte



Carlo investigation to determine . . . the behavior of the IRT based methods with respects to false positives and false negatives (p. 310).

Despite these early calls, a recent (1/27/2021) search of the literature on *scholar.google.com* using the exact phrase terms “item response theory” and “differential item functioning” in the title returned 110 results, the majority of which focused on specific applications or software. When “Type-I” was added, there was only one citation, a dissertation concerning missing data. Another search using the terms “misspecification” and “item response theory” in the title returned only 2 results, neither related to DIF. The existing MC studies of DIF which have examined IRT and non-IRT DIF methods offer varied and sometimes conflicting research recommendations (Cohen et al., 1996; Cohen & Kim, 1993; DeMars, 2010; Kim et al., 1994; Herrera & Gómez, 2008; Lautenschlager & Park, 1988; Li et al., 2012; Lim & Drasgow, 1990; McLaughlin & Drasgow, 1987; Paek, 2010; Rudner et al., 1980; Sari & Huggins, 2015; Shepard et al., 1985; Wang & Yeh, 2003; Wells et al., 2009). Therefore, there still remain unknown aspects regarding these DIF methods such as IRT model fit, IRT model specification and misspecification, sample size, item discrimination variability, and item impact, which are addressed in this study and fill in the gap identified by Raju et al. (1993).

The main purpose of this study was to investigate and compare Type I error rates of DIF detection using LC, the LRT, and the MH procedures. Using multiple DIF methods, a form of psychometric triangulation, is a useful approach to investigate DIF in practice because each DIF detection method has different strengths and this adds to the research literature by allowing for comparisons to be made across DIF methods. Type I error was evaluated based on Bradley's (1978) stringent criterion interval  $[0.045, 0.055]$ , which is equivalent to  $\alpha \pm 0.1\alpha$ , when  $\alpha = .05$ . Within the main purpose, two additional research interests guided this study: (1) the role of correct or incorrect IRT model specification in LC and the LRT, which was addressed using two simulations and (2) the role of the continuity correction in the MH procedure, which was addressed using one simulation. This MC study will add to and clarify the existing literature by determining the importance of correctly or incorrectly choosing the IRT model when computing LC and the LRT and comparing those results to the MH procedure. Correct and incorrect IRT model specification was added to enrich this study by providing guidance and recommendations to not only applied researchers but also to evaluators. In applied research determining the true and best IRT model to select when using LC and the LRT for a given dataset is never deterministically known (as it is in MC research) but is statistically assessed. Hence, these findings are useful to theoretical and applied researchers.

### 1.3. Variables in Monte Carlo DIF Studies

In the present study, the number of test items, IRT parameter estimation method, and item parameter equating were fixed while the IRT model used to generate data, IRT model used to fit the data, sample size, and impact varied based on the existing literature. For each DIF method, the MC literature surrounding the relationship between these variables of interest and the DIF method is discussed.

#### 1.3.1. Number of test items

To obtain an accurate measure of ability for an individual, tests should include a sufficient number of items (Rogers & Swaminathan, 1993). From an IRT perspective, the minimum number of items needed to ensure accurate sampling error for item discrimination in the three-parameter logistic (3PL) model is 50 (Lord, 1968). From a CTT perspective, as test length increases standard error decreases resulting in a more accurate measurement of an examinees' ability using the observed score (Rogers & Swaminathan, 1993).

For much of the MC literature on DIF, the number of test items varied from 20 to 60. For LC, Wells et al. (2009) found test length did not influence Type I error rate while Cohen and Kim

(1993) found that Type I error was inflated for a 20 item test. For the LRT, Finch (2005) found an inconsistent relationship between Type I error and test length. For the MH procedure, Finch (2005) simulated test lengths of 20 and 50 items finding Type I error was consistently conservative but closer to the nominal level with the longer test. DeMars (2009) simulated three test lengths of 20, 40 and 60 items finding shorter tests led to higher Type I error rates for MH. Paek (2010) found Type I error for MH with the continuity correction was consistently conservative regardless of the number of test items. Similarly, Paek and Wilson (2011) found Type I error was consistently conservative regardless of the number of test items.

### **1.3.2. Item parameter equating**

Item parameter equating is needed for LC only. The literature did not provide sufficient evidence to discern how item parameter equating influenced Type I error rate for two reasons. First, several studies fixed the item parameter equating technique (Kim & Cohen, 1995; Kim et al., 1994; Wells et al., 2009). Second, Candell and Drasgow (1988) investigate two equating methods but simulated impact in every cell of their design so their results could be confounded by impact. They used the weighted mean and sigma method developed by (Linn et al., 1981) and the test characteristic method of equating (Stocking & Lord, 1983) finding Type I error rate was consistently higher with this method.

### **1.3.3. Sample size**

Sample size has varied widely in the MC DIF literature, ranging from 250 to 20,000. To obtain an accurate measure of the item parameters for both the 1PL and 2PL a sample size of 500 is adequate (Holland & Wainer, 1993; Sari & Huggins, 2015). For all three DIF methods the research literature suggests different and sometimes conflicting findings regarding the relationship between sample size and Type I error. For LC, there was no difference in Type I error across sample size (Wells et al., 2009). Kim et al. (1994) found more accurate results with larger sample sizes, while other studies found more accurate results with smaller sample sizes (Lim & Drasgow, 1990; McLaughlin & Drasgow, 1987). Finally, two studies did not find consistent results (Candell & Drasgow, 1988; Kim & Cohen, 1992).

For the LRT, Cohen et al. (1996) generated sample sizes of 250 and 1,000 while Stark et al. (2006) generated sample sizes of 500 and 1,000 both finding no marked differences in Type I error rates across sample sizes. Two additional studies found Type I error rate depended on whether group sample size was balanced or not. Finch (2005) found Type I error depended on the number of test items and group ability difference for the balanced condition. For the unbalanced condition Type I error was within the nominal level. Finch and French (2007) found with balanced group sample sizes of 250 and unbalanced group sample size results were within the nominal level. With balanced group sample size of 500, Type I error was conservative. However, with unbalanced sample size Type I error was inflated.

For the MH procedure, Narayanan and Swaminathan (1996) found Type I error was maintained across sample size. Other research has found conflicting results. Some studies have shown that larger sample sizes led to Type I error inflation (DeMars, 2009; Herrera & Gómez, 2008; Roussos & Stout, 1996) while other studies have shown both smaller and larger sample sizes resulted in conservative Type I error (Finch, 2005; Güler & Penfield, 2009; Herrera & Gómez, 2008; Paek 2010; Paek & Wilson, 2011; Rogers & Swaminathan, 1993). Due to the conflicting evidence and the need to obtain accurate item parameter estimates for the IRT DIF methods, the present study generated sample sizes of 500 and 1,000.

### **1.3.4. Model misspecification**

The MC research literature was sparse concerning how IRT model selection affected Type I error rates. Most MC studies generated data and analyzed DIF based on fitting the true

underlying IRT model (e.g., Candell & Drasgow, 1988; Cohen & Kim, 1993; DeMars, 2009; Finch & French, 2007; Wang & Yeh, 2003). For LC, Lautenschlager and Park (1988) addressed model misspecification but the findings were not applicable to the current study because they generated multidimensional ability values (Lautenschlager & Park, 1988). For the LRT, Bolt (2002) addressed model misspecification for polytomous item response data finding that the IRT model selected impacted Type I error rate especially in the presence of impact.

The MH procedure uses the observed score to match the groups on ability so the underlying IRT model used to simulate data matters because the observed score is only a sufficient statistic for ability,  $\theta$ , when the data follow the Rasch model and the one-parameter logistic (1PL) model (de Ayala, 2009; Zwick, 1990). Therefore, using the observed score in place of  $\theta$  may cause problems for two-parameter logistic (2PL) and 3PL data. Narayanan and Swaminathan (1996) found Type I error was within Bradley's (1978) stringent criterion using 3PL data for all but one instance (i.e., reference group sample size of 500 and focal group sample size of 1,000). Roussos and Stout (1996) found Type I error was maintained when impact was not present but inflated when impact was present using 3PL data. Conversely, Rogers and Swaminathan (1993) found 2PL and 3PL data did not impact the number of Type I errors. The present study fit data using both the correct (same model used for data generation) and incorrect (different model used for data generation) IRT model. Note that when data created using the 1PL model are fitted using the 2PL model there is a case of overfitting and when data created using the 2PL model are fitted using the 1PL model there is a case of underfitting.

### 1.3.5. Impact

Impact occurs when the ability distribution of the groups being analyzed is not the same (Camilli, 2006; Camilli & Shepard, 1994; Clauser & Mazor, 1998; Dorans & Holland, 1993). Zumbo (1999) defined impact as different group probabilities of getting the item right because of true group ability differences on the underlying latent trait designed to be measured by the item. The MC research literature suggested an inconsistent relationship between impact and Type I error rate for LC, the LRT, and the MH procedure. In some instances, Type I error was conservative or maintained while in other cases it was inflated. For LC, Cohen and Kim (1993) did not find a clear relationship between impact and Type I error because their results depended on the nominal alpha level and estimation method. For the LRT, Finch and French (2007) found Type I error did not depend on impact. Finch (2005) found Type I error was generally closer to the nominal level when impact was present. Stark et al. (2006) found Type I error depended on impact and sample size. For MH, when simulating impact from 0.0 for 1.0 with intervals of 0.25 or 0.1 SD unit, Type I error increased as impact increased (DeMars, 2009; Li et al., 2012). However other studies found Type I error was conservative or maintained (Finch, 2005; Narayanan & Swaminathan, 1996; Paek, 2010; Paek & Wilson, 2011). The present study used three levels of impact: 0.0, 0.5, and 1.0.

## 2. METHOD

The open-source software R (R Core Team, 2013) was used to generate the data, run statistical analyses, and compute Type I error while BILOG-MG 3 (Zimowski et al., 2003) was used to estimate IRT models for LC and the LRT. In BILOG-MG, the number of cycles and quadrature points were both changed from the default of 10 to 20 and the number of Newton cycles was changed from the default of two to five to aid in more accurate item parameter estimates. The convergence criterion was changed from the default of 0.01 to 0.1 to aid model convergence for the LRT. Generally, the  $-2LL$  value was greater than 1,000 so this small change did not greatly change the test statistic. In BILOG-MG, neither marginal maximum likelihood estimation (MMLE) nor Bayesian estimation can provide estimates for perfect items (proportion correct of 0.0 or 1.0). A condition was added to exclude datasets with perfect items.

Bayesian estimation, maximum marginal a posterior estimation, was chosen for parameter estimation. Four factors, sample size, group ability differences, IRT model used to generate data, and IRT model used to estimate item parameters, were manipulated in this study. The values selected for each factor were based upon theoretical and empirical rationale. This methodology fits the current trend for replication by providing sufficient detail, which will be discussed. It is at best difficult to compare the results of studies that do not provide sufficient, or sufficiently precise, details needed for replication or comparison. For all simulated conditions, the number of replications was fixed at 10,000 which is a relatively large number of replications as the number of replications in the literature ranges from one to 10,000 (Candell & Drasgow, 1988; Kim & Cohen, 1992; Li et. al, 2012).

In the present study,  $N_R$  and  $N_F$  denoted the number of examinees in the reference group and focal group, respectively. Two conditions,  $N_R = N_F = 500$  and  $N_R = N_F = 1,000$ , were selected to represent moderate and large sample sizes. IRT DIF methods require larger sample sizes to accurately compute the variance-covariance matrix and the  $-2LL$  value. Due to the complexity of computing these IRT DIF statistics, larger sample sizes were used.

Three levels of group mean ability difference, denoted  $\mu_j$ , were manipulated. Theoretically, DIF analyses with group ability differences should not result in Type I error inflation, but prior research has shown that Type I error increased as impact increased (DeMars, 2009; Li et al., 2012). In this study, the reference group mean of the ability distribution was 0.0, 0.5, and 1.0 while the focal group mean of the ability distribution was fixed at 0. In all conditions SD was set at 1.0 for both groups.

A function was written in R to simulate dichotomous item response data (0 for incorrect and 1 for correct) based on a 50 item test with no DIF items for the reference and focal group separately with specified item parameters ( $a_i$ ,  $b_i$ , and  $c_i$ ) and person parameter ( $\theta_j$ ) following the 1PL and 2PL models. Test length was fixed at 50 items, the outer range of previous research (Cohen et al., 1996; Finch, 2005; McLaughlin & Drasgow, 1987). The higher number of items was simulated to obtain an accurate measure of ability for an individual and item difficulty and discrimination estimates for LC and the LRT (Lord, 1968; Rogers & Swaminathan, 1993). The 3PL model was not included as it would constitute another larger paper. The item difficulty parameter function inputs for the 1PL model were generated to follow a normal distribution while the pseudo guessing parameter was fixed at zero, which was consistent with prior research (Herrera & Gomez, 2008; Paek, 2010). For the 2PL model, the item discrimination parameter followed a normal distribution ( $M = 1.1$ ,  $SD = 0.25$ ), which was similar to Paek (2010) (i.e., a normal distribution with ( $M = 1.0$ ,  $SD = 0.3$ )) This produced a range from 0.35 to 1.85 for 99% of values making it highly unlikely to encounter a negatively discriminating item. The item difficulty parameter followed a standard normal distribution, which was consistent with prior research (Herrera & Gomez, 2008; Paek, 2010). For the 1PL model, the item discrimination parameter was fixed at 1.1. This value was chosen for consistency because it was the mean of the item discrimination parameter in the 2PL model. Furthermore, this selection did not introduce any complications when comparing results across models. That is, if the item discrimination parameter had been chosen to be fixed at another value such as 0.8 or 1.2 it would have been more difficult to compare findings based on the IRT model due to the misalignment of item discrimination. In addition, this enhanced the generalizability of findings as data were generated from a different set of parameters each time as opposed to generating item response data based on a single test (Cohen et al., 1996; Sari & Huggins, 2015; Wang & Yeh, 2003). As previously noted, the number of items was fixed at 50 and no DIF items were simulated. For DIF detection, the choices for several parameters for data simulation are, admittedly, arbitrary. IRT model parameters were estimated using both the correct IRT model and incorrect IRT model.

Simulation I examined Type I error rates for LC and the LRT based on the 1PL and 2PL models under varied levels of sample size and impact when fitting the correct IRT model. There were two types of correct model-data fit: (a) generating 1PL model data and fitting the 1PL model (hereafter denoted GEN1FIT1) and (b) generating 2PL model data and fitting the 2PL model (hereafter denoted GEN2FIT2). Fully crossing sample size, impact, and correct IRT model fit to data resulted in 12 cells for Simulation I, which are displayed in [Table 1](#).

**Table 1.** Summary of data collection procedure.

Cell	IRT Model	Sample Size	Impact
1	1PL model	500	0.0
2	1PL model	500	0.5
3	1PL model	500	1.0
4	1PL model	1,000	0.0
5	1PL model	1,000	0.5
6	1PL model	1,000	1.0
7	2PL model	500	0.0
8	2PL model	500	0.5
9	2PL model	500	1.0
10	2PL model	1,000	0.0
11	2PL model	1,000	0.5
12	2PL model	1,000	1.0

To compute Type I error the first item was fixed and selected as the studied item reflecting previous studies (Güler & Penfield, 2009; Li et al., 2012; Roussos & Stout, 1996). Because the data are generated to be in random order, choosing to study the first item is equivalent to choosing a random item. The function *difLord* in the R package *difR* (Magis et al., 2010), was used for LC since simulation within R is advantageous for speed, efficiency, and potential replication. Item parameter estimates from BILOG-MG 3 were used as the inputs to compute LC for both the 1PL and 2PL models. A mean-sigma equating was used to place the focal group item parameter estimates onto the scale of the reference group (Cook & Eignor, 1991). Item parameter equating method was deliberately fixed to control both complexity of the study and the time needed to conduct the simulation. For the LRT, the  $-2LL$  of the compact and augmented model denoted  $L(C)$  and  $L(A)$ , respectively, from BILOG-MG were each saved as vectors in R. All the test items except the studied item were used as the anchor. The LRT was computed by comparing the difference of the two models ( $G^2 = L(C) - L(A)$ ) to a  $\chi^2$  test with 1 *df* and 2 *df* for the 1PL and 2PL model, respectively. When  $-2LL$  differences were negative, implying the counterintuitive result that the compact model provided better fit, results were retained. The *p* value for these negative items was always 1.0 implying they were never rejections. This method was chosen for its consistency with results from IRTLTDIF (Thissen, 2001). As with LC, R was used to compute the LRT. For the MH procedure, the *difMH* function in R package *difR* (Magis et al., 2010) was used with the default of total score, or thin matching, to match the reference and focal group (item purification was judged unnecessary as no DIF items were simulated).

Simulation II examined Type I error rates for LC and the LRT based on the same levels of sample size and group ability difference used in Simulation I, but with the incorrect model fitted. Fully crossing sample size, impact, and incorrect IRT model fit to data also resulted in 12 cells for Simulation II. The only difference between Simulations I and II was whether the



specified IRT model was correctly fitted. There are two types of incorrect model-data fit: (a) generating 1PL model data and fitting the 2PL model (or overfitting, hereafter denoted GEN1FIT2) and (b) generating 2PL model data and fitting the 1PL model (or underfitting, hereafter denoted GEN2FIT1). Based on the previous literature, there was little guidance concerning how incorrect IRT model selection influenced IRT DIF analyses.

Lastly, Simulation III addressed the role of the continuity correction in the MH procedure. Simulation III examined Type I error rates in the MH procedure with and without the continuity correction under the same conditions, used in Simulations I and II. Given the conservative findings of Paek (2010), we included both forms of the MH procedure for completeness. Fully crossing sample size, impact, and IRT model used to generate data resulted in 12 cells for Simulation III. The code is available upon request from the authors.

### 3. RESULTS / FINDINGS

For Simulation I, the results for LC are displayed in Table 2. First, for GEN1FIT1, LC was consistently conservative, pretty stable, and not far from .05 regardless of sample size and impact using Bradley's (1978) stringent criterion. Second, for GEN2FIT2 Type I error rate increased as impact increased ranging from 0.042 to 0.098. Third, for GEN2FIT2 Type I error increased for LC as sample size increased regardless of impact. For GEN2FIT2 Type I error increased as both sample size and impact increased.

**Table 2.** Type I error rates for LC and the LRT when fitting the correct IRT model.

IRT Model	Sample Size	Impact	LC	LRT	MH	MH_CC
1PL model	500	0.0	0.041*	0.049	0.049	0.040*
		0.5	0.043*	0.056**	0.053	0.041*
		1.0	0.041*	0.064**	0.045	0.033*
	1,000	0.0	0.043*	0.050	0.050	0.042*
		0.5	0.040*	0.067**	0.048	0.040*
		1.0	0.043*	0.095**	0.050	0.042*
2PL model	500	0.0	0.042*	0.045	0.049	0.037*
		0.5	0.052	0.053	0.051	0.039*
		1.0	0.077**	0.059**	0.049	0.038*
	1,000	0.0	0.050	0.048	0.050	0.041*
		0.5	0.060**	0.063**	0.051	0.044*
		1.0	0.098**	0.075**	0.052	0.044*

Note. MH = the MH procedure without continuity correction; MH\_CC = the MH procedure with continuity correction. Values marked with an \* are conservative based on Bradley's (1978) stringent criterion; Values marked with \*\* are inflated based on Bradley's (1978) stringent criterion. The degrees of freedom are 1 and 2 for the 1PL and 2PL models, respectively.

The results for the LRT are also displayed in Table 2. First, when the groups were matched on ability Type I error was maintained, ranging from 0.045 to 0.050, for all four combinations of IRT model and sample size. Second, when the groups were not matched on ability (impact of 0.5 and 1.0) Type I error was inflated in all instances except one. The exception was GEN2FIT2 with a group sample size of 500 and impact of 0.5, which resulted in maintained Type I error. The actual Type I error rates ranged from 0.053 to 0.095 when impact was present. Third, Type



I error increased as sample size increased for all conditions in the LRT. The same results (conservative, maintained, or inflated) were seen across sample size in all but one condition. The exception was GEN2FIT2 with impact of 0.5. For this condition Type I error was maintained with a sample size of 500 but inflated with a sample size of 1,000.

For Simulation II, the results for LC are displayed in Table 3. First, in both cases of model misspecification Type I error for LC increased as impact increased across sample size ranging from 0.049 to 0.244. Second, in both cases of model misspecification Type I error rate for LC was consistently higher for the larger sample size condition compared to the smaller sample size condition. Third, Type I error was higher for GEN2FIT1 compared GEN1FIT2 in all combinations of sample size and impact except one. The exception was a sample size of 1,000 with impact of 0.0. Moreover, Type I error ranged from 0.041 to 0.091 compared to 0.044 to 0.244 based on GEN1FIT2 and GEN2FIT1, respectively. Fourth, the same Type I error results (conservative, maintained, and inflated) were seen for LC in all combinations of sample size and impact of GEN1FIT2 and GEN2FIT1 except one. The exception was a sample size of 500 and impact of 0.5 where Type I error was maintained for GEN1FIT2 but inflated for GEN2FIT1.

**Table 3.** Type I error rates for LC and the LRT when fitting the incorrect IRT model.

IRT Model Used to Generate Data	IRT Model Fit to Data	Sample Size	Impact	LC	LRT	MH	MH_CC
1PL model	2PL model	500	0.0	0.041*	0.044*	0.049	0.037*
			0.5	0.049	0.053	0.051	0.039*
			1.0	0.081**	0.055	0.049	0.038*
		1,000	0.0	0.047	0.048	0.050	0.041*
			0.5	0.059**	0.056**	0.051	0.044*
			1.0	0.091**	0.079**	0.052	0.044*
2PL model	1PL model	500	0.0	0.044*	0.073**	0.049	0.040*
			0.5	0.081**	0.131**	0.053	0.041*
			1.0	0.152**	0.212**	0.045	0.033*
		1,000	0.0	0.045	0.100**	0.050	0.042*
			0.5	0.110**	0.227**	0.048	0.040*
			1.0	0.244**	0.347**	0.05	0.042*

Note. MH = the MH procedure without continuity correction; MH\_CC = the MH procedure with continuity correction. Values marked with an \*are conservative based on Bradley’s (1978) stringent criterion; Values marked with \*\* are inflated based on Bradley’s (1978) stringent criterion. The degrees of freedom are based on the IRT model fit to the data (i.e., 1 and 2 for the 1PL and 2PL models, respectively).

The results for the LRT are also displayed in Table 3 demonstrating one clear pattern: for all conditions Type I error rate increases as impact increases. There are five additional points worth noting. First, for GEN2FIT1 Type I error was inflated for all conditions of sample size and impact. Second, for GEN1FIT2 Type I error conclusions depended on sample size and impact. That is, with a sample size of 500 Type I error was conservative when groups were matched on ability but maintained when impact was present. However, with a sample size of 1,000 Type I error was maintained when the groups were matched on ability but inflated when impact was present. Third, Type I error rates were larger in GEN2FIT1 compared to GEN1FIT2 for all conditions of sample size and impact. Fourth, within each model misspecification category Type I error was higher for the larger sample size condition compared to the smaller sample

size condition. Fifth, the difference in Type I error rate from the larger sample size to the smaller sample size was larger in GEN2FIT1 compared to GEN1FIT2 regardless of impact.

For Simulation III, the results for the standard MH procedure using the continuity correction and the MH procedure without the continuity correction are given in [Tables 1](#) and [2](#). In both tables the MH results are the same and were added to facilitate comparisons among the three DIF methods. For both forms of the MH procedure there was one consistent finding when using Bradley's (1978) stringent criterion across all simulated conditions: Type I error rates were conservative for the traditional MH procedure while Type I error rates were maintained for the MH procedure without the continuity correction. Furthermore, for both forms of the MH procedure impact, sample size, IRT model used to generate the data, and no combination of these three variables influenced Type I error rates to any great extent.

#### **4. DISCUSSION and CONCLUSIONS**

This study adds to the research literature by investigating IRT model specification or correct and incorrect model-data fit. Portions of the simulation results agree with previous research while other portions disagree. It is difficult to compare the results of this study with prior literature because many studies do not provide the methodological details needed for replication and comparison (e.g., two studies that examined LC [Lim & Drasgow, 1990; McLaughlin & Drasgow, 1987] did not mention item parameter equating). Overall, the results demonstrated two conclusions. First, when using large sample sizes of 500 and 1,000 per group regardless of impact and IRT model used to generate data the MH procedure is the preferred DIF method due to its Type I error performance consistency. Second, when using IRT DIF methods correct and incorrect IRT model specification and the effect of group differences cannot be ignored.

For LC in Simulation I GEN1FIT1 Type I error rates did not depend on sample size and impact using Bradley's (1978) stringent criterion, which is consistent with previous research (Wells et al., 2009). For GEN2FIT2, Type I error increased as sample size increased, which is consistent with research by Lim and Drasgow (1990) and McLaughlin and Drasgow (1987), but inconsistent with Kim et al. (1994). For GEN2FIT2 Type I error rate increased with impact, which did not agree with previous literature by Cohen and Kim (1993).

For the LRT in Simulation I Type I error was reasonably maintained when the groups were matched on ability for all four conditions of model-data fit and sample size. This was inconsistent with previous research (Finch, 2005; Finch & French, 2007; Stark et al., 2006). When the groups were not matched on ability (impact of 0.5 and 1.0) Type I error was inflated in seven of the eight model-data fit and sample size conditions. The exception was GEN2FIT2 with a group sample size of 500 and impact of 0.5 in which Type I error was maintained. These results were reasonably consistent with previous research (Finch, 2005; Stark et al., 2006). When impact was present with GEN2FIT2 Finch (2005) found Type I error was maintained with a group sample size of 500 while Stark et al. (2006) found Type I error somewhat conservative with a group sample size of 1,000. Although Type I error increased as sample size increased for all conditions, similar conclusions were generally made across sample size. This was inconsistent with some previous research (Cohen et al., 1996; Finch, 2005; Stark et al., 2006). There are several reasons why the present study may have inconsistencies with prior work. For example, the estimation methods differed for Cohen et al. (1996) and Finch and French (2007) and prior studies used 50-1,000 replications while the present study used 10,000.

For Simulation II, Type I error conclusions for GEN1FIT2 were generally consistent for LC and the LRT across the conditions with one exception. The exception was GEN1FIT2 with a group sample size of 500 and impact of 1.0 in which Type I error was inflated for LC and maintained for the LRT. However, for GEN2FIT1 the Type I error conclusions were only consistent when impact was present. Type I error increased as impact increased for all

conditions in both DIF methods. Type I error was generally inflated in both GEN1FIT2 and GEN2FIT1, but was more pronounced in GEN2FIT1. Finally, Type I error rates were lower for LC than the LRT in GEN2FIT1 but varied in GEN1FIT2. There was little research with which to compare these findings.

In Simulation III, Type I error rate was consistently somewhat conservative for the MH procedure. Impact did not lead to Type I error inflation when other variables were manipulated, which was both consistent and inconsistent with previous research (DeMars, 2009; Finch, 2005; Narayanan & Swaminathan, 1996; Paek, 2010; Roussos & Stout, 1996). Sample size did not influence Type I error rates when other variables were manipulated, which was also both consistent and inconsistent with prior research (DeMars, 2009; Herrera & Gómez, 2008; Narayanan & Swaminathan, 1996; Paek & Wilson, 2011; Roussos & Stout, 1996). The IRT model used to generate the data did not influence Type I error rates when other variables were manipulated. This was an interesting finding because the MH procedure matches the groups on observed score and not the underlying latent variable,  $\theta$ . This finding was consistent with Rogers and Swaminathan (1993) who investigated how the 2PL model and 3PL model impacted the distributional shape of the MH test statistic and found no drastic differences in the number of Type I errors made for model-data fit. The finding also was consistent with Paek (2010) who simulated 1PL, 2PL, and 3PL data finding that Type I error was consistently conservative regardless of the IRT model used to generate data. This is noteworthy because the default method for the MH procedure in SPSS uses the continuity correction. Researchers and practitioners need to be mindful of this when interpreting their DIF results as they may be conservative. Furthermore, this study adds to the literature by extending the findings of Paek (2010). Paek (2010) examined the MH procedure under a variety of conditions while the present study included the MH procedure in conjunction with IRT DIF methods so that comparisons can be made between the two types of methods.

A key observation for the MH procedure was that no combination of IRT model used to generate the data, sample size, and impact, influenced Type I error rates to any great extent. This finding agrees with some previous research (Paek, 2010; Paek & Wilson, 2011), but was inconsistent with other research (DeMars, 2009; Herrera & Gómez, 2008; Narayanan & Swaminathan, 1996; Roussos & Stout, 1996).

#### 4.1. Recommendations

There are six main recommendations based on this study. Recommendation one applies to statistical software for DIF analyses. Recommendations two through four apply to the results of the MH procedure and LC and the LRT using correct and incorrect model-data fit. The fifth recommendation compares the results of all three DIF procedures while recommendation six is a more general reflection on simulation studies.

First, it is important to empirically validate any R packages of interest prior to use. The authors of this study were not able to replicate the item parameter estimates from ltm at the time of data generation (Rizopoulos, 2006) which were used to implement LC. Thus, BILOG-MG was used in place of ltm.

Second, DeMars (2010) pointed out that when groups are not matched on ability Type I error can become inflated for the MH procedure. Previous research has shown this was true (DeMars, 2009; Li et al., 2012; Roussos & Stout, 1996). This study, however, demonstrates that this is not always the case. The Type I error for the MH procedure without the continuity correction was reasonably unaffected by group differences (impact) for all simulated conditions. This finding is important because the MH procedure is theoretically easy to understand, easy to implement, does not require knowledge of IRT, and is often used for DIF detection (Holland & Thayer, 1988; Wainer, 2010). Furthermore, this finding is particularly valuable to

psychometricians and applied researchers because it supports the use of the MH procedure for DIF analyses in the presence of impact based on Type I error rate.

Third, both studied IRT procedures generally showed inflated (and sometimes highly inflated) Type I errors with the combination of item impact and model misspecification. However, if the groups are matched on ability the LRT may be slightly preferred to LC when fitting the correct IRT model. Moreover, when the data are fit to the correct IRT model LC is often too conservative while the LRT is often too inflated based on Bradley's (1978) stringent criterion. When the data are fit to the incorrect IRT model, Type I error increased and became inflated as impact increased for both sample sizes. Therefore, choosing an appropriate IRT model for existing data is an important consideration (e.g., Bolt et al., 2014; Köse, 2014; Maydeu-Olivares, 2013) and, done well, can be arduous. In their chapter on the assessment of model-data fit, Hambleton et al. (1991) recommend a comprehensive set of procedures for assessing IRT model fit including checking model assumptions, parameter invariance, and model predictions. Furthermore, this IRT DIF finding is noteworthy because implementing LC based upon fitting the 1PL model is the only DIF procedure implemented in BILOG-MG 3. Therefore, psychometricians and applied researchers conducting DIF analyses using BILOG-MG 3 must be careful that their data correctly fit the 1PL model. That is, if a psychometrician or applied researcher is using BILOG-MG 3 for DIF analyses and the true underlying IRT model is the 2PL DIF results should be interpreted with caution as serious Type I error inflation can occur especially in the presence of impact.

Moreover, this recommendation is important because Type I error inflation is a serious problem as test items are expensive to construct. Luecht (2005) states that the: "ACPI [average-cost-per-item] typically runs from several hundred to more than fifteen hundred dollars per item" (p. 8). That is, making a Type I error by falsely removing a non DIF item is a serious financial consequence, which cannot be taken lightly.

Fourth, this study did not identify any unique advantage for using IRT methods over CTT methods based on Type I error rates. That is, based on Type I error rates the findings of this study do not support the theoretical advantages of using IRT for DIF analyses despite the recommendations of Camilli and Shepard (1994).

Fifth, taken together, recommendations two, three, and four agree with the principle of parsimony that the simpler method in comparison to the more complex method is better. That is, based on Type I error rate the MH procedure, a non-IRT based DIF method, should be used for DIF analyses instead of the more complex IRT DIF methods (LC and the LRT), which agrees with prior recommendations (Holland & Thayer, 1988; Wainer, 2010). Furthermore, this finding is particularly valuable to psychometricians and applied researchers because it supports a simpler method to implement and does not rely on correct IRT model selection. That is, the MH procedure overcomes the problem of Type I error inflation of LC and the LRT when selecting the incorrect IRT model.

Sixth, it is critically important that simulation studies in all areas provide sufficient detail for both comparison with prior research and replication. More bluntly, Monte Carlo research can be much more than a collection of case studies.

### **Acknowledgments**

The authors received no financial support for the research, authorship, and/or publication of this article. Some pilot research that led to this manuscript was accepted and presented at the annual meeting of the American Educational Research Association in Philadelphia in April 2014.

## Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research and publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

## Authorship contribution statement

**Emily Diaz:** Investigation, Resources, Methodology, Visualization, Software, Formal Analysis, and Writing original draft. **Gordon Brooks:** Methodology, Supervision, Validation, and Writing original draft. **George Johanson:** Methodology, Supervision, and Writing original draft.

## ORCID

Emily Diaz  <https://orcid.org/0000-0001-9460-8647>

Gordon Brooks  <https://orcid.org/0000-0002-2704-2505>

George Johanson  <https://orcid.org/0000-0002-4253-1841>

## 5. REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.
- Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, 15(2), 113-141. [https://doi.org/10.1207/S15324818AME1502\\_01](https://doi.org/10.1207/S15324818AME1502_01)
- Bolt, D. M., Deng, S., & Lee, S. (2014). IRT model misspecification and measurement of growth in vertical scaling. *Journal of Educational Measurement*, 51(2), 141-162. <https://doi.org/10.1111/jedm.12039>
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2), 144-152. [https://doi.org/10.1207/S15324818AME1502\\_01](https://doi.org/10.1207/S15324818AME1502_01)
- Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 220-256). American Council on Education.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage.
- Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, 12(3), 253-260. <https://doi.org/10.1177/014662168801200304>
- Cohen, A. S., & Kim, SH. (1993). A comparison of Lord's  $\chi^2$  and Raju's area measures in detection of DIF. *Applied Psychological Measurement*, 17(1), 39-52. <https://doi.org/10.1177/014662169301700109>
- Cohen, A. S., Kim, S. H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement*, 20(1), 15-26. <https://doi.org/10.1177/014662169602000102>
- Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice*, 10(3), 37-45. <https://doi.org/10.1111/j.1745-3992.1991.tb00207.x>
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.
- Creswell, J. W. (2009). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Sage.
- DeMars, C. E. (2009). Modification of the Mantel-Haenszel and Logistic Regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics*, 34(2), 149-170. <https://doi.org/10.3102/1076998607313923>



- DeMars, C. E. (2010). Type I Error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement*, 70(6), 961-972. <https://doi.org/10.1177/013164410366691>
- Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Lawrence Erlbaum.
- Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, 29(4), 278-295. <https://doi.org/10.1177/0146621605275728>
- Finch, W. H., & French, B. F. (2007). Detection of crossing differential item functioning a comparison of four methods. *Educational and Psychological Measurement*, 67(4), 565-582. <https://doi.org/10.1177/0013164406296975>
- Güler, N., & Penfield, R. D. (2009). A Comparison of the Logistic Regression and Contingency Table Methods for Simultaneous Detection of Uniform and Nonuniform DIF. *Journal of Educational Measurement*, 46(3), 314-329. <https://doi.org/10.1111/j.17453984.2009.00083.x>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.
- Herrera, A. N., & Gómez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel-Haenszel and logistic regression techniques. *Quality & Quantity*, 42(6), 739-755. <https://doi.org/10.1007/s11135-006-9065-z>
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Lawrence Erlbaum.
- Kane, M. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 39-64). Information Age Publishing.
- Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73. <https://doi.org/10.1111/jedm.12000>
- Kim, S. H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement*, 29(1), 51-66. <https://doi.org/10.1111/j.17453984.1992.tb00367.x>
- Kim, S. H., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education*, 8(4), 291-312. <https://doi.org/10.1207/s15324818ame08042>
- Kim, S. H., Cohen, A. S., & Kim, H. O. (1994). An investigation of Lord's procedure for the detection of differential item functioning. *Applied Psychological Measurement*, 18(3), 217-228. <https://doi.org/10.1177/014662169401800303>
- Köse, I. A. (2014). Assessing model data fit of unidimensional item response theory models in simulated data. *Educational Research and Reviews*, 9(17), 642-649. <https://doi.org/10.5897/ERR2014.1729>
- Lautenschlager, G. J., & Park, D. G. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness, and parameter linking. *Applied Psychological Measurement*, 12(4), 365-376. <https://doi.org/10.1177/014662168801200404>
- Li, Y., Brooks, G. P., & Johanson, G. A. (2012). Item discrimination and Type I error in the detection of differential item functioning. *Educational and Psychological Measurement*, 72(5), 847-861. <https://doi.org/10.1177/0013164411432333>



- Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology, 75*(2), 164-174. <https://doi.org/10.1037/0021-9010.75.2.164>
- Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5*(2), 159-173. <https://doi.org/10.1177/014662168100500202>
- Lord, F. M. (1968). An analysis of the verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*, 989-1020. <https://doi.org/10.1177/001316446802800401>
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.
- Luecht, R. M. (2005). Some useful cost-benefit criteria for evaluating computer-based test delivery models and systems. *Journal of Applied Testing Technology, 7*(2), 1-31.
- Magis, D., Beland, S., Tuerlinckx, S., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*, 847-862. <https://doi.org/10.3758/BRM.42.3.847>
- Marañón, P. P., García, M. I. B., & Costas, C. S. L. (1997). Identification of nonuniform differential item functioning: A comparison of Mantel-Haenszel and item response theory analysis procedures. *Educational and Psychological Measurement, 57*(4), 559-568. <https://doi.org/10.1177/0013164497057004002>
- Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives, 11*(3), 71-101. <https://doi.org/10.1080/15366367.2013.831680>
- Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement, 54*(2), 284-291. <https://doi.org/10.1177/013164494054002003>
- McLaughlin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. *Applied Psychological Measurement, 11*(2), 161-173. <https://doi.org/10.1177/014662168701100205>
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. <https://doi.org/10.1037/0003-066X.50.9.741>
- 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>
- Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*(3), 257-274. <https://doi.org/10.1177/014662169602000306>
- National Research Council. (2007). *Lessons learned about testing: Ten years of work at the National Research Council*. The National Academies Press.
- Paek, I. (2010). Conservativeness in rejection of the null hypothesis when using the continuity correction in the MH chi-square test in DIF applications. *Applied Psychological Measurement, 34*(7), 539-548. <https://doi.org/10.1177/0146621610378288>
- Paek, I., & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel-Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement, 71*(6), 1023-1046. <https://doi.org/10.1177/0013164411400734>

- R Core Team (2013). R: A language and environment for statistical computing. [Computer software]. R Foundation for Statistical Computing: Vienna, Austria. <http://www.R-project.org/>.
- Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, 14(2), 197-207. <https://doi.org/10.1177/014662169001400208>
- Raju, N. S., Drasgow, F., & Slinde, J. A. (1993). An empirical comparison of the area methods, Lord's chi-square test, and the Mantel-Haenszel technique for assessing differential item functioning. *Educational and Psychological Measurement*, 53(2), 301-314. <https://doi.org/10.1177/0013164493053002001>
- Rizopoulos, D. (2006). Ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1-25. <https://doi.org/10.18637/jss.v017.i05>
- Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, 17(2), 105-116. <https://doi.org/10.1177/014662169301700201>
- Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement*, 33(2), 215-230. <https://doi.org/10.1111/j.1745-3984.1996.tb00490.x>
- Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). A Monte Carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement*, 17(1), 1-10. <https://doi.org/10.1111/j.1745-3984.1980.tb00810.x>
- Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, 66(1), 63-84. <https://doi.org/10.1177/0013164404273942>
- Sari, H. I., & Huggins, A. C. (2015). Differential item functioning detection across two methods of defining group comparisons: Pairwise and composite group comparisons. *Educational and Psychological Measurement*, 75(4), 648-676. <https://doi.org/10.1177/0013164414549764>
- Shepard, L., Camilli, G., & Williams, D. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement*, 22(2), 77-105. <https://doi.org/10.1111/j.1745-3984.1985.tb01050.x>
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, 91(6), 1292-1306. <https://doi.org/10.1037/00219010.91.6.1292>
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7(2), 201-210. <https://doi.org/10.1177/014662168300700208>
- Thissen, D. (2001). *IRTLRDIF user's guide: software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning*. Retrieved from <http://www.unc.edu/~dthissen/dl.html>
- Thissen, D., Steinberg, L., Pyszczynski, T., & Greenberg, J. (1983). An item response theory for personality and attitude scales item analysis using restricted factor analysis. *Applied Psychological Measurement*, 7(2), 211-226. <https://doi.org/10.1177/014662168300700209>
- Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Lawrence Erlbaum.

- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Lawrence Erlbaum.
- Wainer, H. (2010). 14 conversations about three things. *Journal of Educational and Behavioral Statistics*, 35(1), 5-25. <https://doi.org/10.3102/1076998609355124>
- Wang, WC., & Yeh, YL. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, 27(6), 479-498. <https://doi.org/10.1177/0146621603259902>
- Wells, C. S., Cohen, A. S., & Patton, J. (2009). A range-null hypothesis approach for testing DIF under the Rasch model. *International Journal of Testing*, 9(4), 310-332. <https://doi.org/10.1080/15305050903352073>
- Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3: Item analysis and test scoring with binary logistic models*. [Computer software]. Scientific Software.
- Zumbo, B. (1999). *A handbook on the theory and methods of differential item functioning: Logistic regression modeling as a unitary framework for binary and Likert-type item scores*. Directorate of Human Resource Research and Evaluation, National Defense Headquarters.
- Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational and Behavioral Statistics*, 15(3), 185-197. <https://doi.org/10.3102/10769986015003185>

## Views of Teachers on the Potential Negative Effects of High Stake Tests

Mustafa İlhan <sup>1</sup>, Nese Guler <sup>2</sup>, Gulsen Tasdelen Teker <sup>3,\*</sup>

<sup>1</sup>Dicle University, Ziya Gökalp Education Faculty, Department of Mathematics and Science Education, Diyarbakır, Turkey

<sup>2</sup>İzmir Demokrasi University, Faculty of Education, Department of Measurement and Evaluation, İzmir, Turkey

<sup>3</sup>Hacettepe University, Faculty of Medicine, Department of Medical Education and Informatics, Ankara, Turkey

### ARTICLE HISTORY

Received: May 10, 2020

Revised: Jan. 27, 2021

Accepted: Apr. 05, 2021

### Keywords:

High stake exams,  
Pairwise comparisons,  
Many-facet Rasch model,  
Scaling method

**Abstract:** This study analyses teachers' viewpoints on the potential undesirable influences of high stake exams. Seven stimulants related to undesirable influences of high stake exams on education were given to 191 teachers in a pairwise comparisons format. The participating teachers in this study were asked to choose one undesirable influence by comparing the stimulants given to them in pairwise and to determine the more prominent problem stemming from high stake exams. Data were analyzed via many-facet Rasch model. As a result, it was found that teachers considered the stimulant "school assessments turn into secondary importance in the eyes of students and parents" as the foremost problem stemming from high stake exams. On the other hand, the stimulant "administrators focus on policies for increasing exam scores instead of policies for improving the learning-teaching process" ranked the last in the undesirable influences of high stake exams.

## 1. INTRODUCTION

Examinations are an integral part of educational processes. Individuals' proficiency levels in various fields are determined via examinations and in accordance with the results obtained and decisions are made about the success of students, the functioning of curricula, or the quality of teaching. Yet, the importance of decisions made based on exam results is not always the same for students, parents, teachers, schools, or educational policy-makers. While some of the exam results form the basis for at least one key stakeholder of education to make extremely important decisions, some of the results are used to make decisions with relatively more restricted effects on individuals. Considering the impact of the decisions it creates on students, parents, education administrators, or policy makers and the traces they leave on individuals' lives, exams can be examined under two headings; namely, low stake exams and high stake exams.

Exams whose results are not used to make important decisions for students or educators are called low stake exams. Subject screening tests and quizzes given in order to determine the deficiencies in learning and to plan the improvements to be made in teaching are the most typical examples for low stake exams. Low stake exams have such an important advantage as

---

\*CONTACT: Gülşen Taşdelen Teker ✉ [gulsentasdelen@gmail.com](mailto:gulsentasdelen@gmail.com) 📍 Hacettepe University, Faculty of Medicine, Department of Medical Education and Informatics, Ankara, Turkey

causing no stress to individuals (Simpson, 2016). On the other hand, when individuals know that the test they take will not have important results for them, they can lack motivation (Finn, 2015; Kornhauser et al., 2014), which would result in doubtful approaches towards the validity of information that low stake exams present in relation to examinees' performance (Wise & Demars, 2005).

Tests are defined as high stake exam if individuals gain or lose a lot according to the results obtained from them (Coniam & Falvey, 2007). What makes an exam at high stake is often the impact it may have on the educational life and career opportunities of the candidates (Moses & Nanna, 2007). However, sometimes their impact on teachers, educational administrators, or educational policy makers can make an exam a high stake one. For instance, even though such international exams as the Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS) are low stake ones for students taking the tests, they can be high stake for governments, educational policy makers, and schools (Stobart & Eggen, 2012). In a similar vein, if the scores students receive from the exams they take are used for evaluating teachers' performance and making decisions about their appointment and promotion, they have high risks for teachers but not for students (Dawson, 2012). However, this study focuses on high stake exams that affect primarily students and therefore their parents.

While high stake exams have several advantages that are referred to as justification for large areas of use, on the one hand, they also have a number of disadvantages causing them to be the focus of complaints. A number of major advantages that high stake exams offer are listed as follows: It is thought that the use of the in-class assessment results in order to make decisions about individuals causes justice problems since lecture notes can differ from teacher to teacher and from school to school (Holland, 2001; cited in Brockmeier et al., 2014). However, high stake exams are believed to make relatively fairer measurements (Phelps, 2003). Therefore, making important decisions about individuals according to the results of high stake exams is perceived by society as a more reliable (Çetin & Ünsal, 2018) and conscientious alternative (Baykal, 2014). Another advantage of high stake exams is that they can provide national and even international data on the success of students, schools, and educational systems (Acar Gündür, 2014; Baines & Stanley, 2004). Other positive qualities of high stake exams include raising teachers' sense of responsibility and creating the need for updating themselves (Çetin & Ünsal, 2018), motivating students to study harder on the one hand, and giving feedback to students about their strengths and weaknesses on the other hand (Stecher, 2002).

What turns high stake exams into a controversial subject despite the above-mentioned advantages they offer is that the disadvantages of them are enough to overshadow their advantages. The disadvantages of high stake exams can be summarized as follows: Teaching applications (such as lab work, educational trips, etc.) and lessons not included by the exams are handled superficially (Taylor et al., 2003) and teaching process is regulated in accordance with exam content (Yeh, 2005). These exams put pressure on mainly students, parents, educators, and administrators (Kruger et al., 2007), ensuring that increase in students' exam scores becomes the primary purpose of education (Pbrreault, 2000) and such practice moves education away from collaborative mentality and turns it into a rivalry-oriented system (Polesel et al., 2012)

High stake exams are widely applied in many countries in the world and Turkey is not exceptional in using such exams. In the Turkish educational system, exams administered at a national scale are taken into consideration in making such important decisions as to transition into secondary and higher education, the selection of personnel for employment in the public sector, and determining the individuals who will receive specialisation training in medicine and dentistry. As a result of its widespread use in the educational process, research related to high



stake exams takes a large place in the related literature. When the studies in the literature on the subject are looked into, it is figured out that the effects of high stake exams on students (Amrein & Berliner, 2003; Banks & Smyth, 2015; Segool et al., 2013; Simpson, 2016), teachers (Abrams, 2004; Assaf; 2008; Brady, 2008; Christian, 2010; Dawson, 2012), parents (Polesel et al., 2012; Saito, 2006; Westfall, 2010), teaching-learning process, and curriculum implementation (Amoako, 2019; Finkeldei, 2016; Davis, 2011; Johnson, 2007; Marchant, 2004; Ritt, 2006; Shepard & Dougherty, 1991; Taylor et al., 2002; Togut, 2004; Vogler & Virtue, 2007; Wright, 2002) are generally focused. Current research in the literature provides important information about the undesirable influences of high stake exams in the educational process but does not provide any information on which of the problems stemming from high stake exams is prioritized or which is more in the background. However, in the literature, it is pointed out that the first thing to do in order to overcome the problems caused by an application is to prioritize the problems while the first 20% of the problems are expressed as the causes of the remaining 80%. According to the 80/20 rule (Knapp, 2010), the effectiveness of the steps taken to solve a problem depends on the fact that these steps are related to the first 20% of the problems (Kane, 2014). Therefore, in order to produce the right solutions for the undesirable influences of high stake exams, it is thought that the problems arising from these exams should be sorted out.

### **1.1. The Purpose and Significance of the Study**

The aim of the current study is to analyse teachers' viewpoints on the potential undesirable influences of high stake exams which interest almost all individuals in the society directly or indirectly in terms of their results through pairwise comparisons based on many-facet Rasch model (MFRM). The research is thought to make two important contributions to the literature: Firstly, this study differs from other research in the literature in that it intends to reveal the problems that need to be addressed primarily, beyond identifying the problems caused by high stake exams; the second feature of this research making it important for the literature is that it has a methodological difference. When the studies utilizing the scaling method through pairwise comparisons in the literature are examined (Nartgün, 2006; Anıl & Güler, 2006; Bülbül & Acar, 2012; Ekinçi et al., 2012; Güler & Anıl, 2009; Nalbantoğlu Yılmaz, 2017; İlhan, 2016; Yaşar, 2018), it can be ascertained that the analyses have been performed usually by means of Microsoft Excel and also they have been done on the basis of traditional psychometric approach. Even though the study performed by Güzeller, Eser and Aksu (2016) differs from other studies available in the literature in that it analyses the pairwise data by using R software, it is similar to other pairwise comparison studies in that traditional psychometric approach is dominant in the analysis process. In the present study, however, the collected data through pairwise comparisons are analyzed on the basis of MFRM. When the pairwise comparison data are analyzed within the framework of traditional psychometric approach, statements of indiscrimination are not permitted and the participants are always asked to make a choice between two stimulants (Turgut & Baykul, 1992). In such a case, some of the pairwise comparisons can be left unanswered and thus it becomes difficult to collect data, which may result in a probable loss of data. The first advantage in analyzing the data of pairwise comparisons emerges at this point. When the stimulants compared in pairs are analyzed using the MFRM, participants are not always expected to make a choice. In fact, they are allowed to think equivalently about the two stimulants (Linacre, 2014). Another advantage offered by analyzing the data of pairwise comparisons in MFRM is that statistics which provide evidence for psychometric properties of measurements are reported synchronically with the stimulants' scale values. The fit statistics calculated for the stimulant facet, reliability coefficient, and separation ratio along with the scale values for the stimulants are also available in the many-facet Rasch analysis outputs. There is evidence for both validity and reliability: Fit statistics



provide evidence for the former and reliability coefficient and separation ratio provide evidence for the latter.

In addition, individuals inform us about whether model-data fit exists and whether there are interactions between facets even though they are not used for measurements in pairwise comparisons based on Rasch analysis (Linacre, 2014). Considering all these advantages, it is believed that providing a sample study analyzing pairwise comparisons data according to MFRM for the literature would be important. In this context, the study is also thought to have potential to contribute to the literature in terms of methodology.

## **2. METHOD**

### **2.1. Research Model**

This study is a descriptive survey research. Descriptive research is based on the principle of revealing the present situation without any intervention and is mostly considered as a survey model (Erkuş, 2011). Basic characteristics of survey model include gathering information from individuals to define certain characteristics (attitude, belief, opinion, ability, etc.) of the universe to which they belong (Fraenkel et al., 2012), requiring a large sample selection to represent the universe, presenting standard information obtained by applying the same measurement tool to all individuals in the sample, and collecting quantitative data on which statistical procedures can be performed (Cohen et al., 2007).

### **2.2. The Study Group**

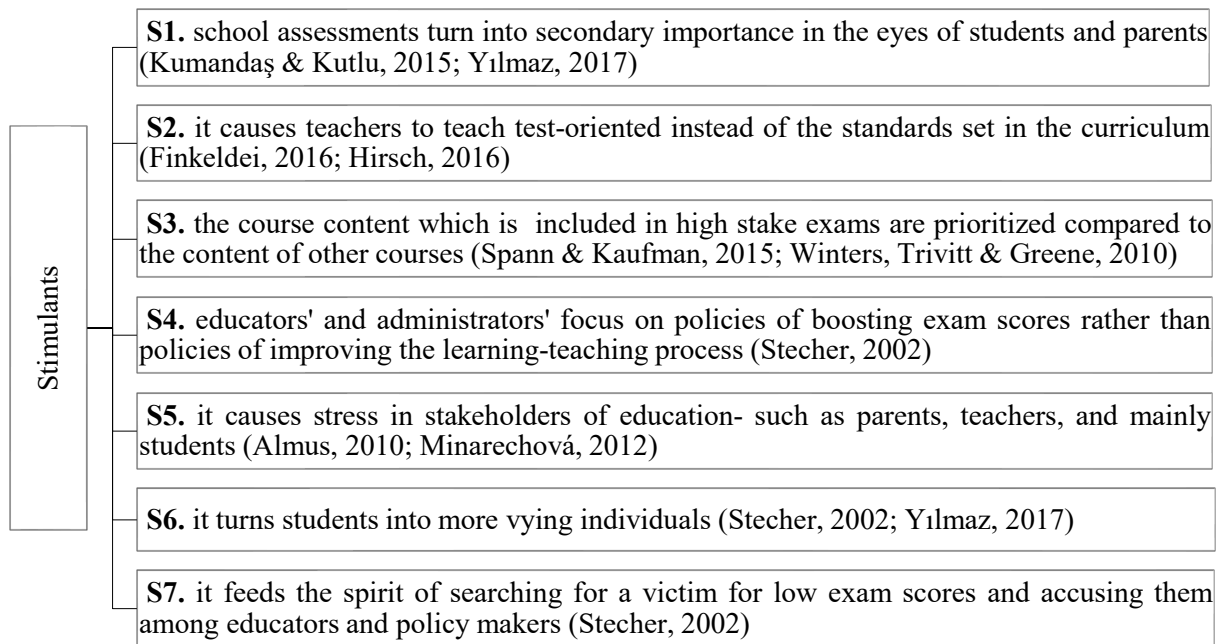
The study was conducted on 191 teachers working in Turkey. Of all the participating teachers 88 (46.07%) were female, 84 (43.49%) were male, and 19 (9.95%) did not mention their gender. The distribution of the teachers according to the stage of education they taught was as follows: 41 (21.47%) primary school teachers, 74 (38.74%) secondary school teachers, and 76 (39.79%) high school teachers. The teachers included in the study group ranged between 22 and 61 years old ( $\bar{X}$ =34.33) and they had been teaching for 1–33 years ( $\bar{X}$ =10.17).

### **2.3. Data Collection Tool**

The relevant literature about the potential undesirable results that high stake exams can yield was reviewed prior to forming the data collection tool. It was seen in the literature review that there were many instructional and affective undesired influences associated with high stake exams. While deciding on the stimulants to be included in the data collection tool, the negative effects mentioned in almost all of the studies were examined. Consequently, the seven stimulants given in [Figure 1](#) coded as and ranked between S1 and S7 and mentioned frequently in most of the examined studies were determined as the major undesirable influences caused by high stake exams in the education process.

The above mentioned seven stimulants were arranged in a way that the teachers involved in the research could make pairwise comparisons and the data collection tool was formed. The data collection tool included 21 comparisons containing the pairwise combinations of the seven stimulants as well as demographic variables of gender, age, branch, and duration of service in the teaching profession. An instruction about the purpose of the study and how to answer the measurement tool was added to the beginning of the instrument. In the directive, the phrase central examination was adopted instead of a high stake exam. Because it was thought that the term of central exam would be more understandable for teachers compared to the concept of high stake exam. In addition, in order to make the statement of the central exam clearer, which exams are included in this scope were exemplified in parentheses; namely, after the central exam phrase a parenthesis was opened and examples of central examination implemented in Turkey was listed.

**Figure 1.** *The Potential Undesirable Influences of High Stake Exams in Education.*



Prior to its use, the data collection tool was presented to the opinion of a total of five experts, each with at least a PhD degree: two measurement and evaluation experts, one curriculum and instruction expert, one training management, one supervision and planning expert, and one psychological counselling and guidance expert. Experts reported that the directive and statements in the data collection tool were understandable and contained the main undesirable influences that could be associated with high stake exams. Then, two experts in the field of language education, one of whom attended doctorate education and the other had the title of associate professor, were interviewed. Data collection phase was started when experts reported that there were no spelling rules, punctuation marks, or problems related to narration.

#### 2.4. Data Collection Procedure

The data were collected during the 2017–2018 and 2019–2020 academic years. When the participating teachers were asked to compare the stimulants pairs given to them in the instrument, they stated the more important problem arising from the high-stakes exams. They were also asked to indicate the comparisons they could not make. The data collection process was completed in this way. No missing or incorrectly filled measurement tool was found in the data set. On examining the data collected, it was detected that the teachers had found it difficult to make a choice between the two stimulants in 256 out of 4011 comparisons [(191 teachers) x (21 pairs of stimulants)]. It was also found that they did not have difficulty in the remaining 3755 comparisons.

#### 2.5. Data Analysis

The MFRM is an extension of the basic Rasch model and is a highly functional analysis for situations where there are different sources of variability that can affect the measurement results other than items and individuals. The MFRM was initially conceptualized as consisting of item, individual, and rater facets. However, later on, the model was expanded to increase the number of facets and the model was started to be used in different problem situations and data sets other than the rater mediated assessments. For example, in order to determine whether the scores of the dependent variable differ in terms of a categorical independent variable, studies (Behizadeha & Engelhard, 2014; Güler & İlhan, 2017; İlhan & Güler, 2017; Ricketts, Engelhard ve Chang, 2015) have been conducted using the MFRM.



included in the study. An examination of [Figure 2](#) shows that the scale values for this study are reported to be in  $\pm 1$  range. The second column of the logit map corresponds to the stimulants. There is a distribution in this column from the top to the bottom extending between the most significant undesirable influence of high stake exams and relatively less significant influences of those exams. Accordingly, S1 was considered as the most significant undesirable influence of high stake exams by teachers, while S4 was considered as a problem less severe than the other six undesirable influences. Column three in the logit map shows the measurements for the facet of individuals. The ranking in terms of examinees' ability levels can be seen by looking at this column in studies aiming to measure individuals' levels of ability. Yet, when pairwise comparison data are analysed in MFRM, scores for abilities cannot be found for individuals as in analyses in Excel. For this reason, all the teachers in the column of individuals are at the same point in this logit map. Column four in the logit map shows information on the scale categories used in this study. After the logit map, the measurement reports for the stimulant facet were analysed and the findings obtained are shown in [Table 1](#).

**Table 1.** *Measurement Reports for the Facet of Stimulant.*

Stimulant	Measure	Model S.H	Infit MnSq	Outfit MnSq
S1	.46	.03	1.00	1.00
S6	.11	.03	1.02	1.02
S2	.02	.03	1.00	1.00
S3	-.01	.03	.98	.98
S5	-.14	.03	.97	.96
S7	-.20	.03	1.04	1.05
S4	-.31	.03	.99	.98
Mean	-.01	.03	1.00	1.00
Standard deviation	.25	.00	.02	.03
Chi square = 335.10	$df = 6$	$p = .00$	Separation Index = 7.75	Reliability = .98

As can be seen from [Table 1](#), the potential undesirable influences of high stake exams were differentiated significantly by the teachers [ $\chi^2_6 = 335.10, p < .001$ ]. Having separation ratio above 2 and reliability coefficient above .80 (Linacre, 2012) indicated that the measures made in the study were reliable. The infit and outfit statistics in [Table 1](#) were found to take on values between .96 and 1.05. The acceptable range for infit and outfit statistics is .5 to 1.5 (Wright & Linacre, 1994). When the number of stimulants seen by the participants as equal is too high, the fit statistics fall below the acceptable range, and in this case, the assumptions about the MFRM are not met and validity issues arise. The fit statistics in the [Table 1](#) are within the recommended range and make a sign that the stimulants ratio seen as equal is not at a size that will negatively affect model-data fit in unfavourable ways. This finding regarding the fit indices provides evidence for the validity of the measurements.

The results of the stimulant presented visually in the logit map are shown numerically in [Table 1](#). Apparently, S1 is the stimulant having the highest scale value (.46) with a considerable difference. Accordingly, the teachers considered the stimulant “school assessments turn into secondary importance in the eyes of students and parents” as the foremost undesirable influence of high stake exams in the process of education, which was followed by the stimulant “it turns students into more vying individuals”. “It causes teachers to teach test-oriented instead of the standards set in the curriculum” ranked the third as an undesirable influence. The stimulant “the course content which is included in high stake exams are prioritized compared to the content of other courses” ranked the fourth and the stimulant “it causes stress in students, teachers and parents” ranks the fifth. The stimulant “it feeds the spirit of searching for a victim for low exam scores and accusing them among educators and policy makers” and the stimulant “educators’

and administrators' focus on policies of boosting exam scores rather than policies of improving the learning-teaching process" ranked the sixth and the seventh, respectively.

#### **4. DISCUSSION and CONCLUSION**

According to the research results, the teachers reported that "school assessments turn into secondary importance in the eyes of students and parents" was the foremost undesirable influence stemming from high stake exams by obvious difference. The clear difference detected means that the primary problem to be resolved regarding the negative effects of high stake exams is that these tests push school assessments to the second plan. Additionally, the fact that this stimulant was placed the first by far can be interpreted as a remarkable consensus among the teachers on the negative effects of these exams. Findings regarding the first stimulant's scale values being significantly different compared to the scale values of other stimulants are in parallel to the ones reported in the literature. Atılgan (2018), in a study on high stake exams administered in transition into the next stages in Turkey historically, points out that schools and curricula have become dysfunctional due to those exams. According to Atılgan (2018), high stake exams have become the goal and schools have become the instrument giving diplomas to achieve the goal in the current system of education. In a similar vein, Can (2017) also states that the great majority of students stated that the primary goal for them was success in the high stake exams and they attended classes in their institutions just to get a diploma. Accordingly, the fact that school assessments are considered as the foremost problem stemming from high stake exams in the eyes of parents and students is a reflection that school assessments are perceived as the tasks which must be fulfilled for high stake exams.

The teachers in the study group put the fact that "it turns students into more vying individuals" in rank two as the undesirable influence of high stake exams. The fact that high stake exams have an almost vital impact on individuals' future lives makes it inevitable that high stake exams trigger rivalry among students. However, the second order among the potential undesirable influences of this rivalry caused by the high stake exams suggests that the psychological impact of these exams on students may be greater than expected. In fact, the vying environment created by high stake exams may make it difficult to transfer the fundamental values such as love for charity, sharing, solidarity, and cooperation (Board of Education and Training, 2017) that the Ministry of National Education of Turkey aims to bring to students.

According to the results obtained in this study, the stimulant "it causes teachers to teach test-oriented instead of the standards set in the curriculum" was ranked the third in the undesirable influences of high stake exams. It was followed by the stimulant "the course content which is included in high stake exams are prioritized compared to the content of other courses" with a scale value very close to the one ranking the third. These stimulants, which ranked the third and fourth, can be regarded as the results of "considering school assessment as of secondary importance" and of "the increase in rivalry between students". More clearly, the fact that the importance students and parents attach to school assessment is shadowed by high stake exams can lead teachers to shape teaching according to high stake exams. Additionally, teachers can choose to plan their teaching according to the test content instead of the curriculum to support their students in the high rivalry environment caused by high stake exams. Thus, which of the stimulant is perceived as the more primary problem stemming from high stake exams can be connected with the cause-effect relationships between the stimulants. This view is supported by the Pareto principle (Jenny, 2007), which argues that priority issues are the cause of other problems with lower priority.

The stimulant "stress caused by high stake exams in students, teachers, and parents" ranked the fifth in the undesirable influences of high stake exams. Accordingly, teachers consider the



affective influences of high stake exams on the stakeholders of education as a problem less important than the effects on learning-teaching process. Yet, it should not be forgotten that the situation might have stemmed from the fact that the study was conducted with the teachers, not with the students or parents. This is because the examinations administered in Turkey form the basis for decisions to be made about students but they are not used in decisions for teachers. Therefore, it is thought that if a study is conducted with students and parents or if such a study is performed in a country where high stake exams influence teachers' wages (Brooke, 2016) and their position (Nichols & Berliner, 2005), the stress caused by exams can be considered as a more important problem. It was found in relevant literature that the stress teachers have due to high stake exams can differ from country/state to country/state. Abrams (2014), for instance, compared the pressure teachers working in and outside Florida were exposed to and found that 80% of the teachers working in Florida felt the pressure caused by high stake exams but that 40% of the teachers working in the other states felt the pressure.

Searching for victims among educators, administrators, and policy makers for low exam scores and accusing them ranked the sixth as an undesirable influence of high stake exams. Administrators' focussing on policies for boosting exam scores instead of policies for improving the learning-teaching process ranked the last. The fact that teachers ranked these two stimulants at the bottom meant that they considered the grade level and school level effects of those exams as more important than the effects on educational policies. This finding was quite different from the one obtained in Adedoyin (2013) who analysed university students' viewpoints on high stake exams in Botswana educational system. It was found that the fact that those exams caused politicians to search for victims for low exam scores and that the exams offered misleading information causing politicians to make inadequate decisions about the process of education were the undesirable influences of high stake exams. On the other hand, it was also found that high stake exams did not have such effects as causing school assessment to lose its importance or causing the subjects included in test content to be prioritised. The great differences between the findings obtained in this study and those in Adedoyin (2013) can be considered as an indication that the effects of high stake exams on the education system vary from country to country.

A review of relevant literature shows that there are studies concluding that the results of high stake exams influence educational policies. For instance, Buyruk (2014) reports that students' achievement in high stake exams is associated with teachers' accomplishment and that the results of those exams are used like instruments informing us of school and teacher performance according to provinces, districts, and schools. Due to this, high stake exams can lead educators and administrators to policies for boosting exam scores to be in a better position in comparisons between schools. Therefore, it would be mistaken to see the results of the study as high stake as exams in Turkey do not have undesirable influences such as searching for victims for low exam scores or educators', administrators', and policy-makers' focusing on policies for boosting exam scores instead of policies for improving the learning-teaching process. The reality is that teachers do not consider these two undesirable influences as primary as the other stimulants.

#### **4.1. Implications for Practice**

This study concentrates on seven undesirable influences of high stake exams. However, the undesirable influences are not restricted to the ones considered in this study. There are several undesirable influences mentioned in the literature that affect all the stakeholders of education such as students, teachers, and administrators. In this context, the undesirable influences of high stake exams should be revised again in the light of scientific studies performed. The attention of stakeholders who can make regulations to reduce the undesirable influences on the educational system should be called to the problems that are considered more primary. Such

exams should no longer be the turning points for individuals so that the undesirable influences could be minimised. Schools should be varied and the differences between schools arising from physical conditions (the number of students per classroom and teacher and specially equipped learning environments such as laboratory, library, and gymnasium), artistic, sporting, cultural and social activities, social-economic conditions of the school district, and teacher qualifications should be reduced, and the meaning attached to exams should be minimized without getting away from the fact that a system without exams is impossible at the moment.

#### 4.2. Future Directions

The first proposal that could be brought into the scope of the study is that the researchers would prefer to use MFRM instead of the traditional psychometric approach when scaling with pairwise comparisons. Since this study aimed to show that paired comparison data can be analyzed with MFRM, MFRM and traditional method comparison was not performed. Another research proposal that can be brought in this context is to test the agreement between the scale values obtained from the paired comparisons performed according to the MFRM and the traditional method. Finally, this study was conducted with a study group of 191 teachers that was not very large. Therefore, it may be suggested that a similar study be conducted with different samples in order to increase the generalizability of the findings obtained in the study.

#### Acknowledgments

The author(s) received no financial support for the research, authorship, and/or publication of this article. The first draft of the paper was presented at VI<sup>th</sup> International Congress on Measurement and Evaluation in Education and Psychology, Prizren, Kosova.

#### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

#### Authorship Contribution Statement

**Mustafa İlhan:** Investigation, Resources, Methodology, Analysis, Writing the original draft.

**Neşe Güler:** Investigation, Resources, Supervision, Writing the original draft.

**Gülşen Taşdelen Teker:** Investigation, Resources, Methodology, Writing the original draft.

#### ORCID

Mustafa İLHAN  <https://orcid.org/0000-0003-1804-002X>

Neşe GÜLER  <https://orcid.org/0000-0002-2836-3132>

Gülşen TAŞDELEN TEKER  <https://orcid.org/0000-0003-3434-4373>

#### 5. REFERENCES

- Abrams, L. M. (2004). *Teachers' views on high-stakes testing: Implications for the classroom*. Arizona: Policy Brief: Education Policy Studies Laboratory. Arizona State University College of Education. <http://epsl.asu.edu/epru/documents/EPSSL-0401-104-EPRU.pdf>
- Acar Güvendir, M. (2014). Student and school characteristics' relation to Turkish achievement in student achievement determination exam. *Education and Science*, 39(172), 163–180. <http://egitimvebilim.ted.org.tr/index.php/EB/article/view/2839>
- Adedoyin, O. O (2013). Public examinations and its influence on the Botswana educational system: Views of undergraduate education students at the University of Botswana. *Asian Journal of Humanities and Social Sciences*, 1(2), 124–134. <https://ajhss.org/pdfs-1/Public%20Examinations%20and%20its.....pdf>

- Almus, K. (2010). *The beliefs of principals and assistant principals regarding high-stakes testing*. [Doctoral dissertation, University of Houston, Houston, United States]. <https://uh-ir.tdl.org/handle/10657/ETD-UH-2010-12-78>
- Amoako, I. (2019). What's at stake in high-stakes testing in Ghana: Implication for curriculum implementation in basic schools. *International Journal of Social Sciences & Educational Studies*, 5(3), 72–82. <https://doi.org/10.23918/ijsses.v5i3p72>
- Amrein, A. L., & Berliner, D. C. (2003). The effects of high-stake testing on student motivation and learning. *Educational Leadership*, 60(5), 32-38. [http://www.wou.edu/~girodm/611/testing\\_and\\_motivation.pdf](http://www.wou.edu/~girodm/611/testing_and_motivation.pdf)
- Anıl, D., & Güler, N. (2006). An example of the scaling study by pair-wise comparison method. *Hacettepe University Journal of Education*, 30, 30-36. <http://www.efdergi.hacettepe.edu.tr/yonetim/icerik/makaleler/722-published.pdf>
- Assaf, L. Z. (2008) Professional identity of a reading teacher: Responding to high-stakes testing pressures. *Teachers and Teaching: Theory and Practice*, 14(3), 239-252. <https://doi.org/10.1080/13540600802006137>
- Atılğan, H. (2018). Transition among education levels in Turkey: Past-present and a recommended model. *Ege Journal of Education*, 19(1), 1-18. <https://doi.org/10.12984/egeefd.363268>
- Banks, J., & Smyth, E. (2015). Your whole life depends on it': academic stress and high-stakes testing in Ireland. *Journal of Youth Studies*, 18(5), 598-616. <http://dx.doi.org/10.1080/13676261.2014.992317>
- Baykal, A. (2014). *Sınavlardan sınav beğen*. *Eğitim sisteminde kademeler arası geçiş ve sınavlar* [Like the exam from the exams. Transition between stages and exams in the education system]. Ege'den Eğitime Bakış Paneli, Ege Üniversitesi, İzmir. [https://www.academia.edu/11610273/SINAVLARDAN\\_SINAV\\_BE%C4%9EEN](https://www.academia.edu/11610273/SINAVLARDAN_SINAV_BE%C4%9EEN)
- Behizadeha, N., & Engelhard, G. (2014). Development and validation of a scale to measure perceived authenticity in writing. *Assessing Writing*, 21, 18-36. <https://doi.org/10.1016/j.asw.2014.02.001>
- Board of Education and Training. (2017). *Our renewal and change studies in curriculum*. [https://tkb.meb.gov.tr/meb\\_iys\\_dosyalar/2017\\_07/18160003\\_basin\\_aciklamasi-program.pdf](https://tkb.meb.gov.tr/meb_iys_dosyalar/2017_07/18160003_basin_aciklamasi-program.pdf)
- Brady, A. L. (2008). *Effects of standardized testing on teachers' emotions, pedagogy and professional interactions with others*. [Doctoral dissertation, Cleveland State University, Cleveland, United States]. <https://engagedscholarship.csuohio.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1038&context=etdarchive>
- Brockmeier, L. L., Green, R. B., Pate, J. L., Tsemunhu, R. T., & Bochenko, M. J. (2014). Teachers' beliefs about the effects of high stakes testing. *Journal of Education and Human Development*, 3(4), 91-104. <http://dx.doi.org/10.15640/jehd.v3n4a9>
- Brooke, N. (2016). High-stakes accountability using teacher salary incentives in Brazil: An update. *Profesorado, Revista de Currículo y Formación del Profesorado*, 20(3), 207–250. <https://recyt.fecyt.es/index.php/profesorado/article/view/54598>
- Bülbül, T., & Acar, M (2012). A pair-wise scaling study on the missions of education supervisors in Turkey. *International Journal of Human Sciences*, 9(2), 623–640. <https://www.j-humansciences.com/ojs/index.php/IJHS/article/viewFile/2327/941>
- Buyruk, H. (2014). Standardized examinations as a teacher performance indicator and performance evaluation in education. *Trakya University Journal of Education*, 4(2), 28–42. <http://dergipark.gov.tr/trkefd/issue/21480/230201>
- Can, E. (2017). Determination of the effects of central exams according to the view of students. *The journal of Academic Social Science*, 5(58), 108-122. <http://dx.doi.org/10.16992/ASOS.12842>

- Çetin, A., & Ünsal, S. (2018). Social, psychological effects of central examinations on teachers and their reflections on teachers' curriculum implementations. *Hacettepe University Journal of Education*, 34(2), 304-323. <http://dx.doi.org/10.16986/HUJE.2018040672>
- Christian, S. C. (2010). *High-stakes testing and its relationship to stress levels of secondary teachers*. [Doctoral dissertation, The University of Southern Mississippi, Hattiesburg, United States]. <https://aquila.usm.edu/dissertations/932/>
- Cohen, L., Manion, L., & Morrison, K. (2007). *Research Methods in Education (6th ed.)*. Routledge Falmer.
- Coniam D., & Falvey P. (2007) High-stakes testing and assessment. In J. Cummins J., & C. Davison (Eds.), *International handbook of English language teaching* (pp. 456-471). Springer.
- Davis, M. F. (2011). *The influence of high-stakes testing on science teacher perceptions and practices*. [Doctoral dissertation, Walden University, Minneapolis, United States]. <https://eric.ed.gov/?id=ED528115>
- Dawson, H. S. (2012). *Teachers' motivation and beliefs in a high-stakes testing context*. [Doctoral dissertation, The Ohio State University, Ohio, United States]. [https://etd.ohiolink.edu/apexprod/rws\\_etd/send\\_file/send?accession=osu1338399669&disposition=inline](https://etd.ohiolink.edu/apexprod/rws_etd/send_file/send?accession=osu1338399669&disposition=inline)
- Ekinci, A., Bindak, R., & Yıldırım, M. C. (2012). A research regarding the empathic approaches of school managers about professional problems of teachers by pair-wise comparisons method. *Gaziantep University Journal of Social Sciences*, 11(3), 759-776. <http://dergipark.gov.tr/download/article-file/223316>
- Erkuş, A. (2011). *Scientific research process for behavioral sciences*. Seçkin.
- Finkeldei, J. (2016). *The influence of high stakes testing on elementary classroom instruction*. [Doctoral dissertation, Wichita State University, United States]. [https://soar.wichita.edu/bitstream/handle/10057/12632/d16008\\_Finkeldei.pdf?sequence=1](https://soar.wichita.edu/bitstream/handle/10057/12632/d16008_Finkeldei.pdf?sequence=1)
- Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series*, 2(2), 1-17. <https://doi.org/10.1002/ets2.12067>
- Fraenkel, J. R., Wallen, N. E. & Hyun, H. H. (2012). *How to design and evaluate research in education*. McGraw-Hill.
- Güler, N., & Anıl, D. (2009). Scaling through pair-wise comparison method in required characteristics of students applying for post graduate programs. *International Journal of Human Sciences*, 6(1), 627-639. <https://www.j-humansciences.com/ojs/index.php/IJHS/article/view/673>
- Güler, N., & İlhan, M. (2017, October). *Comparison of the results obtained from t-test and ANOVA, and many facet Rasch analysis for difference based statistics*. Paper presented at the 13th International Conference on Social Sciences, Vienna.
- Güzeller, C. O., & Eser, M. T., & Aksu, G. (2016). Pair-wise comparison method application via R project and Microsoft Excel. *Journal of Measurement and Evaluation in Education and Psychology*, 7(1), 96-108. <http://dx.doi.org/10.21031/epod.80072>
- Hirsch, P. J. (2016). *High stakes testing and its effect on teacher methodologies*. [Master's thesis, Caldwell University].
- İlhan, M. (2016). An analysis of researchers' difficulties in quantitative data analysis with the use of pairwise comparisons. *Journal of Measurement and Evaluation in Education and Psychology*, 7(1), 73-84. <http://dx.doi.org/10.21031/epod.28154>
- İlhan, M., & Güler, M. (2017). The use of Rasch model in Likert types scales: An application on the fear of negative evaluation scale-Student form (FNE-SF). *Trakya Journal of Education*, 8(4), 756-775. <http://dx.doi.org/10.24315/trkefd.357367>
- Johnson, P. (2007). *High stakes testing and No Child Left Behind: Conceptual and empirical considerations*. Long Island Economic & Social Policy Institute, Dowling College School



- of Education, Long Island, NY. [http://martincantor.com/files/High%20Stakes\\_LIESPW\\_hitepaperMay4.pdf](http://martincantor.com/files/High%20Stakes_LIESPW_hitepaperMay4.pdf)
- Kane, G. (2014). *Accelerating sustainability using the 80/20 rule*. Oxford: Do Sustaniablity.
- Knapp, D. (2010). *A guide to service desk concepts*. Course Technology, Cengage Learning.
- Kornhauser, Z. G. C., Minahan, J., Siedlecki, K. L., & Steedle, J. T. (2014, April). *A strategy for Increasing student motivation on low-stakes assessments*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia. <https://files.eric.ed.gov/fulltext/ED582123.pdf>
- Kruger, L. J., Wandle, C., & Struzziero, J. (2007). Coping with the stress of high stakes testing. *Journal of Applied School Psychology, 23*(2), 109-128, [https://doi.org/10.1300/J370v23n02\\_07](https://doi.org/10.1300/J370v23n02_07)
- Kumandaş, H., & Kutlu, Ö. (2015). High stake tests. *Journal of Educational Science Research, 5*(2), 63–75. <http://dx.doi.org/10.12973/jesr.2015.52.4>
- Linacre, J. M. (2012). *Many-facet Rasch measurement: Facets tutorial2*. <http://www.winsteps.com/a/ftutorial2.pdf>
- Linacre, J. M. (2014). *A user's guide to FACETS Rasch–model computer programs*. <https://www.winsteps.com/facetman/pairedcomparisons.htm>
- Marchant, G. J. (2004). What is at stake with high stakes testing? A discussion of issues and research. *The Ohio Journal of Science, 104*(2), 2-7. <https://pdfs.semanticscholar.org/ca83/c29a37b1389c737a35b3454808bacaf97128.pdf>
- Minarechová, M. (2012). Negative impacts of high–stakes testing. *Journal of Pedagogy, 3*(1), 82–100. <https://doi.org/10.2478/v10159-012-0004-x>
- Moses, M. S., & Nanna, M. J. (2007). The testing culture and the persistence of highstakes testing reforms. *Education and Culture, 23*(1), 55-72. <https://doi.org/10.1353/eac.2007.0010>
- Nalbantoğlu Yılmaz, F. (2017). Investigation of the factors affecting occupation choices of high school students with paired comparison method. *Journal of Measurement and Evaluation in Education and Psychology, 8*(2), 224–236. <http://dx.doi.org/10.21031/epod.303882>
- Nartgün, Z. (2006). Öğretmenlik meslek bilgisi derslerinin önem düzeyinin ikili karşılaştırmalarla ölçeklenmesi [Scaling the importance level of teaching profession courses using paired comparisons]. *Abant İzzet Baysal University Journal of Faculty of Education, 6*(2), 161–176. <http://efdergi.ibu.edu.tr/index.php/efdergi/article/view/964>
- Nichols, S. L., & Berliner, D. C. (2005). *The inevitable corruption of indicators and educators through high–stakes testing*. Education Policy Research Unit (EPRU), Education Policy Studies Laboratory, Arizona State University. <https://files.eric.ed.gov/fulltext/ED508483.pdf>
- Pbrreault, G. (2000). The classroom impact of high-stress testing. *Education, 120*(4), 705-710.
- Phelps, R. P. (2003). *Kill the messenger: The war on standardized testing*. New Brunswick, Transaction.
- Polesel, J., Dulfer N., & Turnbull, M. (2012). *The experience of education: The impacts of high stakes testing on school students and their families*. Sydney: Whitlam Institute Report, University of Western Sydney, [https://www.researchgate.net/publication/265752469\\_The\\_Experience\\_of\\_Education\\_The\\_impacts\\_of\\_high\\_stakes\\_testing\\_on\\_school\\_students\\_and\\_their\\_families\\_Literature\\_Review](https://www.researchgate.net/publication/265752469_The_Experience_of_Education_The_impacts_of_high_stakes_testing_on_school_students_and_their_families_Literature_Review)
- Ricketts, S.N., Engelhard, G., & Chang, M.L. (2015). Development and validation of a scale to measure academic resilience in mathematics. *European Journal of Psychological Assessment, 33*(2), 79–86. <https://doi.org/10.1027/1015-5759/a000274>
- Ritt, M. (2016). *The impact of high–stakes testing on the learning environment*. [Masters' thesis, St. Catherine University/University of St. Thomas, United States of America]. [https://sophia.stkate.edu/cgi/viewcontent.cgi?article=1660&context=msw\\_papers](https://sophia.stkate.edu/cgi/viewcontent.cgi?article=1660&context=msw_papers)



- Saito, Y. (2006). Consequences of high stakes testing on the family and schools in Japan. *KEDI Journal of Educational Policy (KJEP)*, 3(1), 101–102.
- Segool, N. K., Carlson, J. S., Goforth, A. N., von der Embse, N., & Barterian, J. A. (2013). Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing. *Psychology in Schools*, 50(5), 489-499. <https://doi.org/10.1002/pits.21689>
- Shepard, L., & Dougherty, K. C. (1991, April). *Effects of high-stakes testing on instruction*. Paper presented at the annual meeting of the American Educational Research Association, Chicago. <https://files.eric.ed.gov/fulltext/ED337468.pdf>
- Simpson, C. (2016). *Effects of standardized testing on students' well-being*. [https://projects.iq.harvard.edu/files/eap/files/c.\\_simpson\\_effects\\_of\\_testing\\_on\\_well\\_being\\_5\\_16.pdf](https://projects.iq.harvard.edu/files/eap/files/c._simpson_effects_of_testing_on_well_being_5_16.pdf)
- Spann, P., & Kaufman, D. (2015). *The negative effects of high-stakes testing*. <https://www.luc.edu/media/lucedu/law/centers/childlaw/childed/pdfs/2015studentpapers/Spann.pdf>
- Stanley, G. K. (2004). High stakes hustle: Public schools and the new billion dollar accountability. *The Educational Forum*, 69(1), 8-16. <https://doi.org/10.1080/00131720408984660>
- Stecher, B. M. (2002). Consequences of large-scale, high-stakes testing on school and classroom practice. In L. S. Hamilton, B. M. Stecher, & S. P. Klein (Eds.). *Making sense of test-based accountability in education* (pp. 79-100). RAND Corporation. <https://larrycuban.files.wordpress.com/2012/01/mr1554-ch4.pdf>
- Stobart, G., & Eggen, T. (2012). High-stakes testing – value, fairness and consequences. *Assessment in Education: Principles, Policy & Practice*, 19(1), 1-6. <https://doi.org/10.1080/0969594X.2012.639191>
- Taylor, G., Sheppard, L., Kinner, F., & Rosenthal, J. (2003). *A survey of teachers' perspectives on high-stakes testing in Colorado: What gets taught, what gets lost*. (CSE Tech. Rep. 588). Center for Research on Evaluation, Standards, and Student Testing. <https://files.eric.ed.gov/fulltext/ED475139.pdf>
- Togut, T. D. (2004). High stakes testing: *Educational barometer for success, or false prognosticator for failure*. <https://www.harborhouselaw.com/articles/highstakes.togut.htm>
- Turgut, M. F., & Baykul, Y. (1992). *Ölçekleme Teknikleri [Scaling techniques]*. ÖSYM.
- Vogler, K. E., & Virtue, D. (2007). “Just the Facts, Ma’am”: Teaching social studies in the era of standards and high-stakes. *Testing, The Social Studies*, 98(2), 54–58. <https://doi.org/10.3200/TSSS.98.2.54-58>
- Westfall, D. M. (2010). *Parental perceptions of the effects of the high-stakes TAKS test on the home lives of at-risk fifth grade students*. [Doctoral dissertation, University of Houston, Houston, Texas]. <https://uh-ir.tdl.org/handle/10657/170>
- Winters, M. A., Trivitt, J. R., & Greene, J. P. (2010). The impact of high-stakes testing on student proficiency in low-stakes subjects: Evidence from Florida's elementary science exam. *Economics of Education Review*, 29(1), 138-146. <https://doi.org/10.1016/j.econedurev.2009.07.004>
- Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10(1), 1-17. [https://doi.org/10.1207/s15326977ea1001\\_1](https://doi.org/10.1207/s15326977ea1001_1)
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8, 370. <http://www.rasch.org/rmt/rmt83b.htm>

- Wright, W. E. (2002). The effects of high stakes testing in an inner-city elementary school: The curriculum, the teachers, and the English language learners. *Current Issues in Education*, 5(5). <https://cie.asu.edu/ojs/index.php/cieatasu/article/view/1622/663>
- Yaşar, M. (2018). Scaling of ideal teachers characteristics with pairwise comparison judgments according to pre-service teachers' opinions. *International Journal of Assessment Tools in Education*, 5(1), 130–145. <https://doi.org/10.21449/ijate.369233>
- Yeh, S. S. (2005). Limiting the unintended consequences of high-stakes testing. *Education Policy Analysis Archives*, 13(43). <http://dx.doi.org/10.14507/epaa.v13n43.2005>
- Yılmaz, S. (2017). *Evaluating the reflections of high stakes tests on school culture*. [Master's thesis, Trakya University].

## The Views of Pre-Service Elementary Teachers About Online and Traditional Peer Assessment

Ahmet Oguz Akcay <sup>1,\*</sup>, Ufuk Guven <sup>2</sup>, Engin Karahan <sup>3</sup>

<sup>1</sup>Eskişehir Osmangazi University, Faculty of Education, Department of Elementary and Early Childhood Education, Eskişehir

<sup>2</sup>Düzce University, Faculty of Education, Department of Elementary and Early Childhood Education, Düzce

<sup>3</sup>Eskişehir Osmangazi University, Faculty of Education, Department of Curriculum and Instruction, Eskişehir

### ARTICLE HISTORY

Received: July 01, 2020

Revised: Feb. 16, 2021

Accepted: Apr. 09, 2021

### Keywords:

Online peer assessment,  
Elementary Teacher  
Education,  
Traditional peer-assessment

**Abstract:** The goal of this study is to compare traditional peer evaluation and online peer evaluation in order to identify which method is more effective in evaluating peers. Qualitative research method was used in this study to understand pre-service teachers' opinions on different peer evaluation techniques. The study was carried out in a state university in Turkey. The sample consisted of 58 second year pre-service teachers majoring in primary school teacher program who enrolled in "Instructional Technologies and Material Development" course. Pre-service teachers were divided into 11 groups, with five or six students in each group. Participation was voluntary and the students in each group actively participated in the traditional and online peer assessment activities. The analyses of the data were done via content analysis, by creating categories and then themes. The themes that emerged as a result of the analysis of the data collected within the study were (1) objectivity, (2) evaluation criteria, (3) interaction, and (4) attributes of the online evaluation platform. The study concluded that a combination of peer and instructor evaluation and even self-assessment can give a better validity and objectivity of assessment.

## 1. INTRODUCTION

Every teaching and learning process contains a strategic assessment system because assessment is an essential part of every instructional method (Taras, 2005). Assessment informs students about their learning level and advises teachers about student performance, instructional methods, and areas where students need more help to better understand subjects (The National Council for Teacher Education [NCTE], 2013). Individual, peer or group assessment, projects, and tests often enhance the learning experiences of students. Assessment also provides feedback to students and teachers. With these feedbacks, students may correct their knowledge on a subject (Roediger et al., 2011). Moreover, teachers have an important role in organizing the measurement and evaluation, planning of when the work to be done, determining how to use

---

\*CONTACT: Ahmet Oğuz Akçay ✉ [aoguzakcay@gmail.com](mailto:aoguzakcay@gmail.com) 📍 Eskişehir Osmangazi University, Faculty of Education, Department of Elementary and Early Childhood Education, Eskişehir, Turkey

the obtained data, and encouraging participation of students in classroom works (Özenç & Çakır, 2015). Kampen (2020) stated that there are six types of assessments: diagnostic, formative, summative, ipsative, norm-referenced, and criterion-referenced assessments; however, teachers usually use summative and formative assessment methods only in their classroom. Bhat and Bhat (2019) defined summative assessments as assessment of learners that has a main goal of measuring the outcome of a curriculum. They are used to evaluate learning level, skill acquisition and achievement level of an intervention, which can be a project, course, workshop, program, or an academic year. Formative assessment is part of instructional process, and it is used to get accurate feedback about students' learning during the teaching process and to arrange teaching methods. In addition, the formative assessment focuses on organizing and improving students' learning in the process, helping students find the answer to the question, "How do I learn?" Exit slips, projects, homeworks, question-answer technique, summarization, concept maps, quizzes, criteria and goal setting, observations, self and peer assessment, and student record keeping are some of the instructional strategies that can be used for formative assessment (Kampen, 2020).

Peer assessment is one of the methods that use constructivist approach, where the student is responsible for his/her own learning and the teacher plays the role of a guide or facilitator/organizer of activities who support the student rather than being a transferrer of knowledge in the teaching-learning process. In the constructivist approach, the teacher is expected to use different methods, techniques, and technologies to assist students in structuring information, as well as various assessment and evaluation tools to enhance their learning and understanding (Şahin & Kalyon, 2018). Hence, measurement and evaluation have a very important role for students' learning. An alternative assessment measures applied proficiency instead of student knowledge. Portfolios, project work, and other assessments that require a form of rubric are typical examples of alternative assessment (Bradley, 2020). Self-assessment and peer assessment are also types of alternative assessment. Self-assessment is the process where students make evaluations about their own learning and products (Brown & Harris, 2014). Peer assessment methods are widely used in classrooms. Peer assessment is the process of providing formative or summative feedback to their peers about their work (Chin, 2007). In the peer assessment, one or more individuals in a group evaluate their peer(s) and students take responsibility for peer assessment and actively participate in the learning process. Students evaluate their peers' work and performance using pre-defined criteria. Moreover, students see each other as resources for understanding and checking quality work against previously established criteria (Garrison & Ehringhaus, 2007).

While there might be some classroom management issues during the implementation of peer assessment, there are important advantages of peer assessment such as installing autonomy in learners, empowering learners in a learning environment, developing learners' confidence in assessment through practice, activating learners on self-evaluation and reflection, greater understanding of what is required by teachers in assessment, creating an interactive classroom environment, improving information and understanding, providing a clear and open marking system, and creating an effective way to assess a large amount of students' work and provide specific feedback (Langan & Wheeler, 2003),

As with all the other assessment methods, peer assessment also has its own disadvantages. The main disadvantage of peer evaluations is the non-objective evaluation of peers due to personal relationships and peer pressure. Moreover, it is really hard to keep the reliability and validity of peer assessment at an acceptable level (Ashenafi, 2019). Ashenafi (2019) also sees validity and reliability issues of peer assessments as a barrier for teachers to implement this strategy more often. Peer pressure and the possibility of affecting personal relationships are especially common in traditional evaluations, and they are barriers to implement peer evaluation method.

On the other hand, with the advancements in educational technologies, teachers can use online peer assessment methods to avoid barriers that exist in traditional peer evaluation methods.

We are living in the technology era, so web-based or online peer assessment is also valuable to inform learners about their learning. While there are similarities between traditional and online peer assessment methods, there are also some differences. In both peer assessment methods, students assess their peers and provide feedback based on a rubric or a pre-defined standards. Students provide face-to-face feedback and assessment in traditional peer assessment, which may affect the quality and objectivity of the assessment because of the relationships between students. On the other hand, online peer assessment is made through web 2.0 tools and smartphone apps that anonymize students' names; thus, students can provide objective feedback because they do not feel any pressure from other students. Wen and Tsai (2008) stated the importance of online peer assessment in helping the learner pursue learning. Falchikov (2001) highlighted the importance of online peer assessment as,

In peer assessment, members of a class grade the work or performance of their peers using relevant criteria ... In peer feedback, students engage in reflective criticism of the work or performance of other students using previously identified criteria and supply feedback to them ... In peer learning, students learn with and from each other, normally within the same class or cohort ... (pp. 2–3).

Moreover, students take responsibility for peer assessment and participate actively in the learning process (Ndoye, 2017). This is valid for both traditional and online peer assessment methods. It is well-known that students like to use their phones in classroom activities. Online peer assessment can be done through web 2.0 tools and smartphone apps so that students are eager to participate in assessment processes with their devices. Online peer assessment tools enable students not only to grade their peer's work but also to provide feedback. These features are easily applicable in traditional peer assessment but some tweaking is required to utilize these features in online format. Online peer assessment tools also anonymize student names that removes peer pressure while assessing peer's work. This is very hard in traditional peer assessment. The goal of this study is to analyze traditional peer evaluation and online peer evaluation based on students' views.

## **2. METHODOLOGY**

Qualitative research method was used in this study to understand pre-service teachers' opinions on different peer evaluation techniques. Qualitative research seeks to understand phenomena in context-specific settings, such as "real world setting [where] the researcher does not attempt to manipulate the phenomenon of interest" (Patton, 2002, p. 39).<sup>[SEP]</sup>The intent of conducting qualitative research is to explore human behaviors within the natural context in which it occurs (Hatch, 2002) and to focus on process and meaning (Merriam, 1998). Hence, the study investigated the opinions of the participants derived from their experiences within the context of a semester-long class.

### **2.1. Participants**

The study was carried out in a public university in Turkey. The sample consisted of 58 second year pre-service teachers (45 girls, 13 boys) majoring in primary school teacher program who enrolled in the "Instructional Technologies and Material Development" course. Pre-service teachers were divided into 11 groups, with five or six students in each group. Participation was voluntary and the students in each group actively participated in the traditional and online peer assessment activities. Among the participants of the study, six students who worked in different groups were randomly selected in order to conduct the semi-structured interviews.



## 2.2. Peer Assessment Procedure

The aim of the "Instructional Technologies and Material Development" course is to introduce the characteristics of various instructional technologies, their importance and use in the teaching and learning environment, the development of instructional materials, and the evaluation of materials of varied qualities. The instructional process of this course was organized with instructional methodologies and instructional materials. In this course, pre-service elementary teachers were asked to prepare materials based on given primary school level's standards. The teacher candidates prepared their materials in groups. In the evaluation process that lasted for six weeks, the groups presented the materials they prepared while the other groups evaluated the presenting group according to the given criteria: (1) expediency, (2) educational and pedagogical value, (3) promoting motivation and engagement, (4) user friendliness, (5) robustness and durability, (6) portability, (7) adaptability, and (8) design based on material principles. Each group can receive one to three points (low, medium, high) for each criterion, and the maximum total score a student can get was 24 points. The course was conducted face to face, but the evaluation process was implemented as one week face to face and one week online respectively during the six weeks. In addition, each group was required to give an oral presentation in class and upload pictures and videos of materials they prepared for the online platform.

## 2.3. Data Collection Tools

Data can be obtained from different sources, like observations and interviews, in qualitative research method (Yıldırım & Şimşek, 2011). The data collection tools in this present study include online assessment forms (OAF), observations and semi-structured interviews. Students were required to develop an instructional material that can be used to teach a curriculum standard. Students presented their materials to the whole class. In OAF, students were asked to give a grade for their peers' material based on pre-defined evaluation criteria and write comments if they desire. Students' identities were kept anonymous in online peer assessment activity. Only instructors were able to see peer grades. Instructors also observed students in face-to-face peer assessment and online assessment activities. For semi-structured interviews, six open-ended questions were asked to six students to deeply understand the differences between traditional peer evaluation and online peer evaluation.

## 2.4. Data Analysis

Content analysis method was applied in the analysis of the data. In this process, the categories and themes that emerged with the coding of the data were interpreted. In this study, content analysis was used in four stages of processing qualitative research data from documents: 1) coding of the data, 2) finding themes, 3) editing codes and themes, and 4) identification and interpretation of the findings (Yıldırım & Şimşek, 2011). Miles and Huberman (1994) explained how to provide intercoder reliability as follows: "Check coding not only aids definitional clarity but also is a good reliability check...The best advice here is for more than one person to code, separately" (p. 64). The coding data process was completed by three researchers separately to determine whether reliability and consistency were achieved. The congruity among these three code sets were higher than 80%. Also, the interview and observation data were examined to see if they supported each other to improve the validity and reliability of the study (Yıldırım & Şimşek, 2011). In order to support or disprove the validity of the analysis process, codes that appeared within one data source were considered with other data sources, effectively triangulating the code against multiple data sources. Additionally, the level of transparency was increased by providing rich and thick descriptions that allow readers to draw their own conclusions.

### 3. RESULTS

The themes that emerged as a result of the analysis of the data collected within the study were (1) objectivity, (2) evaluation criteria, (3) interaction, and (4) attributions of the online evaluation platform. Theme and code list used in the research are shown in [Table 1](#).

**Table 1.** Theme and Code List.

Themes	Code
Objectivity	<ul style="list-style-type: none"> <li>● The effect of personal relationships on peer review</li> <li>● The effect of the anonymous answering system</li> <li>● Consistency in evaluation</li> </ul>
Evaluation criteria	<ul style="list-style-type: none"> <li>● Evaluation criteria</li> </ul>
Interaction	<ul style="list-style-type: none"> <li>● Peer pressure</li> <li>● Instant interaction</li> <li>● Face-to-face communication</li> </ul>
Attributions of the online evaluation platform	<ul style="list-style-type: none"> <li>● Positive attributions</li> <li>● Negative attributions</li> </ul>

#### 3.1. Objectivity

##### 3.1.1. The Effect of Personal Relationships on Peer Review

When the data collected from the participants of the study were examined, objectivity emerged as one of the points underlined primarily. Participants insistently emphasized their advantages and disadvantages in terms of objectivity in the online evaluation processes. Unfair evaluation was discussed by many participants in different ways. For example, they stated that due to the competition among the participants, they gave one another lower scores than usual and it negatively affected the fairness of the evaluation.

*Other groups were unfair. This is because they want the highest score for their group (OAF).*

*It was unfair because some of them deliberately gave high scores to other groups. Others deliberately gave low scores to other groups just to be the winner (OAF).*

*Some groups gave high scores for other groups if they have given them high regardless of whether the material met the criteria (Interview-Student 2).*

Another participant argued that competition among students poses an obstacle to the fairness of students.

*We can say that these evaluations depend on the conscience of the group members. I don't think everyone is fair. Sometimes they are competitive and not fair (OAF).*

*Some friends gave low scores to other groups in order to be the first (Interview- Student 5)*

Another point that should be emphasized under the theme of objectivity is that the participants do not find themselves and their friends ready for peer assessment. Many participants stated that during the evaluation, the students gave unfair scores by being negatively affected by the criticism.

*The evaluations were absolutely unfair. Negative reviews and low scores were given against the negative reviews. To get higher scores, the materials were evaluated less than their value (OAF).*

One of the emphases that came to the fore in the point of objectivity was the highlight of friendships in the assessment. While the participants stated that good friends gave each other high scores, they also thought that their friends may be offended when they gave low scores.

---

*I mostly think about that problem. I was torn between giving low scores or taking risk to offend my friends by giving them low scores. But still, I tried to be fair while giving scores (OAF).*

Another point that drew attention to objectivity in the evaluation was that each group attempted to perform a fair evaluation process by giving the same points given to them or by giving a similar score to all groups. On the other hand, this approach has been severely criticized by the participants. For example, one of the participants stated that the same points they gave to the groups were given to them, so the teacher evaluation would be more accurate than the peer evaluation.

*I do not think it is a fair evaluation by any means. Each group gave us the same score that we gave them. No group's material reviews were fair. It would be more appropriate for the teacher to evaluate it, not the students (OAF).*

### **3.1.2. The Effect of the Anonymous Answering System**

While participants highlighted the role of personal relationships in the evaluation process, they pointed to the anonymous evaluation feature in the online evaluation platform as a solution to this. Accordingly, many participants shared that they did not feel peer pressure during the evaluation process, and they performed their evaluations more comfortably. Therefore, the groups who did not see the scores given to them defended that they gave the other groups the points they deserved.

*Online environment was more effective because nobody felt under pressure when evaluating or we didn't think like "the other group gave me this and I'll give the same score" (OAF).*

*Since we did not see who gave us how many points, we gave them the points we think they deserve (OAF).*

It was stated that anonymous evaluation not only provides objectivity but also enables the participants to carry out more comfortable evaluations, wherein the participants share their opinions more freely and comprehensively. Another important point emphasized by the participants was that the anonymous answering system ensures confidentiality between the teacher and the student. In this way, it was demonstrated that the participants performed a fairer and honest assessment.

*Online environment. Because the answers and thoughts of the people remain confidential between the teacher and the student. Questions can be answered more honestly and undoubtedly (OAF).*

*It was observed that students gave more honest answers in the anonymous evaluation, because the students knew that their score was only visible to the instructor (Observation).*

According to the participants, another advantage of the anonymous evaluation is to prevent conflicts and communication problems that may arise in the classroom as a result of negative evaluation. In this way, it was shared that the participants were able to make negative evaluations for their friends without worrying about any trouble.

*Online environment is more appropriate in this regard because it is not appropriate to use a hard language in the classroom while criticizing a material. If it is not liked, it should be expressed appropriately, but thanks to this application, low scores can be given as desired (OAF).*

While the participants highlighted many advantages of anonymous evaluation, some stated that keeping the identities hidden during the evaluation process could negatively affect the objectivity of the evaluations.

*I think face-to-face evaluation is more objective because people gave random points as anyone didn't see what score he/she got in the Online environment (OAF).*

Similarly, most of the participants stated that the social pressure felt in face-to-face evaluation contributed positively to the objectivity of the evaluations. Thus, the participants emphasized that face-to-face interaction-based evaluation is much fairer and more realistic compared to the anonymity feature of the online evaluation system.

*I find the face-to-face assessment more objective and realistic because, in Online environment, undeserved scores can be given because people give scores over the internet, it is difficult to control this (OAF).*

Finally, the participants stated that the evaluation process could not be carried out regardless of the evaluation environment and method, and they identified the competition among students as the main reason for this.

*Neither of them was objective. Because my dear classmates gave low scores to everyone as if they were in the competition program (OAF).*

*I don't think they are objective in either of them, because everyone scored low each other to be 1st and no one's work was taken into account (OAF).*

### **3.1.3. Consistency in Evaluation**

One of the important points that emerged in the opinions of the participants about objectivity was the inconsistencies they observed during the evaluation process. In the evaluation of the same material, the score differences obtained from different groups emerged as an important criticism in this process. As can be seen from the examples given below, when the participants examined the scores given to themselves and the other groups, they emphasized that the differences between the scores were much higher than they should have been.

*When we talked with the other groups after the lesson, we realized that the score ranges are very high (OAF).*

*There were huge differences between the points. I think any criteria was not considered (OAF).*

*We have seen unbalanced score distributions (OAF).*

*When one group gave 30 to a material and another group gave 12 to the same material, it shows the incompatibility clearly (OAF).*

The inconsistencies that emerged during the evaluation process were compared to the questionnaires filled without reading, and it was underlined that the evaluated material was not even considered in the evaluation.

*I think there was no consistency. Scoring was sometimes very irrelevant, like a survey or scales filled out without reading (OAF).*

## **3.2. Interaction**

### **3.2.1. Peer Pressure**

Participants emphasized the pressure in many points while examining the evaluation process. Some participants stated that the environment in which the assessment was made turned into a place where people could not express their opinion freely because of their friends who did not accept criticism. They even shared that commending one another can develop good relationships, while giving negative comments can develop a negative perception toward friends.

*It depends on the person. Of course, I am not afraid to rate the person who can accept criticism, but the person who cannot take it must learn to accept criticism (OAF).*

*While criticizing, I realized that nobody could express their opinions freely and they just made good comments to be good with their friends (OAF).*

---

*Giving points in a classroom setting can strain the student or think negatively towards people who make a negative assessment (OAF).*

### **3.2.2. Instant Interaction**

One of the points considered as indispensable by the participants in the peer assessment was instant interaction. While the instant interaction was emphasized, some of the participants stated that online evaluation provided this better and the others think that face-to-face evaluation provided the instant interaction. For example, one of the participants thought that a more effective evaluation process was carried out because they had the opportunity to instantly convey their criticism in face-to-face evaluation.

*I think face-to-face evaluation was more effective because we were able to criticize each other instantly and it became more effective (OAF).*

Another participant argued that online assessment practices are much more effective as they offer the opportunity to interact instantly, regardless of location.

*When an online assessment application that can be applied simultaneously inside or outside the class is finished, we have the chance to receive reports instantly based on class, student or question (OAF).*

### **3.2.3. Face-to-face Communication**

In addition to demonstrating the advantages of the online evaluation, the participants frequently emphasized the importance of face-to-face interaction during the evaluation process. Underlining the effectiveness of face-to-face interaction, the participants stated that people can express themselves more clearly and comprehensively by face-to-face interaction. Also, it has been added that more spontaneous interaction can be achieved through face-to-face communication.

*The assessment was more effective when it was face to face. Above all, the basic elements of communication are gestures and facial expressions. The realization of interpersonal interaction while evaluating is a necessary skill for us as a teacher candidate. We can measure the reactions of people face to face more easily (OAF).*

*In fact, instead of explaining here, I think it will be in the heat of the moment and more realistic, maybe I could not express myself here as I want (OAF).*

Another reason for the participants to prefer face-to-face evaluation over online evaluation was the direct interaction of the people who evaluated and were evaluated during the face-to-face communication process. It is said that if the assessor is known, assessments can be taken more seriously, and due to the pressure to respond, it may be necessary to think about feedback.

*Face to face evaluation, because we can see who is saying the mistakes, we can decide whether they are realist enough to be considered (OAF).*

*Face to face evaluation. Because the assessed person or the assessors see the answer given, the obligation to give a more logical answer is felt (OAF).*

One of the important points mentioned by the participants is that people evaluated during the face-to-face evaluation both defend themselves and realize their deficiencies with more concrete feedback.

*I think face-to-face was more effective because at that time we had the opportunity to defend the material we prepared and at the same time see our deficiencies (OAF).*

## **3.3. Evaluation Criteria**

Another theme that emerges as a result of the analysis of the data is the evaluation criteria. While the participants stated that they have many different criteria, they emphasized that these



criteria strongly affect the evaluation. When the evaluation criteria are examined, one of the most prominent criteria is related to the extent to which the material meets the targeted gains. Therefore, one of the main criteria determined by the participants was suitability for the gains.

*Which gains the material is made and its suitability to this gain, the usefulness of the material (OAF).*

*The clear criteria helped us to evaluate the materials objectively (Interview- Student 3).*

Another point emphasized by the participants in the evaluation criteria was the usefulness of the developed products. Many of the participants stated that they evaluated the materials by prioritizing their usefulness and functionality.

*I tried to give points by paying attention to all evaluation criteria. Mostly, I paid attention to usefulness (OAF).*

*In particular, I evaluated the materials according to whether they are useful in primary education (Interview-Student 6).*

Another point that stood out in the evaluation criteria of the participants was the presentation of the material. The participants who evaluated the presentations as a kind of marketing method argued that the features the groups highlighted during the presentation were considered more important by the evaluators, and this directly affected the evaluation process.

Besides, some participants stated that instead of focusing on a single aspect of the materials they evaluated, they could approach the evaluation process more fully with a rubric developed in line with the material evaluation criteria. They stated that at the end of the process, the rubric used for evaluation provided a strong argument for the evaluator to justify that s/he evaluated correctly and thus provided a fair evaluation.

Although the groups determined certain criteria and made their evaluations according to the given criteria, some of the participants shared that they made their evaluations based on the evaluations of other groups. Therefore, it is possible to say that there are situations in which participants are affected by others in their evaluations. For example, some of the participants stated that they carried out their evaluations by averaging the scores given by other groups so that they would not affect the overall evaluation positively or negatively.

*We have divided the number of points determined by everyone by 10 and divided it into our group number and said the result. That's exactly how we decided (OAF).*

*We calculated the points given by each group member separately and scored them by taking the average (OAF).*

### **3.4. Attributions of the Online Assessment Platform**

As the participants were not familiar with conducting an evaluation process online, they shared positive and negative thoughts when asked for their opinions about this process.

#### **3.4.1. Positive Attributions**

In this context, the positive feature the participants highlighted is that the evaluation process, which is carried out via mobile phones, can be carried out anytime and anywhere independent by the requirements of the age. The participants who argued that the notifications received through the application are valuable in terms of carrying out the process in a timely and effective manner said that the communication with the one responsible of the course and other students who took the course through this platform made the process more efficient. Thanks to this platform, the participants had constant access to their products and their friends' products.

*I think it has a lot of advantages. Using the application, getting information about the course and seeing the homework of our other friends is an advantage (OAF).*

---

*I could follow the assignments all the time and I had access to the assignments everywhere (OAF).*

The participants who underlined that this evaluation platform is always with them because it is a mobile application stated that it is a great convenience to upload and access this platform via mobile phones instead of carrying and sharing products and materials with them.

*It is more advantageous to submit assignments in the online environment because we have been freed from carrying materials constantly (OAF).*

Another feature highlighted by the participants is that with this evaluation platform, all products and materials prepared by the whole class became a portfolio that is available for future use.

*It is very useful. The assignments that we uploaded there will be useful for us in the future (OAF).*

*In addition, as we uploaded our assignments to the online platform, we were able to access the materials other friends made (Interview- Student 4).*

### **3.4.2. Negative Attributions**

The participants highlighted not only advantages but also disadvantages of the online evaluation platform. Some of the participants stated that they are not yet fully prepared for this technology, as they encountered difficulties in trying it for the first time.

*I do not think we are fully ready for applications made on the internet. We need to improve on this (OAF).*

On the other hand, some of the participants shared that they got used to the application over time and that they did not experience any problems related to use.

*At first, I thought it was a difficult application but with time, I got used to it. Easy and simple application (OAF).*

The technical difficulties, including internet problem, encountered during the use of the evaluation platform were also regarded as disadvantages. Many of the participants stated that they could not use the platform efficiently enough because they did not have enough quality internet access. Also, they shared that the mistakes made while using the phone screen are troublesome.

*Since we made the scoring on the phone, touching accidentally sometimes caused trouble, and I think it was difficult to log in separately for each scoring (OAF).*

*Some of the students had problems connecting to the internet. In addition, students whose phones were old could not use the program as they wanted due to freezing of the screens during the evaluation process (Observations).*

## **4. CONCLUSION**

This study aimed to compare traditional peer evaluation and online peer evaluation in order to investigate which method is more effective in peer assessment strategy. The results of the study showed that students make more objective peer evaluation when online assessment tools are used in comparison to traditional tools. Students' identity was not known during the online peer evaluation process, so students expressed their opinions liberally in online assessment since they did not feel any pressure from their peers. Anonymity, on the other hand, was not possible during the face-to-face assessment. Therefore, online assessment provided more advantages than traditional methods in terms of anonymity of the students and objectivity of the peer evaluations. However, the study also found some disadvantages of assessing their peers in an online format. Students mentioned the difficulty of writing in handheld devices and not having the internet on their devices when using online tools. While it is easy to conclude that teachers

and instructors should implement online peer evaluation methods in order to have an effective feedback mechanism, they also need to keep in mind that students may not be able assess their peers in depth due to challenges of handheld devices and lack of internet on their devices.

The study revealed that students find the online peer assessment environment as a way that enables them to express their opinions objectively. Based on the results of the research, objectivity has emerged as one of the most emphasized points. In a similar context, Kali and Ronen (2005) expressed different arguments. Kali and Ronen (2005) found that there were differences between student and instructor scores; students were not objective in evaluating their peers since they had bias based on personal stand. In addition, according to Herbert (2007), some students did not make an objective evaluation when evaluating their peers. Since this study allowed the participants to compare face-to-face and online evaluation processes, participants found the online evaluation process relatively objective. Therefore, this study reached conclusions different from the literature.

This study demonstrated that the well-organized criteria help pre-service teachers evaluate their peers objectively and systematically. The statements under the evaluation theme provided that the students were more able to approach the evaluation process comprehensively. Also, following the criteria helped them evaluate their peers' materials in a more objective manner. Similar results were found by Chen and Tsai (2009); they mentioned that explained criteria help the instructor to maintain students' attitudes toward the class. It can be concluded that whether it is online peer assessment or traditional face-to-face assessment, pre-defined criteria for evaluation help students make accurate evaluations.

Another highlighted point in this research is the importance of interaction. Students cannot interact with each other in online assessment environments. However, although face-to-face evaluation methods enable interaction among students, pre-service teachers stated that peer pressure may be common in face-to-face evaluation. Hence, the anonymity in the online environment allowed the students to interact with their peers better. Tsai, Lin, and Yuan (2002) identified that students can freely express their thoughts about their peers' work in online peer assessment. McConnell (2002) also finalized similar results and stated that utilizing online peer assessment can furnish students with a mysterious domain to unreservedly communicate their considerations and thoughts regarding others' work.

To summarize, the students were asked to evaluate the materials prepared by their peers as a group. In this context, the opinions of students about online and face-to-face evaluation were taken. Within the scope of this aim, interviews, observation, and online assessment forms data obtained from participants in this study revealed the positive attitudes towards online peer assessment[A1]. The result of the study indicated that students evaluated their peers more objectively online than face to face. Further studies are needed to investigate online peer assessment practices in various aspects. The study recommends that a combination of peer and instructor assessment and even self-assessment can give a better validity of the peer assessment[A3]. Since this study focused on the use of a particular online peer assessment tool, the effectiveness and ease of use affected students' peer assessment experience. The advancement in web 2.0 tools generated various tools for online peer assessment. Therefore, further studies can investigate the effects of various peer assessment tools. The study was conducted with pre-service elementary teachers that took several assessment and evaluation courses; thus, it may be easier to integrate assessment strategies with education faculty students but not with other departments. Lastly, the study found advantages and disadvantages of online peer assessment but with some training for instructors and students. Online peer assessment strategies can be a useful method, especially during Covid-19 pandemic.

## Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

## Authorship Contribution Statement

**Ahmet Oğuz Akçay:** Investigation, research design, literature review, data collection, data analysis, and writing the manuscript. **Ufuk Güven:** Research design, literature review, methodology, and writing the manuscript. **Engin Karahan:** Research design, methodology, data analysis, and writing the manuscript.

## ORCID

Ahmet Oğuz Akçay  <https://orcid.org/0000-0003-2109-976X>

Ufuk Güven  <https://orcid.org/0000-0003-1977-6426>

Engin Karahan  <https://orcid.org/0000-0003-4530-211X>

## 5. REFERENCES

- Ashenafi, M. M. (2019). *Online peer-assessment datasets*. Preprint. <https://arxiv.org/pdf/1912.13050.pdf>
- Bhat, B. A., & Bhat, G. J. (2019). Formative and summative evaluation techniques for improvement of learning process. *European Journal of Business & Social Sciences*, 7(5), 776-785.
- Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education*, 18(5), 529-549.
- Bradley, B. (2020). *Using alternative assessment*. <https://ctl.byu.edu/using-alternative-assessments>
- Brown, G. T. L., & Harris, L. R. (2014). The future of self-assessment in classroom practice: Reframing self-assessment as a core competency. *Frontline Learning Research*, 3, 22-30. <https://doi.org/10.14786/flr.v2i1.24>
- Chen, Y.-C., & Tsai, C. (2009). An educational research course facilitated by online peer assessment. *Innovations in Education and Teaching International*, 46, 105-117. <https://doi.org/10.1080/14703290802646297>
- Chin, P. (2007). Peer assessment. *New Directions in the Teaching of Physical Sciences*, 3, 13-18. <https://doi.org/10.29311/ndtps.v0i3.410>
- Davies, P. (2002). Using students reflective self-assessment for awarding degree classifications. *Innovations in Education and Teaching International*, 39(4), 307-319. <https://doi.org/10.1080/13558000210161034>
- Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer, and co-assessment in higher education: A review. *Studies in Higher Education*, 24(3), 331-350. <https://doi.org/10.1080/03075079912331379935>
- Duran, M., Mihladiz, G., & Balliel, B. (2013). İlköğretim öğretmenlerinin alternatif değerlendirme yöntemlerine yönelik yeterlilik düzeyleri [Adequacy levels of primary school teachers towards alternative assessment methods]. *Mehmet Akif Ersoy Üniversitesi Eğitim Bilimleri Enstitüsü Dergisi*, 2(2), 26-37.
- Falchikov, N. (1995). Peer feedback marking developing peer assessment. *Innovations in Education and Training International*, 32, 175-187. <https://doi.org/10.1080/1355800950320212>
- Falchikov, N. (2001). *Learning together: Peer tutoring in higher education*. Routledge Falmer.

- Garrison, C., & Ehringhaus, M. (2007). *Formative and summative assessments in the classroom*. [http://ccti.colfinder.org/sites/default/files/formative\\_and\\_summative\\_assessment\\_in\\_the\\_classroom.pdf](http://ccti.colfinder.org/sites/default/files/formative_and_summative_assessment_in_the_classroom.pdf)
- Hatch, A. J. (2002). *Doing qualitative research in education settings*. State University of New York Press.
- Herbert, N. (2007, January). Quantitative peer assessment: Can students be objective?. In *Proceedings of the Ninth Australasian Conference on Computing Education-Volume 66* (pp. 63-71). Australian Computer Society, Inc.
- Kali, Y., & Ronen, M. (2005, May). Design principles for online peer-evaluation: Fostering objectivity. In *Proceedings of the 2005 Conference on Computer Support for Collaborative Learning: Learning 2005: The next 10 years!* (pp. 247-251). International Society of the Learning Sciences.
- Kampen, M. (2020). *The 6 type of assessments*. <https://www.prodigygame.com/main-en/blog/types-of-assessment/>
- Langan, M.A., & Wheeler, C.P. (2003). Can students assess students effectively?. Some insights into peer assessment. *Learning and Teaching in Action*, 2(1). <https://www.celt.mmu.ac.uk/ltia/issue4/langanwheater.pdf>
- Linn, R. L., & Gronlund, N. E. (1995). *Measurement and assessment in teaching* (7th ed.). Pentice-Hall Inc.
- McConnell, D. (2002). The experience of collaborative assessment in E-learning. *Studies in Continuing Education*, 24, 73-102. <https://doi.org/10.1080/01580370220130459>
- Merriam, S. B. (1998). *Qualitative research and case study application in education*. Jossey-Bass.
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). Sage.
- Miller, L., & Ng, R. (1996). *Autonomy in the classroom: peer assessment*. In R. Pemberton, eds. pp. 133-146.
- Ndoye, A. (2017). Peer/Self assessment and student learning. *International Journal of Teaching and Learning in Higher Education*, 29 (2), 255-269.
- Özenç, M., & Çakır, M. (2015). Sınıf öğretmenlerinin alternatif ölçme ve değerlendirme yeterliklerinin belirlenmesi [Exploring primary school teachers' competencies of alternative assessment and evaluation]. *Elementary Education Online*, 14(3), 914-933. <http://dx.doi.org/10.17051/ieo.2015.22900>
- Patton, M. Q. (2002). *Qualitative evaluation and research methods* (3rd ed.). Sage Publications, Inc.
- Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test enhanced learning in the classroom: long-term improvements from quizzing. *Journal of Experimental Psychology. Applied*, 17(4), 382-95. <https://doi.org/10.1037/a0026252>
- Şahin, M. G., & Kalyon, D. Ş. (2018). Öğretmen adaylarının öz-akran-öğretmen değerlendirmesine ilişkin görüşlerinin incelenmesi [Investigation of preservice teachers' opinions about self-, peer- and teacher assessment]. *Kastamonu Eğitim Dergisi*, 26(4), 1055-1068. <https://doi.org/10.24106/kefdergi.393278>
- Taras, M. (2005). Assessment—summative and formative—some theoretical reflections. *British Journal of Educational Studies*, 53(4), 466-478. <https://doi.org/10.1111/j.1467-8527.2005.00307.x>
- The National Council for Teacher Education [NCTE]. (2013). *Formative assessment that truly informs instruction*. [http://www.ncte.org/positions/statements/formative-assessment/formative-assessment\\_full](http://www.ncte.org/positions/statements/formative-assessment/formative-assessment_full)
- Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, 68(3), 249-276. <https://doi.org/10.2307/1170598>



- 
- Topping, K. J. (2009). Peer Assessment. *Theory into Practice*, 48(1), 20-27. <https://doi.org/10.1080/00405840802577569>
- Tsai, C.-C., Lin, S.S.J., and Yuan, S.-M. (2002). Developing science activities through a networked peer assessment system. *Computers & Education* 38, 241–252. [https://doi.org/10.1016/S0360-1315\(01\)00069-0](https://doi.org/10.1016/S0360-1315(01)00069-0)
- Wen, M. L., & Tsai, C. (2008) Online peer assessment in an inservice science and mathematics teacher education course. *Teaching in Higher Education*, 13, 55-67. <http://doi.org/10.1080/13562510701794050>
- Yildirim, A., & Simsek, H. (2011). *Sosyal bilimlerde nitel araştırma yöntemleri* (8th ed.). Seckin Yayınevi.

## A Guide for More Accurate and Precise Estimations in Simulative Unidimensional IRT Models

Fulya Baris Pekmezci <sup>1,\*</sup>, Asiye Sengul Avsar <sup>2</sup>

<sup>1</sup>Bozok University, Faculty of Education, Department of Educational Sciences, Measurement and Evaluation in Education, Yozgat, Turkey

<sup>2</sup>Recep Tayyip Erdoğan University, Faculty of Education, Department of Educational Sciences, Measurement and Evaluation in Education, Rize, Turkey

### ARTICLE HISTORY

Received: Sep. 04, 2020

Revised: Mar. 30, 2020

Accepted: Apr. 11, 2021

### Keywords:

Monte carlo simulation study,  
Replication,  
Unidimensional item response theory models,  
Bias estimation,  
Type I error inflation.

**Abstract:** There is a great deal of research about item response theory (IRT) conducted by simulations. Item and ability parameters are estimated with varying numbers of replications under different test conditions. However, it is not clear what the appropriate number of replications should be. The aim of the current study is to develop guidelines for the adequate number of replications in conducting Monte Carlo simulation studies involving unidimensional IRT models. For this aim, 192 simulation conditions which included four sample sizes, two test lengths, eight replication numbers, and unidimensional IRT models were generated. Accuracy and precision of item and ability parameter estimations and model fit values were evaluated by considering the number of replications. In this context, for the item and ability parameters; mean error, root mean square error, standard error of estimates, and for model fit;  $M_2$ ,  $RMSEA_2$ , and Type I error rates were considered. The number of replications did not seem to influence the model fit, it was decisive in Type I error inflation and error prediction accuracy for all IRT models. It was concluded that to get more accurate results, the number of replications should be at least 625 in terms of accuracy of the Type I error rate estimation for all IRT models. Also, 156 replications and above can be recommended. Item parameter biases were examined, and the largest bias values were obtained from the 3PL model. It can be concluded that the increase in the number of parameters estimated by the model resulted in more biased estimates.

## 1. INTRODUCTION

To make sense of human behavior, individuals need to be observed and evaluated accurately. According to these evaluations, it is important to make decisions about individuals or to direct them towards their needs in a true way. Therefore, the psychometric properties of the measurement tools used for evaluations must be at satisfactory levels.

Test theories are used to assess the psychometric properties of measurement tools. Test theories can be considered as a study area where research is conducted to investigate problems affecting

---

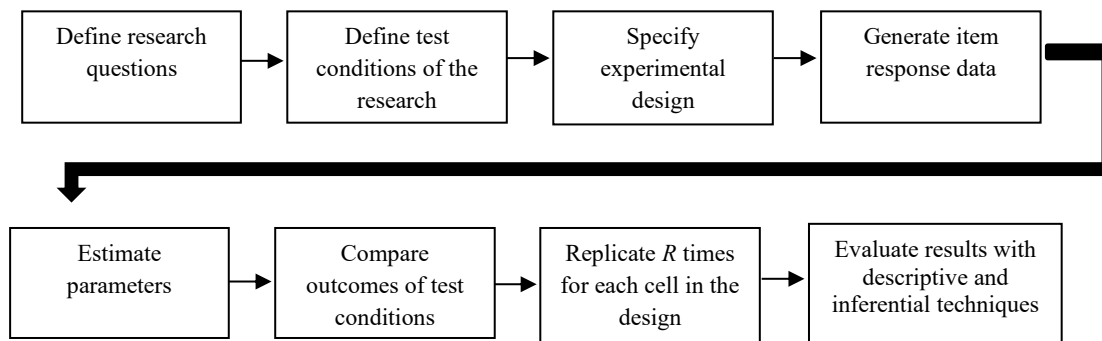
\*CONTACT: Fulya BARIŞ PEKMEZCİ ✉ [fulyabaris@gmail.com](mailto:fulyabaris@gmail.com) 📍 Bozok University, Department of Educational Sciences, Measurement and Evaluation in Education, Yozgat, Turkey

psychological measurements and to achieve valid and reliable measurement results by trying to reduce these problems as much as possible (Crocker & Algina, 1986). In the literature, the classical test theory and item response theory (IRT) are the most studied theories in the psychometric area.

Human behavior is the main subject of social sciences. It is very important to measure human characteristics, which are very variable, validly, and reliably. The measurement of human behavior is different from the measurements made in natural science. Ideal laboratory conditions are created to achieve the most accurate results in natural science, but it is very difficult to apply these in social sciences. One of the best ways to achieve accuracy in social sciences is through simulation studies. Simulation studies have been used since 1900 as a solution to statistical problems (Harwell et al., 1996).

IRT has strong assumptions that differ according to dimensionality, linearity, or scoring type (McDonald, 1982). In cases where the IRT assumptions are not met, the results of the analysis and estimates will be inaccurate. Monte Carlo (MC) simulation studies provide solutions to the problems that can be encountered by creating ideal data sets that meet the assumptions required for IRT (Han, 2007). MC simulation studies are used for many purposes such as the evaluation of new parameter estimation procedures, comparison of different item analysis programs, and parameter estimation in multidimensional data sets (Harwell, 1997). MC studies perform statistical sampling experiments via computers for solutions to statistical problems (Mundform et al., 2011). How MC studies are structured in IRT (Harwell et al., 1996) is shown in Figure 1.

**Figure 1.** Steps of a MC Simulation Study in IRT.



The MC process starts with defining the research question, as seen in Figure 1. When the research questions defined in psychometry are related to theoretical studies, especially include comparing different conditions at the same time, a simulation study is inevitable in obtaining the appropriate data sets. Then, it is important to define the test conditions of the research. These conditions consist of dependent variables such as item and ability parameters, and independent variables like test lengths, sample size, or distribution of the sample. After specifying the experimental design, the item response data are generated by the IRT model which is chosen by the researchers. Item and ability parameters are estimated from the generated data sets. Results obtained from different test conditions are compared. This process is replicated  $R$  times and all outcomes are evaluated for answering the research questions using descriptive and inferential techniques.

One of the important issues to be considered in MC simulation studies is the number of replications. With insufficient replications, estimations can be inaccurate (Mundform et al., 2011). Besides, replication is often confused with iteration in the literature. Hence, it is important to clarify the difference between replication and iteration in simulation studies. Iteration is defined as a statistical routine. This routine starts with the first estimate and

continues until a satisfactory estimate, which means the convergence criterion is met, is obtained by working with some statistical rules on this first estimation (Fu, 2019; Thompson, 2004; 2006).

Iterations are needed for convergence of statistical algorithms (Hair et al., 2019). Some iteration algorithms which are used for parameter estimation in IRT are: the Broyden-Fletcher-Goldfarb-Shanno Algorithm, the Bisection Method, the Expectation-Maximization Algorithm, Fisher Scoring, the Gibbs Sampling Algorithm, the Markov Chain Monte Carlo Algorithm, the Newton-Gauss Algorithm, and the Newton-Raphson Algorithm (Cai & Thissen, 2014; Chalmers, 2012; Hanson, 1998; Patsias et al., 2009; Tavares et al., 2004; Thompson, 2009; van der Linden, 2018; Weismann, 2013).

As for replication, this is defined as the repeated administration of an experiment with selected changes in parameters or test conditions being made by the researcher (Hair et al., 2019; Rubinstein, 1981). Replications give an estimate of the stability of the predictions made in simulation studies (Feinberg & Rubright, 2016). Because the number of replications affects the accuracy and reliability of parameter estimates (Feinberg & Rubright, 2016), it is stated that the number of replications is an important factor for statistical results (Kleijnen, 1987; Rubinstein, 1981). These estimations are directly related to the implications to be reached in simulation studies. When conducting a MC simulation study, it is important to answer the question of how many replications are needed for accurate estimations. So, the number of replications should be determined carefully by the researchers. Within the context of unidimensional IRT models, various studies that are conducted on the MC method with a different number of replications are given in [Table 1](#).

**Table 1.** Literature review about the number of replications for unidimensional IRT models.

Studies	Number of Replication
Sheng & Wikle, 2007	10
Roberts et al., 2002	30
Sen et al., 2016	50
Crışan et al., 2017; Lee et al., 2017; Park et al., 2016; Yang, 2007; Zhang, 2008	100
Matlock Cole & Paek, 2017	200
Feinberg & Rubright, 2016	250
Matlock & Turner, 2016	500
Ames et al., 2020; Reise et al., 2011	1000
Baldwin, 2011; Mundform et al., 2011	5000
Babcock, 2011	10000

As is seen from [Table 1](#), the different number of replications ranges between 10 and 10000. It is usual for a different number of replications to be made in varying test conditions for accurate parameter estimations by different IRT models. However, it is not clear what the appropriate number of replications should be under varying test conditions for unidimensional IRT models. In addition, it is important to determine a sufficient number of replications according to test conditions that are specified by the researchers. Although simulative studies provide convenience to theoretical studies, they are time-consuming processes.

To establish a rule for what ideal replication number should be, Feinberg and Rubright (2016) had provided a formula about replication number, which is given in Equation 1:

$$\sigma_M = \frac{\hat{\sigma}}{\sqrt{R-1}} \tag{Equation 1}$$

where  $\hat{\sigma}$  is the standard deviation of the estimated parameter across replications and  $R$  is the number of replications and  $\sigma_M$  is the SE of the mean.

According to their formula, they suggested calculating the ideal number of replications by using the standard deviation of the estimated parameters across replications. To determine the ideal replication number, firstly, the researchers must replicate data, and secondly, the ideal replication number must be calculated according to replicated samples' standard deviation. Starting replication number will be the determiner of the ideal replication number. This seems a time-consuming process. Because, firstly, data need to be replicated, and then the ideal replication number must be calculated. Doing more replication will result in a smaller standard deviation of replicated samples or vice versa. Hence, the calculation of ideal replication number according to Feinberg and Rubright (2016) will tend to be smaller due to using that standard deviation. Large standard deviations will recommend more replications. Lastly, there is no exact rule about what the ideal standard deviation of replicated samples should be (see for details Feinberg & Rubright, 2016). Therefore, using Equation 1 does not seem very practical.

In this study, the number of replications required for the most accurate parameter estimations in various sample sizes and test lengths according to unidimensional IRT models (1PL model, 2PL model, and 3PL model) was determined.

The purpose of the current study is to develop guidelines for the adequate number of replications in conducting MC simulation studies involving unidimensional IRT models with different test conditions. Based on this purpose, answers to the following research questions were sought:

1. How are the estimations of item parameters obtained from varying sample sizes and test lengths affected by varying numbers of replications?
2. How are the estimations of ability parameters obtained from varying sample sizes and test lengths affected by varying numbers of replications?
3. How are the estimations of model fit obtained from varying sample sizes and test lengths affected by varying numbers of replications?

## 2. METHOD

### 2.1. Study Design Factors

The purpose of this study is to develop guidelines for the adequate number of replications in conducting MC simulation studies involving unidimensional IRT models with different test conditions. According to this aim, different sample sizes and test lengths were studied to determine the adequate number of replications to obtain more accurate and precise estimations.

In line with this purpose, firstly, studies which implemented unidimensional IRT models and MC simulation studies were reviewed. According to the literature review (Baldwin, 2011; Mundform et al., 2011), 5000 was selected as a starting replication number for this study. In determination of other numbers of replications, the method which Preecha (2004) implemented in his study was used. Considering this method, if the bias difference between two consecutive replication numbers is large, this interval should be halved, and the analysis should be repeated. If not, then the last replication number should be halved, the analysis should be repeated, and the bias statistics should be calculated.

After determining the maximum replication number as 5000, bias analyses were performed. Half of the 5000 replications were taken, and the analyses were re-run for 2500 replications. This process was performed until the number of replications was 78. Additionally, the minimum number of replications was determined as 20. In some nonparametric IRT studies, 20 is used as the minimum number of replications (Şengül Avşar & Tavşancıl, 2017; van Onna, 2004). Therefore, in this study, the adequacy of 20 replications was also tested.



Within the scope of this study, a literature review was also done for the test lengths and sample sizes which are given in Table 2. In IRT studies, there are no exact rules for adequate sample sizes for accurate and precise estimation (De Ayala, 2009; Kirisci et al., 2001; Reise & Yu, 1990). At this point, it is important to explain what accuracy and precision are.

Accuracy indicates how close the measured values are to known values. For example, if in the laboratory one measures a given object as 132.2 cm, but the known height is 150 cm, then the measurement of the given object is inaccurate. In this case, the measurement is not close to the known value. Precision indicates how two or more measurements are close to each other. Using the aforementioned example, if one measures a given object ten times, and obtains 132.2 cm each time, then the measurement of that object is very precise. Any measurement can be very precise but inaccurate, as described above, while it can also be accurate but imprecise (Barış Pekmezci & Gülleroğlu, 2019).

Sample sizes and test lengths are the other independent variables of this research besides number of replications and IRT models. In order to determine which sample sizes and test lengths were commonly used in unidimensional IRT studies, literature was reviewed. The literature review results are given in Table 2.

**Table 2.** Literature review about sample sizes and test lengths for unidimensional IRT models.

Studies	Sample Size			Test Lengths
	1PL model	2PL model	3PL model	
Lord, 1968	1000	-	-	50
Hulin et al., 1982	-	-	500/1000	30/ 60
Thissen & Wainer, 1982	1000	2500	-	
Goldman & Raju, 1986	250	1000	-	
Yen, 1987	-	-	1000	10/ 20/40
Patsula & Gessaroli, 1995	-	-	1000	20/40
Baker, 1998	-	500	-	50
De La Torre & Patz, 2005	-	-	1000	10/30/ 50
Gao & Chen, 2005	-	-	500/ 2000	10/ 30/ 60
Yang, 2007	100/500/1000	-	-	15/ 30/ 45
Babcock, 2011	-	1000/2500/4000	-	54/62/70
Chuah et al., 2006	-	-	500/1000	20
Sahin & Anil, 2017	150/ 250/ 350/ 500/ 750/ 1000/2000/ 3000/ 5000	150/ 250/ 350/ 500/ 750/ 1000/2000/ 3000/ 5000	150/ 250/ 350/ 500/ 750/ 1000/2000/ 3000/ 5000	10/20/30
Matlock Cole & Paek, 2017	-	1500	3000	20/40
Ames et al., 2020	-	250/500/1000	250/500/1000	10/40

Sample sizes and test lengths differ as can be seen from Table 2. Accordingly, sample sizes are varied between 150 and 5000, while test lengths are varied between 10 and 70. Minimum sample size was determined as 500, medium sample sizes were determined as 1000 and 2000, and maximum sample size was determined as 3000 for this research. Test lengths were selected as 25 items for short tests and 50 items for long tests for this research.

To begin the simulation, the item difficulty parameters ( $b$ ), the item discrimination parameters ( $a$ ), the item lower asymptote parameters or guess parameters ( $c$ ), and the ability parameters ( $\theta$ ) were chosen according to the literature review. In this study, the  $b$  parameters are normally distributed [ $b \sim N(0.50, 1.50)$ ]; the  $a$  parameters are uniformly distributed [ $a \sim U(1.5, 2.0)$ ], the  $c$  parameters are beta distributed [ $c \sim Beta(20, 90)$ ], and the ability parameters ( $\theta$ ) are normally

distributed [ $\theta \sim N(0, 1)$ ] (Bahry, 2012; Bulut & Sünbül, 2017; Cohen et al., 1993; DeMars, 2002; Feinberg & Rubright, 2016; Jiang et al., 2016; Harwell & Baker, 1991; Mislevy & Stocking, 1989; Mooney, 1997). According to these parameters, dichotomous response patterns were generated for selected conditions (3x2x4x8), which are shown in Table 3. The generation of the data sets in the test conditions, determined in the research, by two computers with 2.7 GHz Intel Core i5 8 GB RAM and 1.8 GHz Turbo Intel Core i7 16 GB RAM took approximately a month.

**Table 3.** Simulation conditions.

IRT Models	Test lengths	Sample Size	Number of Replications								
			20	78	156	312	625	1250	2500	3000	
1PL model	25 items	500	✓	✓	✓	✓	✓	✓	✓	✓	✓
		1000	✓	✓	✓	✓	✓	✓	✓	✓	✓
		2000	✓	✓	✓	✓	✓	✓	✓	✓	✓
		3000	✓	✓	✓	✓	✓	✓	✓	✓	✓
	50 items	500	✓	✓	✓	✓	✓	✓	✓	✓	✓
		1000	✓	✓	✓	✓	✓	✓	✓	✓	✓
		2000	✓	✓	✓	✓	✓	✓	✓	✓	✓
		3000	✓	✓	✓	✓	✓	✓	✓	✓	✓
2PL model	25 items	500	✓	✓	✓	✓	✓	✓	✓	✓	✓
		1000	✓	✓	✓	✓	✓	✓	✓	✓	✓
		2000	✓	✓	✓	✓	✓	✓	✓	✓	✓
		3000	✓	✓	✓	✓	✓	✓	✓	✓	✓
	50 items	500	✓	✓	✓	✓	✓	✓	✓	✓	✓
		1000	✓	✓	✓	✓	✓	✓	✓	✓	✓
		2000	✓	✓	✓	✓	✓	✓	✓	✓	✓
		3000	✓	✓	✓	✓	✓	✓	✓	✓	✓
3PL model	25 items	500	✓	✓	✓	✓	✓	✓	✓	✓	✓
		1000	✓	✓	✓	✓	✓	✓	✓	✓	✓
		2000	✓	✓	✓	✓	✓	✓	✓	✓	✓
		3000	✓	✓	✓	✓	✓	✓	✓	✓	✓
	50 items	500	✓	✓	✓	✓	✓	✓	✓	✓	✓
		1000	✓	✓	✓	✓	✓	✓	✓	✓	✓
		2000	✓	✓	✓	✓	✓	✓	✓	✓	✓
		3000	✓	✓	✓	✓	✓	✓	✓	✓	✓

## 2.2. Simulating Model Parameters and Item Responses

All parameters were simulated based on the null (ideal) model. Any departure from the null model can cause misfit or non-fit of the data, therefore; misspecified models are not in the scope of this research. To simulate dichotomous item responses and estimate the item parameters based on the unidimensional IRT models, the “itemrecovery” function, which is composite of R functions and defined by Bulut and Sünbül (2017), was revised for this study and used. This function, which generates item parameters, simulates item responses concerning parameters, estimates the item parameters of related IRT models, and computes bias statistics, was adapted to the current study. IRT model parameters and model fit values were estimated using the mirt package (Chalmers, 2012) in R. After all bias statistics had been calculated, the relevant graphics were drawn by using the lattice package (Sarkar, 2008) in R.

### 2.3. Estimation of Model Parameters and Type I Error Rates

The evaluation of the accuracy and precision of item and ability parameter estimations throughout the replications was carried out via mean error (ME), root mean square error (RMSE) and standard error of estimates (SE). Mean Error (ME) measures the average magnitude of the errors. ME is the average of the differences between the model’s predicted and actual values, where all individual differences have equal weight. ME is given in Equation 2:

$$ME = \frac{1}{N} \sum_{i=1}^N y_i - \hat{y}_i \tag{Equation 2}$$

where  $N$  is the total test length,  $\hat{y}_j$  is the estimated item parameter for item  $i$  ( $i = 1, 2, \dots, N$ ), and  $y_j$  is the true item parameter for item  $i$ .

RMSE is the square root of the variance of the residuals. It indicates the fit of the model, which is the closeness of the observed data points to the model’s predicted values. RMSE can range from 0 to  $\infty$  and lower values mean better fit. The errors are squared before they are averaged. RMSE should be used when undesirable large errors exist because, in the calculation, RMSE gives a relatively high weight to large errors.

RMSE is in the same unit as the response variable and can be interpreted as the variation of unexplained variance. RMSE is an important criterion of estimation accuracy, and it is important when the interest is in the model prediction. There is no one best model fit measure; researchers should choose depending on their objectives, and more than one is often useful. RMSE is given in Equation 3:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \tag{Equation 3}$$

Standard Error (SE), like standard deviation, is a measure of dispersion. However, while the standard deviation is a measure of dispersion from sample values, the standard error is a measure of dispersion from the sampling distribution, which belongs to the population of interest. SE is the measure of how accurate and precise the sample is. SE is not only a measure of dispersion and accuracy of the sample statistic but also an important indicator of reliability of estimation of the population parameter. SE is given in Equation 4:

$$SE = \frac{1}{N} \sum_{i=1}^N (\hat{y} - \frac{\sum_{i=1}^N \hat{y}}{N})^2 \tag{Equation 4}$$

In addition to bias estimation of model parameters, Type I error rates for model fit were calculated in this study. Glass et al. (1972) emphasize that sampling error contaminates empirical Type I error and statistical power. Therefore, in comparing Type I error, they highly recommended taking this sampling error into account. Glass et al. (1972) suggested Equation 5 (Type I error rates) about standard error of a sampling proportion by using the number of replications as a sample size:

$$\hat{\sigma}_p = \sqrt{\frac{(1-P)P}{R}} \tag{Equation 5}$$

where  $R$  denotes the number of replications,  $P$  is the nominal or theoretical Type I error (.05 for this study), and  $p$  is the empirical or the observed Type I error. Glass et al. (1972) advise against considering the difference between a particular observed  $p$  value and the theoretical  $P$  value significant, if departure is less than two standard errors of that  $p$ .

To estimate accuracy of error rate, the MC variance of an estimate of Type I error rate ( $\frac{\hat{\sigma}_p}{\sqrt{R}}$ ) was used, where  $\hat{\sigma}_p$  is the simulated standard deviation of the  $p$  values, and  $R$  is the number of replications.

### 3. RESULT / FINDINGS

Findings are given in the order of the research questions. Most parts of the analysis outputs are given in the [Supplementary](#) file due to the excessive number of simulation conditions (in total 192 conditions from [Table 3](#)). Only the most remarkable findings are given via figures and tables in the findings section. For detailed information, the [Supplementary](#) file can be reviewed.

#### 3.1. The Effect of the Number of Replications on Estimation of Item Parameters with Varying Sample Sizes and Test Lengths

Bias estimations of item parameters obtained by examining simulation conditions are given in this section. [Figure 2](#), [Figure 3](#), and [Figure 4](#) summarize for RMSE values according to IRT models. Besides, ME, RMSE, and SE values are given in the [Supplementary](#) file.

When ME, RMSE, and SE values according to item parameters were examined, the same pattern was seen for all IRT models. Therefore, findings were interpreted in a way that concerns all IRT models. Increasing the sample size resulted in decreasing RMSE values for  $b$  parameters in all simulation conditions. When the RMSE values were examined in terms of sample sizes in detail, for all replication numbers, it was seen that the bias differences between samples were quite large. Contrary to this, when each sample was analyzed within itself, a slight difference was found in regard to replication number. For example, for the simulation condition with the 1PL model with a test length of 25 items and sample size of 1000, RMSE values obtained from 5000 replications and 78 replications were compared, the difference between them was found to be 0.001. This indicates that parameter estimation accuracy was mostly affected by sample size rather than by the number of replications. Results of ME, RMSE, and SE can be seen in [Table 4](#).

**Table 4.** Accuracy and precision of  $b$  parameters.

IRT Models	Test lengths	Bias statistics	Number of replications			
			20		5000	
Sample size			500	3000	500	3000
1PL model	25 items	ME	0.024	0.007	0.002	0.000
		RMSE	0.024	0.055	0.134	0.054
		SE	1.496	1.453	1.465	1.457
	50 items	ME	-0.002	-0.003	0.002	0.000
		RMSE	0.139	0.053	0.135	0.055
		SE	1.510	1.484	1.489	1.483
2PL model	25 items	ME	-0.01	-0.009	0.008	0.003
		RMSE	0.196	0.073	0.191	0.075
		SE	1.517	1.479	1.481	1.458
	50 items	ME	0.037	-0.015	0.009	0.003
		RMSE	0.199	0.085	0.190	0.075
		SE	1.529	1.479	1.509	1.482
3PL model	25 items	ME	-0.030	-0.002	-0.050	-0.004
		RMSE	0.628	0.228	0.588	0.203
		SE	1.731	1.470	1.621	1.463
	50 items	ME	-0.046	-0.005	-0.047	-0.003
		RMSE	0.553	0.172	0.519	0.185
		SE	1.668	1.450	1.602	1.489

When ME and SE statistics were examined, although the average ME and SE did not change as much as RMSE values according to the sample size, the highest bias values were observed in the smallest sample size for both test lengths. Additionally, except for the 3PL model in terms

of RMSE values,  $b$  parameter estimation biases were found to be quite similar for both test lengths. According to SE values, it can be said that the precision of  $b$  parameter estimates were not much affected by the number of replications. Accuracy and precision of  $b$  parameters, which was obtained with the minimum replication number (20) and the largest sample size (3000), could not be obtained with the maximum replication number (5000) and the minimum sample size (500).

For  $a$  parameters, bias statistics were examined and interpreted in detail according to both test lengths. In regard to  $a$  parameters, increasing the sample size resulted in decreased bias statistics (ME, RMSE, SE) for both test lengths except one condition. Estimation of  $a$  parameter accuracy and precision is directly related with sample size. Accuracy and precision of  $a$  parameters, which was obtained with the minimum replication number (20) and the largest sample size (3000), could not be obtained with the maximum replication number (5000) and the minimum sample size (500). Related findings can be seen in Table 5.

**Table 5.** Accuracy and precision of  $a$  parameters.

IRT Models	Test lengths	Bias statistics	Number of replications				
			20		5000		
			500	3000	500	3000	
2PL model	25 items	ME	0.018	0.002	0.022	0.004	
		RMSE	0.214	0.086	0.215	0.085	
		SE	0.256	0.167	0.249	0.163	
	50 items	ME	0.003	0.006	0.020	0.003	
		RMSE	0.215	0.082	0.207	0.082	
		SE	0.248	0.163	0.244	0.163	
	3PL model	25 items	ME	0.201	0.008	0.179	0.008
			RMSE	0.735	0.194	0.649	0.194
			SE	0.703	0.229	0.631	0.229
50 items		ME	0.168	0.019	0.146	0.016	
		RMSE	0.592	0.163	0.544	0.168	
		SE	0.583	0.211	0.538	0.217	

Regardless of the sample size, bias statistics (ME, RMSE, SE) were not substantially affected by the number of replications. For example, for the 2PL model with a test length of 50 items and sample size of 500, the SE of the  $a$  parameters obtained from 5000 replications and 20 replications were compared, and the difference between them was found to be 0.004. Regardless of the sample size, parameter estimation bias (ME, RMSE, SE) of  $a$  parameters were not affected by the number of replications, as in  $b$  parameters. In summary, it was seen that the sample size had the largest effect rather than the number of replications in the estimation of both  $a$  and  $b$  parameters in both test lengths.

For  $c$  parameters, bias statistics were examined, and it was seen that as the sample size increased, SE and RMSE decreased. When the sample size was the largest (3000), the estimation accuracy and precision obtained with the minimum replication number (20) could not be obtained with the smallest sample size (500) and maximum replication number (5000). Related findings can be seen in Table 6. In summary, like the other item parameters ( $a$  and  $b$ ), sample size had a greater effect on  $c$  parameter estimation bias than replication number.

When the effect of test lengths on parameter estimation bias was examined, it was seen that, for  $a$  parameters, increasing the length of the test provided more accurate and precise parameter estimation in all sample sizes and replication numbers. Increasing the length of the test



decreased the estimation bias of  $a$  parameters. For  $b$  parameters, increasing the test lengths resulted in increased SE. In terms of RMSE values, there was no remarkable change in the accuracy of  $b$  parameter estimations. In general, increasing the test lengths resulted in increased accuracy and precision of  $c$  parameters.

**Table 6.** Accuracy and precision of  $c$  parameters.

IRT Model	Test lengths	Bias statistics	Number of replications			
			20		5000	
Sample size			500	3000	500	3000
3PL model	25 items	ME	-0.001	-0.007	0.002	-0.002
		RMSE	0.134	0.076	0.141	0.077
		SE	0.144	0.082	0.141	0.083
	50 items	ME	0.001	-0.004	0.004	-0.002
		RMSE	0.134	0.072	0.131	0.071
		SE	0.134	0.078	0.133	0.078

**Figure 2.** RMSE values for IPL model.

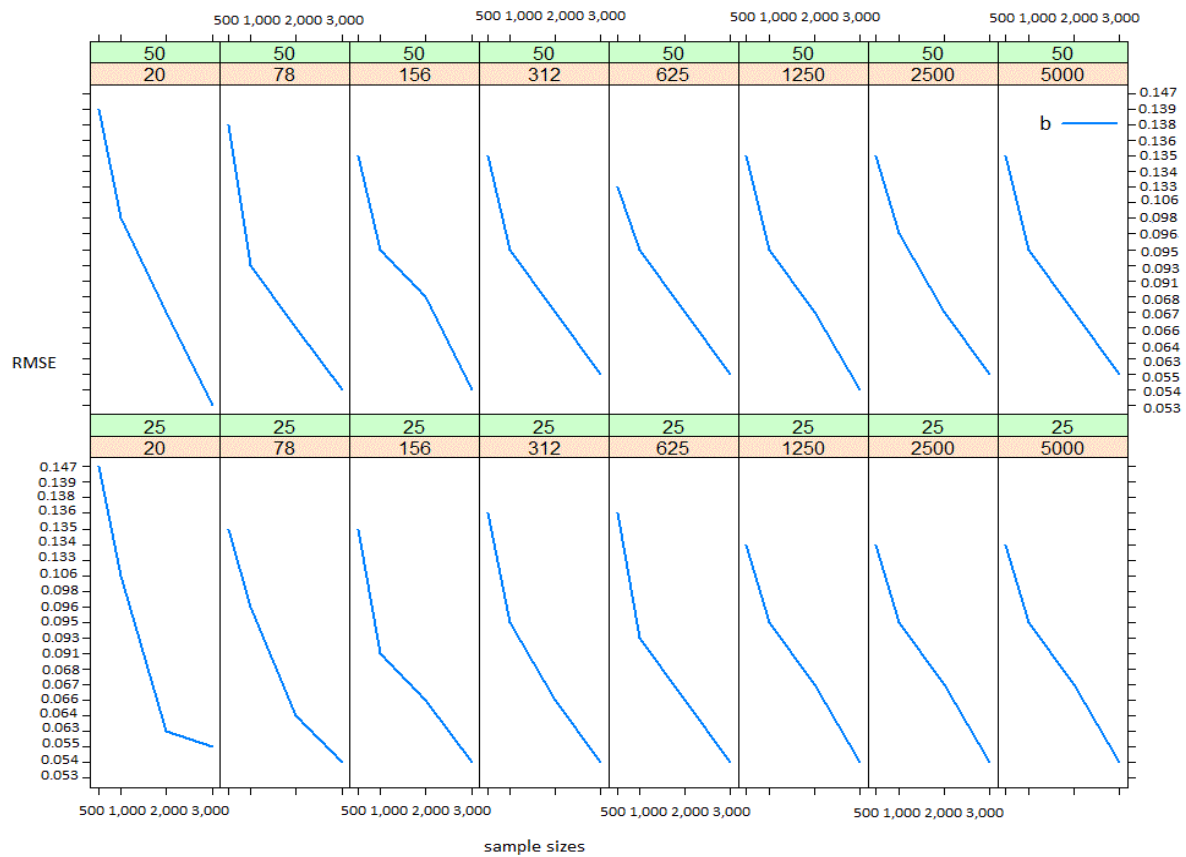


Figure 3. RMSE values for 2PL model.

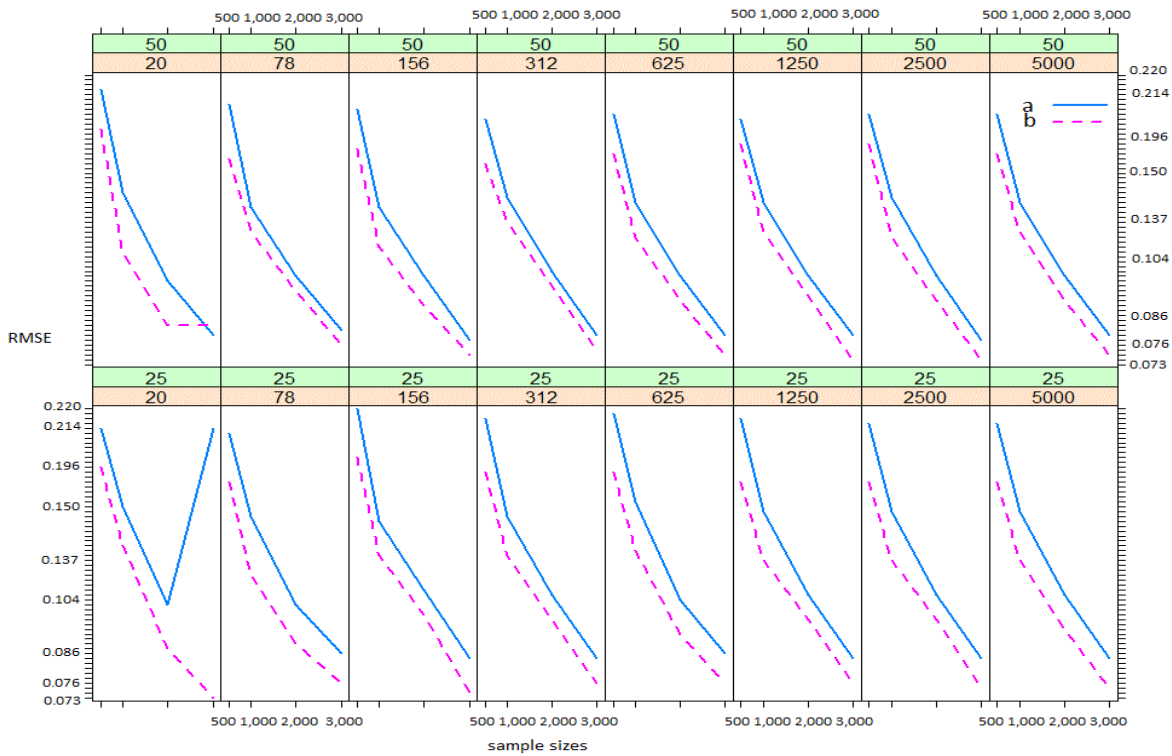
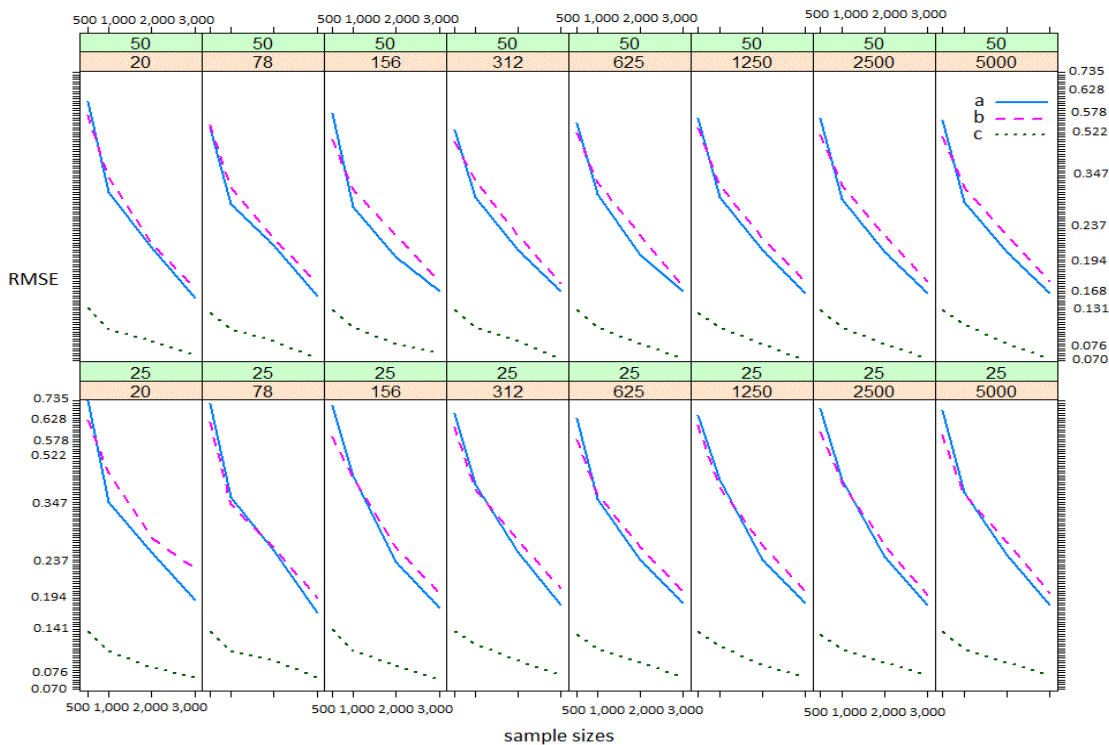


Figure 4. RMSE values for 3PL model.



As a result, it was seen that the sample size had a greater effect rather than the number of replications in the estimation of item parameters ( $a$ ,  $b$ , and  $c$ ). The parameter accuracy and precision obtained with the minimum replication number when the sample size was the largest could not be obtained with the maximum replication number when the sample size was the

smallest. When item parameter biases were examined among IRT models, the largest bias values were obtained from the 3PL model. It can be concluded that the increase in the number of parameters estimated by the model resulted in more biased estimates.

### 3.2. The Effect of the Number of Replications on Estimation of Ability Parameters with Varying Sample Sizes and Test Lengths

Bias estimations of ability parameters ( $\theta$ ) obtained by examining simulation conditions are given in this section. Besides, all bias statistics are given in the [Supplementary](#) file. The ability parameter ( $\theta$ ) estimation accuracy and precision did not change much according to test lengths within all IRT models. Apart from this finding, between all IRT models, some minor differences occurred in terms of bias statistics.

For the 1PL model, when the bias statistics were inspected in detail, it was seen that in general, estimation accuracy for  $\theta$  parameters increased if sample size was increased. When SE values were investigated in terms of estimation precision, the largest sample size (3000) and minimum replication number (20) conditions (0.071 and 0.044, respectively for test lengths 25 and 50 items) were superior to the smallest sample size (500) and maximum replication number (5000) conditions (0.176 and 0.087, respectively for test lengths 25 and 50 items). In other words,  $\theta$  parameters with the minimum sample size and maximum replication number were not predicted as accurately as with the large sample size and minimum replication number. Increasing the sample size would provide more precise  $\theta$  parameters. Lastly, when the RMSE values for  $\theta$  parameters were analyzed, it can be said that the accuracy of  $\theta$  parameters increased as the sample size was increased.

For the 2PL model, when the RMSE values regarding  $\theta$  parameters were examined, it can be said that the accuracy of  $\theta$  parameter estimations increased as sample size was increased for both test lengths. When the SE statistics were analyzed, it was detected that  $\theta$  parameters were estimated most precisely in the 2000-sample size for both test lengths. When the effects of test length in the estimation of  $\theta$  parameters were examined, there were not seen many differences in terms of bias statistics.

For the 3PL model, the estimation accuracy of  $\theta$  parameters increased with increasing sample size for both test lengths. In general, regardless of the sample size, the number of replications did not have a remarkable effect on the accuracy and precision of  $\theta$  parameters. However, the number of replications did have an important effect on the precision of  $\theta$  parameter estimations when for the test length of 50 items and the sample size was 1000. According to findings the sample size had a greater effect on the estimation accuracy of  $\theta$  parameters than the number of replications for all IRT models.

### 3.3. The Effect of the Number of Replications on Estimation of Model Fit with Varying Sample Sizes and Test Lengths

Model fit statistics ( $M_2$  and  $RMSEA_2$ ) were evaluated for all IRT models.  $M_2$  and  $RMSEA_2$  statistics are given respectively in [Figure 5](#) and [Figure 6](#) for all IRT models. According to  $M_2$  values, increasing the test length did not show improvement on the model fit. Additionally, when  $RMSEA_2$  values were examined for the 1PL model, the best model fit was seen in the largest sample size for both test lengths. For both the 2PL model and 3PL model, increasing the test length resulted in decreased/poor model fit in terms of  $M_2$  values. Although not much change was seen,  $RMSEA_2$  values decreased to some extent regardless of the sample size for both the 2PL model and 3PL model. Lastly, it was also detected that regardless of the sample size, the number of replications had no effect on model fit values for both test lengths for all IRT models.

Figure 5.  $M_2$  values for all IRT models.

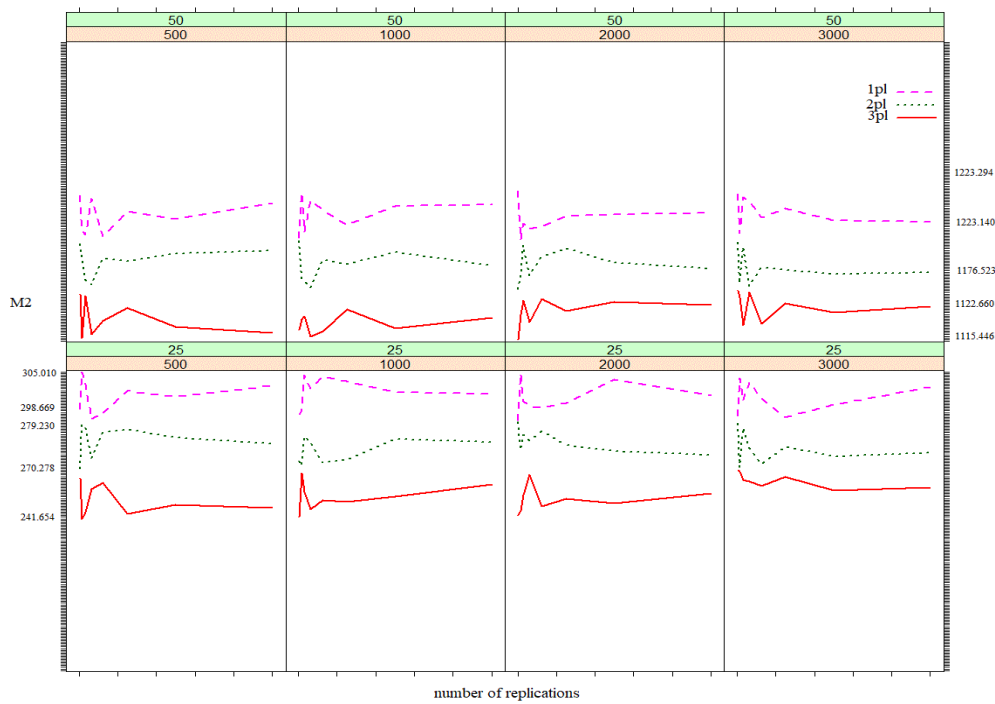
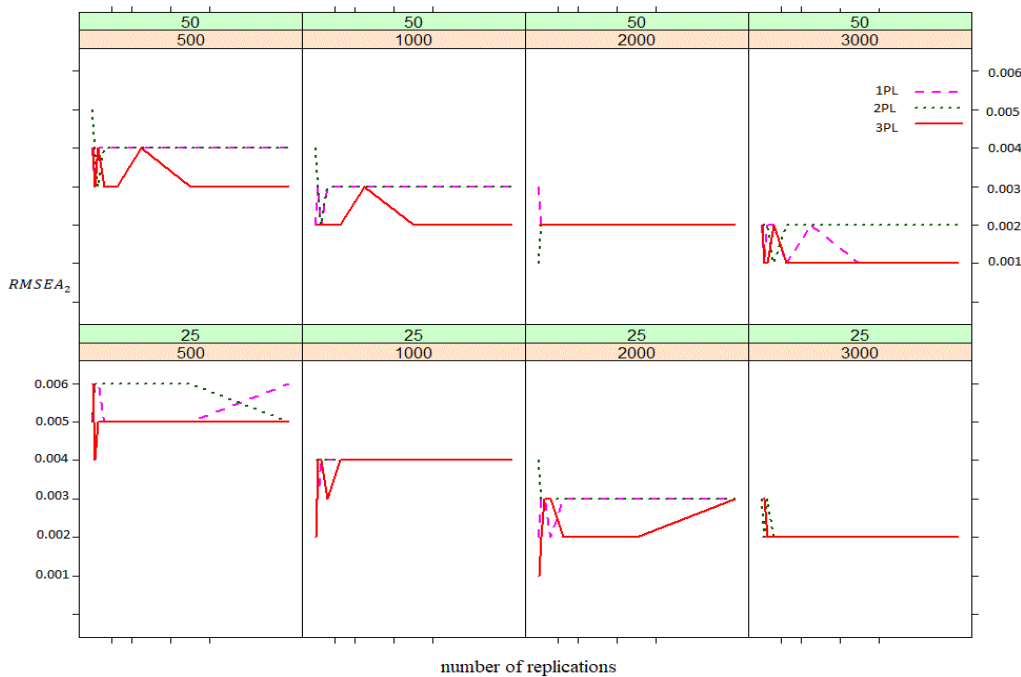


Figure 6.  $RMSEA_2$  values for all IRT models.



In addition to these findings, Type I error inflation rates were calculated according to Glass et al. (1972), and these are presented in Appendix A, Appendix B, and Appendix C. The difference between a particular empirical alpha ( $p$ ) value and the nominal alpha ( $P$ ) value was indicated as significant if departure was two standard errors of  $p$ . When Type I error inflation rates are examined in Appendix A, it is seen that Type I error inflation was only seen at 20 replications for the 1PL model in all sample sizes and test lengths. Also, when test length of 50 items, 78 replications were enough for actual model fit interpretations for the sample sizes 500 and 1000.

For the 2PL model, Type I error inflation, given in [Appendix B](#), was only seen at 20 replications for the test length of 25 items in all sample sizes except when the sample size was 3000. When the sample size was 3000, Type I error inflation was seen at 78 replications also. Type I error inflation was seen at only 20 replications in all sample sizes for the test length of 50 items.

For the 3PL model, in all sample sizes and test lengths Type I error inflation was seen at 20 replications. Additionally, Type I error inflations, given in [Appendix C](#), were seen at 78 replications in both 500 and 3000 sample sizes for the test length 25 items. For the test length 50, Type I error inflation was 78 replications in only 2000 sample sizes. In summary, Type I error rates were not affected except at 20 and 78 replications for all IRT models.

Accuracy of error rate estimation and confidence intervals of empirical alpha ( $p$ ), given in Appendices (A, B, and C), were examined and the same results were achieved for all IRT models. It is important to underline the finding that accuracy of error rate estimation did not change according to either test length or sample size and was affected more by the replication number. The lowest accuracy of error rate was seen at 20 replications for the 1PL model, and at 20 and 78 replications for both the 2PL model and 3PL model. Lastly, the largest confidence interval of empirical alpha ( $p$ ) was seen in the smallest replication number, and that is important in terms of supporting inferences about accuracy.

The main concern of this study is determining a suitable replication number for simulations different test conditions. When test conditions which are determined in this research are considered, findings show that the number of replication effects Type I error inflation. Type I error inflation was seen at 20 and 78 replications. In general, it can be thought that 156, 312, or 625 replications may enough for avoiding Type I error inflation (see [Supplementary](#) file for details). However, other factors, such as item parameter estimation and model fit considered together, it is suggested that at least 625 replications should be performed in terms of Type I error rates.

#### 4. DISCUSSION and CONCLUSION

The purpose of the current study was to determine the required number of replications for the most accurate and precise parameter estimations in conducting MC simulation studies involving unidimensional IRT models. In line with research purpose, different sample sizes and different test lengths were defined as test conditions besides the number of replications.

The first major finding was that neither the test length nor the replication numbers had an effect on item parameter estimation accuracy and precision for all IRT models. On the contrary, the sample size had the largest effect rather than the number of replications in estimation of item parameters. It can be concluded that when the sample is large, even with the smallest number of replications, item parameters can be estimated with adequate precision and accuracy.

Consistent with the current research, Hulin et al. (1982) showed that in the studies of item bias which place emphasis on accuracy, large numbers of items were not necessarily needed. However, they recommended using large samples to obtain accurate item parameter estimates. Besides, they proved that a sample size of 500 for the 2PL model and 1000 for the 3PL model was needed, but also underlined that the more accurate results appeared with a sample size of 2000. Also, consistent with the present study, Ames et al. (2020) found that difficulty parameters had smaller mean bias as sample size was increased for the 2PL model. However, contrary to the present study, they found that increasing the sample size increased the mean bias of discrimination parameters.

When item parameter biases were examined among IRT models, the largest bias values were obtained from 3PL model. It can be concluded that the increase in the number of parameters estimated by the model resulted in more biased estimates.



The study also showed that the best way to increase estimation accuracy of  $\theta$  parameters was to increase the sample size. Contrary to this,  $\theta$  parameters were most precisely estimated among other samples only with 2000 for the 2PL model and 3000 for the 3PL model, and increased test length had no effect on estimation precision like the 1PL model. For the 2PL model and 3PL model, only sample size had an effect on estimation in terms of estimation accuracy of  $\theta$  parameters. The largest sample size had a larger effect on estimation accuracy than the number of replications in both test lengths for all IRT models. This is also consistent with the findings of Hulin et al. (1982), who reported that ability estimates were less accurate in small sample sizes for the 3PL model.

The second major finding was that although the number of replications did not seem to have an effect on the model fit, it was decisive in Type I error inflation and error prediction accuracy for all IRT models. Besides, the most determining factor in model fit was the sample size and long tests had relatively better fit values than short tests. This finding is consistent with that of Schumacker et al. (1994), who found no differences between Rasch item and ability fit statistics based on the number of replications, and the Type I error rates were close to expected values. In accordance with the present study, they recommended being more sensitive to the sample size and test length.

The most obvious finding to emerge from this study was that the sample size had the most important effect on estimation bias for both item parameters and model fit statistics. However, the number of replications was found to be effective on Type I error inflation. Generally, when the number of replications is 20 and 78, Type I error inflation was seen much as per other conditions. When all test conditions determined in this study, especially the accuracy of error rate estimate were evaluated together, accuracy of error rate estimate was seen too close to zero for 625 replications. Besides, also 156 replications and above can be recommended but if the researchers want to get more accurate results, should perform at least 625 replications.

The present study investigated the effect of replication number on the estimation of item and ability estimations and model fit statistics in the MC method based on unidimensional IRT models. It was concluded that the number of replications was not a very impressive factor in the test conditions determined in this study for unidimensional IRT models. In particular, it is seen that sample size is the most effective factor in the estimation of the item and ability parameter and model fit. However, it was concluded that the number of replications is effective in estimating Type I error inflation and accuracy of error rate estimate. In general, as a conclusion of this study, when studying with unidimensional IRT models, it is highly recommended that researchers use large samples instead of studying with small samples and excessive replications.

This study showed that an increase in the number of parameters estimated by the model resulted in increased bias. Therefore, it should be taken into consideration that the adequate number of replications would differ in multi-dimensional models because of increasing estimations of the number of parameters. Similarly, since this study focused on IRT models used with dichotomous items, similar studies could be carried out with polytomous items. All simulations and analyses were performed according to the null (ideal) model. Further research can focus on determining the ideal replication number for misfit data. Due to the fact that it is a simulation study, it is suggested that new studies are conducted on the same condition for generalizations.

## Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

## Authorship Contribution Statement

**Fulya Baris-Pekmezci:** Investigation, Resources, Visualization, Software, Analyze, and Writing. **Asiye Sengul-Avsar:** Investigation, Methodology, Analyze, Supervision, Validation, and Writing.

## ORCID

Fulya BARIŞ PEKMEZCİ  <https://orcid.org/0000-0001-6989-512X>

Asiye ŞENGÜL AVŞAR  <https://orcid.org/0000-0001-5522-2514>

## 5. REFERENCES

- Ames, A. J., Leventhal, B. C., & Ezike, N. C. (2020). Monte Carlo simulation in item response theory applications using SAS. *Measurement: Interdisciplinary Research and Perspectives*, 18(2), 55-74. <https://doi.org/10.1080/15366367.2019.1689762>
- Babcock, B. (2011). Estimating a noncompensatory IRT model using Metropolis within Gibbs sampling. *Applied Psychological Measurement*, 35(4), 317-329. <http://dx.doi.org/10.1177/0146621610392366>
- Bahry, L. M. (2012). Polytomous item response theory parameter recovery: an investigation of nonnormal distributions and small sample size [Master's Thesis]. ProQuest Dissertations and Theses Global.
- Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement*, 22(2), 153-169. <https://doi.org/10.1177/01466216980222005>
- Baldwin, P. (2011). A strategy for developing a common metric in item response theory when parameter posterior distributions are known. *Journal of Educational Measurement*, 48(1), 1-11. Retrieved December 9, 2020, from <http://www.jstor.org/stable/23018061>
- Barış Pekmezci, F., & Gülleroğlu, H. (2019). Investigation of the orthogonality assumption in the bifactor item response theory. *Eurasian Journal of Educational Research*, 19(79), 69-86. <http://dx.doi.org/10.14689/ejer.2019.79.4>
- Bulut, O., & Sünbül, Ö. (2017). Monte Carlo simulation studies in item response theory with the R programming language. *Journal of Measurement and Evaluation in Education and Psychology*, 8(3), 266-287. <https://doi.org/10.21031/epod.305821>
- Cai, L., & Thissen, D. (2014). *Modern Approaches to Parameter Estimation in Item Response Theory from: Handbook of Item Response Theory Modeling, Applications to Typical Performance Assessment*. Routledge.
- Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1-29. <https://doi.org/10.18637/JSS.V048.I06>
- Chuah, S. C., Drasgow, F., & Luecht, R. (2006). How big is big enough? Sample size requirements for CAST item parameter estimation. *Applied Measurement in Education*, 19(3), 241-255. [https://doi.org/10.1207/s15324818ame1903\\_5](https://doi.org/10.1207/s15324818ame1903_5)
- Cohen, A. S., Kim, S. H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement*, 17(4), 335-350. <https://doi.org/10.1177/014662169301700402>

- Crişan, D. R., Tendeiro, J. N., & Meijer, R. R. (2017). Investigating the practical consequences of model misfit in unidimensional IRT models. *Applied Psychological Measurement*, 41(6), 439-455. <https://doi.org/10.1177/0146621617695522>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando: Harcourt Brace Jovanovich Inc.
- De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.
- De La Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, 30(3), 295-311. <https://www.jstor.org/stable/3701380>
- DeMars, C. E. (2002, April). Recovery of graded response and partial credit parameters in multilog and parscale. Annual meeting of American Educational Research Association, Chicago. <https://commons.lib.jmu.edu/cgi/viewcontent.cgi?article=1034&context=gradpsych>
- Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, 35(2), 36-49. <https://doi.org/10.1111/emip.12111>
- Fu, J. (2019). *Maximum marginal likelihood estimation with an expectation–maximization algorithm for multigroup/mixture multidimensional item response theory models* (No. RR-19-35). ETS Research Report Series, <https://doi.org/10.1002/ets2.12272>
- Gao, F., & Chen, L. (2005). Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three-parameter logistic model. *Applied Measurement in Education*, 18(4), 351-380. [https://doi.org/10.1207/s15324818ame1804\\_2](https://doi.org/10.1207/s15324818ame1804_2)
- Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 42(3), 237-288. <https://doi.org/10.3102/00346543042003237>
- Goldman, S. H., & Raju, N. S. (1986). Recovery of one-and two-parameter logistic item parameters: An empirical study. *Educational and Psychological Measurement*, 46(1), 11-21. <https://doi.org/10.1177/0013164486461002>
- Hair, J. F., Black W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis*. (8th edition). Annabel Ainscow.
- Han, K. T. (2007). WinGen: windows software that generates irt parameters and item responses. *Applied Psychological Measurement*, 31(5), 457-459. <https://doi.org/10.1177/0146621607299271>
- Hanson, B. A. (1998, October). IRT parameter estimation using the EM algorithm. <http://www.b-a-h.com/papers/note9801.pdf>
- Harwell, M. (1997). Analyzing the results of monte carlo studies in item response theory. *Educational and Psychological Measurement*, 57(2), 266-279. <https://doi.org/10.1177/0013164497057002006>
- Harwell, M. R., & Baker, F. B. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: A didactic. *Applied Psychological Measurement*, 15(4), 375–389. <https://doi.org/10.1177/014662169101500409>
- Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement*, 20(2), 101-125. <https://doi.org/10.1177/014662169602000201>
- Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement*, 6(3), 249–260. <https://doi.org/10.1177/014662168200600301>
- Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in psychology*, 7, 109. <https://doi.org/10.3389/fpsyg.2016.00109>

- Kirisci, L., Hsu, T. C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25*(2), 146-162. <https://doi.org/10.1177/01466210122031975>
- Kleijnen, J. P. (1987). *Statistical tools for simulation practitioners*. Marcel Dekker.
- Lee, S., Bulut, O., & Suh, Y. (2017). Multidimensional extension of multiple indicators multiple causes models to detect DIF. *Educational and Psychological Measurement, 77*(4), 545–569. <https://doi.org/10.1177/0013164416651116>
- Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*(4), 989-1020. <https://doi.org/10.1177/001316446802800401>
- Matlock, K. L., & Turner, R. (2016). Unidimensional IRT item parameter estimates across equivalent test forms with confounding specifications within dimensions. *Educational and Psychological Measurement, 76*(2), 258-279. <https://doi.org/10.1177/0013164415589756>
- Matlock Cole, K., & Paek, I. (2017). PROC IRT: A SAS procedure for item response theory. *Applied Psychological Measurement, 41*(4), 311-320. <https://doi.org/10.1177/0146621616685062>
- McDonald, R. P. (1982). Linear Versus Models in Item Response Theory. *Applied Psychological Measurement, 6*(4), 379-396. <https://doi.org/10.1177/014662168200600402>
- Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*(1), 57-75. <https://doi.org/10.1177/014662168901300106>
- Mooney, C. Z. (1997). *Monte Carlo simulation*. Thousand Oaks, CA: Sage.
- Mundform, D. J., Schaffer, J., Kim, M. J., Shaw, D., Thongteeraparp, A., & Supawan, P. (2011). Number of replications required in Monte Carlo simulation studies: A synthesis of four studies. *Journal of Modern Applied Statistical Methods, 10*(1), 19-28. <https://doi.org/10.22237/jmasm/1304222580>
- Park, Y. S., Lee, Y. S., & Xing, K. (2016). Investigating the impact of item parameter drift for item response theory models with mixture distributions. *Frontiers in Psychology, 7*, 255. <https://doi.org/10.3389/fpsyg.2016.00255>
- Patsias, K., Sheng, Y., & Rahimi, S. (2009, September 24-26). A high performance Gibbs sampling algorithm for item response theory. 22nd International Conference on Parallel and Distributed Computing and Communication Systems, Kentucky, USA.
- Patsula, L. N., & Gessaroli, M. E. A (1995, April). Comparison of item parameter estimates and iccs produced. <https://files.eric.ed.gov/fulltext/ED414333.pdf>
- Preecha, C. (2004). Numbers of replications required in ANOVA simulation studies [Doctoral dissertation, University of Northern Colorado]. ProQuest Dissertations and Theses Global.
- Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*(2), 133-144. <https://www.jstor.org/stable/1434973>
- Reise, S., Moore, T., & Maydeu-Olivares, A. (2011). Target rotations and assessing the impact of model violations on the parameters of unidimensional item response theory models. *Educational and Psychological Measurement, 71*(4), 684-711. <https://doi.org/10.1177/0013164410378690>
- Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2002). Characteristics of MML/EAP parameter estimates in the generalized graded unfolding model. *Applied Psychological Measurement, 26*(2), 192-207. <https://doi.org/10.1177/01421602026002006>



- Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo method*. John Wiley and Sons, New York. <https://doi.org/10.1002/9780470316511>
- Sahin, A., & Anil, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice*, 17(1), 321-33. <https://doi.org/10.12738/estp.2017.1.0270>
- Sarkar, D. (2008). *Lattice: multivariate data visualization with R*. Springer, New York.
- Schumacker, R. E, Smith, R. M., & Bush, J. M. (1994, April). *Examining replication effects in Rasch fit statistics*. American Educational Research Association Annual Meeting, New Orleans.
- Sen, S., Cohen, A. S., & Kim, S. H. (2016). The impact of non-normality on extraction of spurious latent classes in mixture IRT models. *Applied Psychological Measurement*, 40(2), 98-113. <https://doi.org/10.1177/0146621615605080>
- Sheng, Y., & Wikle, C. K. (2007). Comparing multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement*, 67(6), 899-919. <https://doi.org/10.1177/0013164406296977>
- Tavares, H. R., Andrade, D. F. D., & Pereira, C. A. D. B. (2004). Detection of determinant genes and diagnostic via item response theory. *Genetics and Molecular Biology*, 27(4), 679-685. <https://doi.org/10.1590/S1415-47572004000400033>
- Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, 47(4), 397-412. <https://doi.org/10.1007/BF02293705>
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis*. Amer Psychological Assn.
- Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. Guilford Press.
- Thompson, N. A. (2009). Ability estimation with item response theory. *Assessment Systems Corporation*. [https://assess.com/docs/Thompson \(2009\) - Ability estimation with IR T.pdf](https://assess.com/docs/Thompson%20(2009)%20-%20Ability%20estimation%20with%20IRT.pdf)
- Şengül Avşar, A., & Tavşancıl, E. (2017). Examination of polytomous items' psychometric properties according to nonparametric item response theory models in different test conditions. *Educational Sciences: Theory & Practice*, 17(2). <https://doi.org/10.12738/estp.2017.2.0246>
- van der Linden, W. J. (Ed.). (2018). *Handbook of item response theory, three volume set*. CRC Press.
- van Onna, M. J. H. (2004). Ordered latent class models in nonparametric item response theory. [Doctoral dissertation]. University of Groningen.
- Weissman, A. (2013). Optimizing information using the EM algorithm in item response theory. *Annals of Operations Research*, 206(1), 627-646. <https://doi.org/10.1007/s10479-012-1204-4>
- Yang, S. (2007). A comparison of unidimensional and multidimensional RASCH models using parameter estimates and fit indices when assumption of unidimensionality is violated [Doctoral dissertation, The Ohio State University]. ProQuest Dissertations and Theses Global.
- Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika*, 52(2), 275-291. <https://doi.org/10.1007/BF02294241>
- Zhang, B. (2008). Application of unidimensional item response models to tests with items sensitive to secondary dimensions. *The Journal of Experimental Education*, 77(2), 147-166. <https://doi.org/10.3200/JEXE.77.2.147-166>

## 6. APPENDIX

## 6.1. Appendix A

**Table A1.** Type I error rate and accuracy of error estimate from 25 items for IPL model.

Sample size	Number of Replication	Empirical alpha ( $p$ )	Empirical alpha ( $p$ )-nominal ( $P$ ) alpha	$p-2\hat{\sigma}_p$	$p+2\hat{\sigma}_p$	Accuracy of error rate estimate
3000	5000	0.493	0.443	0.486	0.499	0.000
	2500	0.502	0.452	0.493	0.511	0.000
	1250	0.515	0.465	0.503	0.527	0.000
	625	0.495	0.445	0.478	0.512	0.001
	312	0.488	0.438	0.463	0.512	0.002
	156	0.493	0.443	0.459	0.528	0.003
	78	0.485	0.435	0.436	0.535	0.006
	20	0.515	0.465	0.418	0.613	0.025
2000	5000	0.497	0.447	0.491	0.503	0.000
	2500	0.489	0.439	0.480	0.497	0.000
	1250	0.501	0.451	0.489	0.514	0.000
	625	0.501	0.451	0.483	0.518	0.001
	312	0.502	0.452	0.478	0.527	0.002
	156	0.497	0.447	0.462	0.532	0.003
	78	0.464	0.414	0.415	0.513	0.006
	20	0.521	0.471	0.424	0.619	0.025
1000	5000	0.496	0.446	0.490	0.502	0.000
	2500	0.495	0.445	0.486	0.504	0.000
	1250	0.499	0.449	0.486	0.511	0.000
	625	0.488	0.438	0.476	0.501	0.000
	312	0.481	0.431	0.464	0.499	0.001
	156	0.489	0.439	0.465	0.514	0.002
	78	0.470	0.420	0.435	0.505	0.003
	20	0.508	0.458	0.458	0.557	0.006
500	5000	0.491	0.441	0.484	0.497	0.000
	2500	0.495	0.445	0.486	0.504	0.000
	1250	0.491	0.441	0.478	0.503	0.000
	625	0.507	0.457	0.490	0.525	0.001
	312	0.527	0.477	0.502	0.552	0.002
	156	0.490	0.440	0.455	0.525	0.003
	78	0.434	0.384	0.385	0.484	0.006
	20	0.508	0.458	0.411	0.605	0.025



**Table A2.** Type I error rate and accuracy of error estimate from 50 items for IPL model.

Sample size	Number of Replication	Empirical alpha ( $p$ )	Empirical alpha ( $p$ )-nominal ( $P$ ) alpha	$p-2\hat{\sigma}_p$	$p+2\hat{\sigma}_p$	Accuracy of error rate estimate
3000	5000	0.505	0.455	0.498	0.511	0.000
	2500	0.504	0.454	0.496	0.513	0.000
	1250	0.499	0.449	0.487	0.511	0.000
	625	0.501	0.451	0.483	0.518	0.001
	312	0.480	0.430	0.455	0.505	0.002
	156	0.464	0.414	0.429	0.499	0.003
	78	0.526	0.476	0.477	0.575	0.006
	20	0.464	0.414	0.367	0.562	0.025
2000	5000	0.502	0.452	0.496	0.508	0.000
	2500	0.501	0.451	0.492	0.510	0.000
	1250	0.500	0.450	0.488	0.513	0.000
	625	0.507	0.457	0.490	0.524	0.001
	312	0.506	0.456	0.481	0.530	0.002
	156	0.510	0.460	0.475	0.545	0.003
	78	0.558	0.508	0.509	0.608	0.006
	20	0.411	0.361	0.313	0.508	0.025
1000	5000	0.491	0.441	0.484	0.497	0.000
	2500	0.498	0.448	0.489	0.506	0.000
	1250	0.505	0.455	0.492	0.517	0.000
	625	0.499	0.449	0.482	0.517	0.001
	312	0.477	0.427	0.452	0.502	0.002
	156	0.523	0.473	0.488	0.558	0.003
	78	0.426	0.376	0.376	0.475	0.006
	20	0.539	0.489	0.441	0.636	0.025
500	5000	0.486	0.436	0.480	0.492	0.000
	2500	0.483	0.433	0.474	0.491	0.000
	1250	0.489	0.439	0.476	0.501	0.000
	625	0.489	0.439	0.471	0.506	0.001
	312	0.487	0.437	0.462	0.511	0.002
	156	0.491	0.441	0.456	0.526	0.003
	78	0.516	0.466	0.466	0.565	0.006
	20	0.445	0.395	0.348	0.543	0.025

## 6.2. Appendix B

Table B1. Type I error rate and accuracy of error estimate from 25 items for 2PL model.

Sample size	Number of Replication	Empirical alpha ( $p$ )	Empirical alpha ( $p$ )-nominal ( $P$ ) alpha	$p-2\hat{\sigma}_p$	$p+2\hat{\sigma}_p$	Accuracy of error rate estimate
3000	5000	0.496	0.446	0.490	0.502	0.000
	2500	0.500	0.450	0.491	0.508	0.000
	1250	0.493	0.443	0.481	0.506	0.000
	625	0.519	0.469	0.502	0.536	0.001
	312	0.496	0.446	0.472	0.521	0.002
	156	0.476	0.426	0.441	0.511	0.003
	78	0.528	0.478	0.478	0.577	0.006
	20	0.453	0.403	0.356	0.550	0.025
2000	5000	0.499	0.449	0.493	0.505	0.000
	2500	0.492	0.442	0.483	0.501	0.000
	1250	0.490	0.440	0.478	0.502	0.000
	625	0.475	0.425	0.457	0.492	0.001
	312	0.489	0.439	0.465	0.514	0.002
	156	0.483	0.433	0.448	0.518	0.003
	78	0.501	0.451	0.452	0.551	0.006
	20	0.325	0.275	0.228	0.423	0.023
1000	5000	0.489	0.439	0.483	0.495	0.000
	2500	0.486	0.436	0.477	0.494	0.000
	1250	0.502	0.452	0.490	0.515	0.000
	625	0.511	0.461	0.494	0.529	0.001
	312	0.490	0.440	0.465	0.514	0.002
	156	0.487	0.437	0.452	0.522	0.003
	78	0.509	0.459	0.460	0.558	0.006
	20	0.502	0.452	0.405	0.599	0.025
500	5000	0.491	0.441	0.485	0.498	0.000
	2500	0.486	0.436	0.477	0.495	0.000
	1250	0.476	0.426	0.463	0.488	0.000
	625	0.477	0.427	0.459	0.494	0.001
	312	0.496	0.446	0.472	0.521	0.002
	156	0.452	0.402	0.417	0.487	0.003
	78	0.451	0.401	0.402	0.500	0.006
	20	0.545	0.495	0.447	0.642	0.025

**Table B2.** *Type I error rate and accuracy of error estimate from 50 items for 2PL model.*

Sample size	Number of Replication	Empirical alpha ( $p$ )	Empirical alpha ( $p$ )-nominal ( $P$ ) alpha	$p-2\hat{\sigma}_p$	$p+2\hat{\sigma}_p$	Accuracy of error rate estimate
3000	5000	0.490	0.440	0.483	0.497	0.000
	2500	0.491	0.441	0.482	0.500	0.000
	1250	0.485	0.435	0.473	0.497	0.000
	625	0.483	0.433	0.466	0.500	0.001
	312	0.515	0.465	0.490	0.540	0.002
	156	0.455	0.405	0.420	0.490	0.003
	78	0.513	0.463	0.464	0.562	0.006
	20	0.419	0.369	0.322	0.516	0.025
2000	5000	0.483	0.433	0.477	0.489	0.000
	2500	0.478	0.428	0.469	0.487	0.000
	1250	0.467	0.417	0.454	0.479	0.000
	625	0.476	0.426	0.459	0.494	0.001
	312	0.509	0.459	0.484	0.533	0.002
	156	0.428	0.378	0.393	0.463	0.003
	78	0.492	0.442	0.443	0.541	0.006
	20	0.531	0.481	0.433	0.628	0.025
1000	5000	0.481	0.431	0.475	0.487	0.000
	2500	0.470	0.420	0.461	0.479	0.000
	1250	0.480	0.430	0.468	0.492	0.000
	625	0.474	0.424	0.457	0.491	0.001
	312	0.522	0.472	0.497	0.547	0.002
	156	0.512	0.462	0.477	0.547	0.003
	78	0.510	0.460	0.461	0.559	0.006
	20	0.379	0.329	0.282	0.476	0.024
500	5000	0.471	0.421	0.465	0.478	0.000
	2500	0.471	0.421	0.463	0.480	0.000
	1250	0.478	0.428	0.466	0.490	0.000
	625	0.477	0.427	0.460	0.495	0.001
	312	0.516	0.466	0.492	0.541	0.002
	156	0.513	0.463	0.478	0.548	0.003
	78	0.476	0.426	0.427	0.525	0.006
	20	0.426	0.376	0.329	0.524	0.025

## 6.3. Appendix C

Table C1. Type I error rate and accuracy of error estimate from 25 items for 3PL model.

Sample size	Number of Replication	Empirical alpha ( $p$ )	Empirical alpha ( $p$ )-nominal ( $P$ ) alpha	$p-2\hat{\sigma}_p$	$p+2\hat{\sigma}_p$	Accuracy of error rate estimate
3000	5000	0.517	0.467	0.510	0.523	0.000
	2500	0.519	0.469	0.510	0.527	0.000
	1250	0.517	0.467	0.504	0.529	0.000
	625	0.527	0.477	0.510	0.544	0.001
	312	0.491	0.441	0.466	0.515	0.002
	156	0.530	0.480	0.495	0.565	0.003
	78	0.498	0.448	0.448	0.547	0.006
	20	0.445	0.395	0.348	0.543	0.025
2000	5000	0.514	0.464	0.508	0.520	0.000
	2500	0.510	0.460	0.501	0.519	0.000
	1250	0.518	0.468	0.505	0.530	0.000
	625	0.500	0.450	0.483	0.518	0.001
	312	0.533	0.483	0.508	0.558	0.002
	156	0.511	0.461	0.476	0.546	0.003
	78	0.499	0.449	0.450	0.549	0.006
	20	0.564	0.514	0.467	0.662	0.025
1000	5000	0.523	0.473	0.517	0.530	0.000
	2500	0.530	0.480	0.521	0.539	0.000
	1250	0.515	0.465	0.502	0.527	0.000
	625	0.531	0.481	0.513	0.548	0.001
	312	0.543	0.493	0.518	0.567	0.002
	156	0.525	0.475	0.490	0.560	0.003
	78	0.518	0.468	0.469	0.568	0.006
	20	0.531	0.481	0.434	0.629	0.025
500	5000	0.531	0.481	0.524	0.537	0.000
	2500	0.526	0.476	0.517	0.535	0.000
	1250	0.519	0.469	0.507	0.532	0.000
	625	0.521	0.471	0.503	0.538	0.001
	312	0.539	0.489	0.515	0.564	0.002
	156	0.501	0.451	0.466	0.536	0.003
	78	0.552	0.502	0.502	0.601	0.006
	20	0.467	0.417	0.369	0.564	0.025

**Table C2.** Type I error rate and accuracy of error estimate from 50 items for 3PL model.

Sample size	Number of Replication	Empirical alpha ( $p$ )	Empirical alpha ( $p$ )-nominal ( $P$ ) alpha	$p-2\hat{\sigma}_p$	$p+2\hat{\sigma}_p$	Accuracy of error rate estimate
3000	5000	0.512	0.462	0.506	0.518	0.000
	2500	0.516	0.466	0.507	0.524	0.000
	1250	0.495	0.445	0.482	0.507	0.000
	625	0.510	0.460	0.492	0.527	0.001
	312	0.505	0.455	0.480	0.530	0.002
	156	0.494	0.444	0.459	0.529	0.003
	78	0.455	0.405	0.406	0.504	0.006
	20	0.421	0.371	0.324	0.519	0.025
2000	5000	0.515	0.465	0.509	0.521	0.000
	2500	0.521	0.471	0.512	0.530	0.000
	1250	0.521	0.471	0.509	0.534	0.000
	625	0.521	0.471	0.503	0.538	0.001
	312	0.493	0.443	0.469	0.518	0.002
	156	0.522	0.472	0.487	0.557	0.003
	78	0.549	0.499	0.499	0.598	0.006
	20	0.593	0.543	0.496	0.691	0.025
1000	5000	0.512	0.462	0.505	0.518	0.000
	2500	0.520	0.470	0.511	0.529	0.000
	1250	0.522	0.472	0.510	0.535	0.000
	625	0.519	0.469	0.501	0.536	0.001
	312	0.528	0.478	0.503	0.552	0.002
	156	0.507	0.457	0.472	0.541	0.003
	78	0.498	0.448	0.449	0.548	0.006
	20	0.575	0.525	0.477	0.672	0.025
500	5000	0.526	0.476	0.520	0.532	0.000
	2500	0.520	0.470	0.511	0.528	0.000
	1250	0.541	0.491	0.528	0.553	0.000
	625	0.508	0.458	0.490	0.525	0.001
	312	0.513	0.463	0.488	0.537	0.002
	156	0.542	0.492	0.507	0.577	0.003
	78	0.592	0.542	0.543	0.642	0.006
	20	0.501	0.451	0.403	0.598	0.025



## 7. SUPPLEMENTARY FILE

### 7. 1. Supplementary File for 1PL Model

# of replication	Sample.size	# of item	me.b	se.b	rmse.b	me.theta	se.theta	rmse.theta	M2	M2.p	RMSEA.2	Log-Likelihood	Eq.5	empirical-nominal alpha	-2	2	accuracy of error rate estimate
5000	3000	25	0.000	1.457	0.054	0.000	0.055	0.014	299.673	0.493	0.002	-38193.147	0.003	0.443	0.486	0.499	0.000
2500	3000	25	0.000	1.456	0.054	0.000	0.056	0.015	298.717	0.502	0.002	-38237.706	0.004	0.452	0.493	0.511	0.000
1250	3000	25	0.000	1.453	0.054	0.000	0.055	0.015	297.623	0.515	0.002	-38204.417	0.006	0.465	0.503	0.527	0.000
625	3000	25	-0.002	1.460	0.054	0.001	0.053	0.015	299.390	0.495	0.002	-38236.766	0.009	0.445	0.478	0.512	0.000
312	3000	25	-0.001	1.492	0.054	0.001	0.055	0.015	299.914	0.488	0.002	-37967.818	0.012	0.438	0.463	0.512	0.001
156	3000	25	-0.002	1.449	0.054	0.002	0.049	0.016	299.200	0.493	0.002	-38197.122	0.017	0.443	0.459	0.528	0.001
78	3000	25	0.006	1.439	0.054	-0.004	0.061	0.014	300.422	0.485	0.002	-38496.120	0.025	0.435	0.436	0.535	0.003
20	3000	25	0.007	1.453	0.055	-0.002	0.071	0.021	297.788	0.515	0.002	-38182.517	0.049	0.465	0.418	0.613	0.011
# of replication	Sample.size	# of item	me.b	se.b	rmse.b	me.theta	se.theta	rmse.theta	M2	M2.p	RMSEA.2	Log-Likelihood	Eq.5	empirical-nominal alpha	-2	2	accuracy of error rate estimate
5000	2000	25	0.001	1.456	0.067	0.000	0.062	0.018	299.396	0.497	0.003	-25477.162	0.003	0.447	0.491	0.503	0.000
2500	2000	25	0.000	1.461	0.067	0.000	0.061	0.018	300.012	0.489	0.003	-25457.807	0.004	0.439	0.480	0.497	0.000
1250	2000	25	0.001	1.458	0.067	0.000	0.063	0.018	298.774	0.501	0.003	-25464.042	0.006	0.451	0.489	0.514	0.000
625	2000	25	0.000	1.442	0.066	0.000	0.061	0.017	298.672	0.501	0.003	-25528.717	0.009	0.451	0.483	0.518	0.000
312	2000	25	-0.001	1.446	0.066	0.002	0.057	0.017	298.688	0.502	0.002	-25480.706	0.012	0.452	0.478	0.527	0.001
156	2000	25	-0.001	1.457	0.066	0.002	0.054	0.016	299.054	0.497	0.003	-25528.210	0.017	0.447	0.462	0.532	0.001
78	2000	25	-0.001	1.445	0.064	-0.002	0.060	0.015	304.377	0.464	0.003	-25325.638	0.025	0.414	0.415	0.513	0.003
20	2000	25	-0.001	1.486	0.063	0.002	0.047	0.014	296.433	0.521	0.002	-25390.219	0.049	0.471	0.424	0.619	0.011
# of replication	Sample.size	# of item	me.b	se.b	rmse.b	me.theta	se.theta	rmse.theta	M2	M2.p	RMSEA.2	Log-Likelihood	Eq.5	empirical-nominal alpha	-2	2	accuracy of error rate estimate
5000	1000	25	0.001	1.461	0.095	0.000	0.073	0.025	299.406	0.496	0.004	-12730.860	0.003	0.446	0.490	0.502	0.000
2500	1000	25	0.001	1.466	0.095	0.000	0.074	0.026	299.449	0.495	0.004	-12711.550	0.004	0.445	0.486	0.504	0.000
1250	1000	25	0.003	1.461	0.095	-0.002	0.078	0.025	299.915	0.488	0.004	-12700.168	0.006	0.438	0.476	0.501	0.000
625	1000	25	0.001	1.456	0.093	0.000	0.073	0.025	300.639	0.481	0.004	-12739.327	0.009	0.431	0.464	0.499	0.000
312	1000	25	0.000	1.481	0.095	0.002	0.065	0.024	299.650	0.489	0.004	-12687.959	0.012	0.439	0.465	0.514	0.001
156	1000	25	-0.002	1.437	0.091	0.002	0.071	0.026	301.334	0.470	0.004	-12823.181	0.017	0.420	0.435	0.505	0.001
78	1000	25	0.003	1.501	0.096	-0.001	0.078	0.027	298.567	0.508	0.003	-12605.717	0.025	0.458	0.458	0.557	0.003
20	1000	25	0.010	1.462	0.106	-0.010	0.116	0.035	297.865	0.521	0.003	-12561.626	0.049	0.471	0.424	0.619	0.011
# of replication	sample.size	# of item	me.b	se.b	rmse.b	me.theta	se.theta	rmse.theta	M2	M2.p	RMSEA.2	Log-Likelihood	Eq.5	empirical-nominal alpha	-2	2	accuracy of error rate estimate
5000	500	25	0.002	1.465	0.134	0.001	0.176	0.036	299.756	0.491	0.005	-6355.279	0.003	0.441	0.484	0.497	0.000
2500	500	25	0.001	1.463	0.134	0.001	0.175	0.036	299.393	0.495	0.005	-6366.520	0.004	0.445	0.486	0.504	0.000
1250	500	25	0.003	1.460	0.134	0.001	0.173	0.035	299.465	0.491	0.005	-6363.194	0.006	0.441	0.478	0.503	0.000
625	500	25	0.001	1.460	0.136	0.002	0.176	0.036	298.152	0.507	0.005	-6360.802	0.009	0.457	0.490	0.524	0.000
312	500	25	0.002	1.441	0.136	0.001	0.183	0.037	297.027	0.527	0.004	-6390.473	0.012	0.477	0.502	0.552	0.001
156	500	25	-0.003	1.491	0.135	0.002	0.174	0.036	299.843	0.490	0.006	-6330.255	0.017	0.440	0.455	0.525	0.001
78	500	25	-0.006	1.491	0.135	0.007	0.143	0.032	305.010	0.434	0.006	-6306.223	0.025	0.384	0.385	0.484	0.003
20	500	25	0.024	1.496	0.147	-0.014	0.228	0.043	298.669	0.508	0.005	-6289.242	0.049	0.458	0.411	0.605	0.011

# of replication	Sample.size	# of item	me.b	se.b	rmse.b	me.theta	se.theta	rmse.theta	M2	M2.p	RMSEA.2	Log-Likelihood	Eq.5	empirical-nominal alpha	-2	2	accuracy of error rate estimate
5000	3000	50	0.000	1.483	0.055	0.000	0.056	0.015	1223.140	0.505	0.001	-74852.745	0.003	0.455	0.498	0.511	0.000
2500	3000	50	0.000	1.480	0.055	0.000	0.056	0.014	1223.545	0.504	0.001	-74894.242	0.004	0.454	0.496	0.513	0.000
1250	3000	50	0.001	1.477	0.054	0.000	0.058	0.015	1224.342	0.499	0.002	-75013.808	0.006	0.449	0.487	0.511	0.000
625	3000	50	-0.001	1.481	0.055	0.001	0.053	0.015	1223.860	0.501	0.001	-74771.204	0.009	0.451	0.483	0.518	0.000
312	3000	50	-0.001	1.477	0.055	0.001	0.053	0.015	1226.967	0.480	0.002	-74905.122	0.012	0.430	0.455	0.505	0.001
156	3000	50	0.001	1.473	0.054	-0.001	0.061	0.014	1229.901	0.464	0.002	-75123.100	0.017	0.414	0.429	0.499	0.001
78	3000	50	0.002	1.473	0.054	-0.001	0.061	0.015	1220.230	0.526	0.001	-74769.185	0.025	0.476	0.477	0.575	0.003
20	3000	50	-0.003	1.484	0.053	0.004	0.044	0.012	1232.012	0.464	0.002	-75249.604	0.049	0.414	0.367	0.562	0.011
<b># of replication</b>	<b>sample.size</b>	<b># of item</b>	<b>me.b</b>	<b>se.b</b>	<b>rmse.b</b>	<b>me.theta</b>	<b>se.theta</b>	<b>rmse.theta</b>	<b>M2</b>	<b>M2.p</b>	<b>RMSEA.2</b>	<b>Log-Likelihood</b>	<b>Eq.5</b>	<b>empirical-nominal alpha</b>	<b>-2</b>	<b>2</b>	<b>accuracy of error rate estimate</b>
5000	2000	50	0.000	1.478	0.067	0.000	0.123	0.018	1223.924	0.502	0.002	-49959.823	0.003	0.452	0.496	0.508	0.000
2500	2000	50	0.001	1.476	0.067	0.000	0.123	0.018	1223.882	0.501	0.002	-49988.694	0.004	0.451	0.492	0.510	0.000
1250	2000	50	0.002	1.485	0.067	-0.001	0.126	0.018	1223.874	0.500	0.002	-49928.566	0.006	0.450	0.488	0.513	0.000
625	2000	50	0.000	1.480	0.067	0.001	0.123	0.018	1222.493	0.507	0.002	-49963.840	0.009	0.457	0.490	0.524	0.000
312	2000	50	0.000	1.479	0.067	0.002	0.056	0.017	1222.136	0.506	0.002	-49932.391	0.012	0.456	0.481	0.530	0.001
156	2000	50	0.002	1.490	0.068	-0.002	0.118	0.017	1222.925	0.510	0.002	-49758.711	0.017	0.460	0.475	0.545	0.001
78	2000	50	0.001	1.477	0.066	-0.001	0.126	0.019	1213.061	0.558	0.002	-50098.217	0.025	0.508	0.509	0.608	0.003
20	2000	50	-0.015	1.467	0.067	0.013	0.030	0.021	1242.775	0.411	0.003	-50266.496	0.049	0.361	0.313	0.508	0.011
<b># of replication</b>	<b>Sample.size</b>	<b># of item</b>	<b>me.b</b>	<b>se.b</b>	<b>rmse.b</b>	<b>me.theta</b>	<b>se.theta</b>	<b>rmse.theta</b>	<b>M2</b>	<b>M2.p</b>	<b>RMSEA.2</b>	<b>Log-Likelihood</b>	<b>Eq.5</b>	<b>empirical-nominal alpha</b>	<b>-2</b>	<b>2</b>	<b>accuracy of error rate estimate</b>
5000	1000	50	-0.001	1.482	0.095	0.002	0.145	0.025	1225.421	0.491	0.003	-24952.975	0.003	0.441	0.484	0.497	0.000
2500	1000	50	0.000	1.492	0.096	0.000	0.074	0.026	1225.135	0.498	0.003	-24924.722	0.004	0.448	0.489	0.506	0.000
1250	1000	50	0.001	1.484	0.095	0.000	0.072	0.025	1222.830	0.505	0.003	-24938.207	0.006	0.455	0.492	0.517	0.000
625	1000	50	-0.001	1.485	0.095	0.001	0.071	0.026	1224.304	0.499	0.003	-24909.186	0.009	0.449	0.482	0.517	0.000
312	1000	50	0.002	1.485	0.095	-0.001	0.073	0.023	1227.370	0.477	0.003	-24949.250	0.012	0.427	0.452	0.502	0.001
156	1000	50	-0.002	1.489	0.095	0.004	0.064	0.024	1220.446	0.523	0.002	-24928.376	0.017	0.473	0.488	0.558	0.001
78	1000	50	-0.001	1.483	0.093	0.002	0.069	0.027	1237.680	0.426	0.003	-24993.260	0.025	0.376	0.376	0.475	0.003
20	1000	50	-0.013	1.484	0.098	0.013	0.054	0.034	1214.549	0.539	0.002	-24920.461	0.049	0.489	0.441	0.636	0.011
<b># of replication</b>	<b>Sample.size</b>	<b># of item</b>	<b>me.b</b>	<b>se.b</b>	<b>rmse.b</b>	<b>me.theta</b>	<b>se.theta</b>	<b>rmse.theta</b>	<b>M2</b>	<b>M2.p</b>	<b>RMSEA.2</b>	<b>Log-Likelihood</b>	<b>Eq.5</b>	<b>empirical-nominal alpha</b>	<b>-2</b>	<b>2</b>	<b>accuracy of error rate estimate</b>
5000	500	50	0.002	1.489	0.135	0.000	0.087	0.036	1226.638	0.486	0.004	-12468.483	0.003	0.436	0.480	0.492	0.000
2500	500	50	0.003	1.487	0.135	-0.001	0.089	0.036	1223.561	0.483	0.004	-12437.121	0.004	0.433	0.474	0.491	0.000
1250	500	50	0.002	1.493	0.135	0.001	0.085	0.036	1224.014	0.489	0.004	-12424.144	0.006	0.439	0.476	0.501	0.000
625	500	50	0.004	1.480	0.133	-0.001	0.090	0.035	1217.708	0.489	0.004	-12376.198	0.009	0.439	0.471	0.506	0.000
312	500	50	0.001	1.489	0.135	0.002	0.091	0.040	1227.775	0.487	0.004	-12461.296	0.012	0.437	0.462	0.511	0.001
156	500	50	0.002	1.495	0.135	0.001	0.087	0.037	1217.750	0.491	0.004	-12340.218	0.017	0.441	0.456	0.526	0.001
78	500	50	-0.001	1.521	0.138	0.002	0.082	0.033	1221.873	0.516	0.003	-12380.931	0.025	0.466	0.466	0.565	0.003
20	500	50	-0.002	1.510	0.139	0.009	0.066	0.034	1231.039	0.445	0.004	-12343.249	0.049	0.395	0.348	0.543	0.011

## 7. 2. Supplementary File for 2PL Model

# of replication	Sample.size	# of item	me.a	se.a	rmse.a	me.b	se.b	rmse.b	me.theta	se.theta	rmse.theta	M2	M2.p	RMSEA.2	Log-Likelihood	Eq.5	empirical-nominal alpha	-2	2	accuracy of error rate estimate
5000	3000	25	0.004	0.163	0.085	0.003	1.458	0.075	-0.001	0.110	0.014	275.439	0.496	0.002	-35127.778	0.003	0.446	0.490	0.502	0.000
2500	3000	25	0.004	0.162	0.085	0.003	1.457	0.075	-0.001	0.111	0.014	275.0509	0.500	0.002	-35152.720	0.004	0.450	0.491	0.508	0.000
1250	3000	25	0.004	0.162	0.085	0.002	1.465	0.076	0.000	0.114	0.015	275.7612	0.493	0.002	-35107.61051	0.006	0.443	0.481	0.506	0.000
625	3000	25	0.004	0.164	0.086	0.004	1.469	0.076	-0.001	0.114	0.014	273.6207	0.519	0.002	-35123.35379	0.009	0.469	0.502	0.536	0.000
312	3000	25	0.005	0.162	0.085	0.005	1.462	0.076	-0.001	0.112	0.014	275.6247	0.496	0.002	-35110.32823	0.012	0.446	0.472	0.521	0.001
156	3000	25	0.003	0.164	0.085	0.000	1.467	0.074	0.000	0.109	0.014	277.8721	0.476	0.003	-35033.99359	0.017	0.426	0.441	0.511	0.001
78	3000	25	0.003	0.162	0.086	-0.002	1.451	0.076	0.002	0.111	0.013	272.1231	0.528	0.002	-35076.88341	0.025	0.478	0.478	0.577	0.003
20	3000	25	0.002	0.167	0.086	-0.009	1.479	0.073	0.003	0.095	0.013	279.572	0.453	0.003	-35204.453	0.049	0.403	0.356	0.550	0.011
<b># of replication</b>	<b>Sample.size</b>	<b># of item</b>	<b>me.a</b>	<b>se.a</b>	<b>rmse.a</b>	<b>me.b</b>	<b>se.b</b>	<b>rmse.b</b>	<b>me.theta</b>	<b>se.theta</b>	<b>rmse.theta</b>	<b>M2</b>	<b>M2.p</b>	<b>RMSEA.2</b>	<b>Log-Likelihood</b>	<b>Eq.5</b>	<b>empirical-nominal alpha</b>	<b>-2</b>	<b>2</b>	<b>accuracy of error rate estimate</b>
5000	2000	25	0.005	0.172	0.105	0.003	1.463	0.092	0.000	0.061	0.018	275.122	0.499	0.003	-23400.047	0.003	0.449	0.493	0.505	0.000
2500	2000	25	0.004	0.172	0.105	0.004	1.462	0.094	-0.001	0.062	0.018	275.495	0.492	0.003	-23421.325	0.004	0.442	0.483	0.501	0.000
1250	2000	25	0.005	0.172	0.105	0.005	1.460	0.094	-0.002	0.064	0.018	275.768	0.490	0.003	-23418.543	0.006	0.440	0.478	0.502	0.000
625	2000	25	0.004	0.172	0.104	0.003	1.443	0.091	-0.001	0.061	0.018	277.119	0.475	0.003	-23474.757	0.009	0.425	0.457	0.492	0.000
312	2000	25	0.002	0.170	0.105	0.004	1.467	0.094	-0.001	0.062	0.019	275.991	0.489	0.003	-23415.020	0.012	0.439	0.465	0.514	0.001
156	2000	25	0.006	0.175	0.108	0.004	1.484	0.097	-0.001	0.061	0.018	276.621	0.483	0.003	-23351.905	0.017	0.433	0.448	0.518	0.001
78	2000	25	0.002	0.171	0.102	0.002	1.451	0.089	-0.002	0.061	0.016	275.571	0.501	0.003	-23463.769	0.025	0.451	0.452	0.551	0.003
20	2000	25	0.002	0.161	0.102	0.007	1.482	0.087	-0.003	0.061	0.015	287.456	0.325	0.004	-23384.685	0.049	0.275	0.228	0.423	0.011
<b># of replication</b>	<b>Sample.size</b>	<b># of item</b>	<b>me.a</b>	<b>se.a</b>	<b>rmse.a</b>	<b>me.b</b>	<b>se.b</b>	<b>rmse.b</b>	<b>me.theta</b>	<b>se.theta</b>	<b>rmse.theta</b>	<b>M2</b>	<b>M2.p</b>	<b>RMSEA.2</b>	<b>Log-Likelihood</b>	<b>Eq.5</b>	<b>empirical-nominal alpha</b>	<b>-2</b>	<b>2</b>	<b>accuracy of error rate estimate</b>
5000	1000	25	0.010	0.199	0.149	0.005	1.470	0.132	-0.001	0.146	0.025	275.917	0.489	0.004	-11688.028	0.003	0.439	0.483	0.495	0.000
2500	1000	25	0.010	0.199	0.149	0.004	1.468	0.132	0.000	0.148	0.025	276.2027	0.486	0.004	-11687.15335	0.004	0.436	0.477	0.494	0.000
1250	1000	25	0.012	0.198	0.149	0.004	1.463	0.132	0.000	0.145	0.025	274.7629	0.502	0.004	-11697.04832	0.006	0.452	0.490	0.515	0.000
625	1000	25	0.010	0.201	0.150	0.008	1.460	0.137	-0.002	0.145	0.026	273.7744	0.511	0.004	-11725.39484	0.009	0.461	0.494	0.529	0.000
312	1000	25	0.014	0.198	0.148	0.001	1.466	0.133	0.001	0.148	0.027	275.9167	0.490	0.004	-11702.90608	0.012	0.440	0.465	0.514	0.001
156	1000	25	0.009	0.201	0.147	0.010	1.467	0.133	-0.005	0.154	0.027	276.600	0.487	0.004	-11687.921	0.017	0.437	0.452	0.522	0.001
78	1000	25	0.016	0.198	0.148	-0.002	1.440	0.129	0.004	0.162	0.030	273.579	0.509	0.004	-11754.848	0.025	0.459	0.460	0.558	0.003
20	1000	25	0.007	0.207	0.150	-0.002	1.458	0.139	0.002	0.125	0.021	274.042	0.502	0.004	-11755.487	0.049	0.452	0.405	0.599	0.011
<b># of replication</b>	<b>Sample.size</b>	<b># of item</b>	<b>me.a</b>	<b>se.a</b>	<b>rmse.a</b>	<b>me.b</b>	<b>se.b</b>	<b>rmse.b</b>	<b>me.theta</b>	<b>se.theta</b>	<b>rmse.theta</b>	<b>M2</b>	<b>M2.p</b>	<b>RMSEA.2</b>	<b>Log-Likelihood</b>	<b>Eq.5</b>	<b>empirical-nominal alpha</b>	<b>-2</b>	<b>2</b>	<b>accuracy of error rate estimate</b>
5000	500	25	0.022	0.249	0.215	0.008	1.481	0.191	0.000	0.087	0.036	275.819	0.491	0.005	-5838.883	0.003	0.441	0.485	0.498	0.000
2500	500	25	0.022	0.248	0.215	0.009	1.480	0.191	-0.001	0.087	0.036	276.222	0.486	0.006	-5836.567	0.004	0.436	0.477	0.495	0.000
1250	500	25	0.022	0.249	0.216	0.010	1.484	0.191	-0.001	0.086	0.035	277.128	0.476	0.006	-5834.162	0.006	0.426	0.463	0.488	0.000
625	500	25	0.025	0.250	0.217	0.009	1.497	0.193	0.001	0.084	0.036	276.863	0.477	0.006	-5820.697	0.009	0.427	0.459	0.494	0.000
312	500	25	0.022	0.251	0.216	0.005	1.466	0.193	0.002	0.084	0.037	274.879	0.496	0.006	-5848.941	0.012	0.446	0.472	0.521	0.001
156	500	25	0.029	0.249	0.220	0.008	1.464	0.204	0.002	0.075	0.034	279.230	0.452	0.006	-5861.134	0.017	0.402	0.417	0.487	0.001
78	500	25	0.029	0.244	0.212	0.007	1.472	0.191	0.002	0.087	0.032	279.297	0.451	0.006	-5856.095	0.025	0.401	0.402	0.500	0.003
20	500	25	0.0177	0.256	0.214	-0.01	1.517	0.196	0.005	0.066	0.036	270.278	0.545	0.005	-5753.086	0.049	0.495	0.447	0.642	0.011

# of replication	Sample.size	# of item	me.a	se.a	rmse.a	me.b	se.b	rmse.b	me.theta	se.theta	rmse.theta	M2	M2.p	RMSEA.2	Log-Likelihood	Eq.5	empirical-nominal alpha	-2	2	accuracy of error rate estimate
4232	3000	50	0.003	0.163	0.082	0.003	1.482	0.075	-0.001	0.110	0.014	1176.523	0.49	0.002	-67642.223	0.003	0.440	0.483	0.497	0.000
2500	3000	50	0.003	0.163	0.081	0.004	1.485	0.074	-0.002	0.110	0.014	1176.306	0.491	0.002	-67603.514	0.004	0.441	0.482	0.500	0.000
1250	3000	50	0.003	0.162	0.082	0.004	1.485	0.074	-0.001	0.111	0.015	1177.651	0.485	0.002	-67678.284	0.006	0.435	0.473	0.497	0.000
625	3000	50	0.004	0.164	0.082	0.002	1.491	0.075	-0.001	0.112	0.014	1178.235	0.483	0.002	-67532.975	0.009	0.433	0.466	0.500	0.000
312	3000	50	0.003	0.163	0.082	0.002	1.476	0.076	-0.001	0.111	0.014	1172.113	0.515	0.001	-67731.533	0.012	0.465	0.490	0.540	0.001
156	3000	50	0.003	0.164	0.081	-0.001	1.482	0.075	0.002	0.109	0.015	1184.222	0.455	0.002	-67716.637	0.017	0.405	0.420	0.490	0.001
78	3000	50	0.000	0.166	0.083	-0.002	1.492	0.078	0.002	0.096	0.013	1172.677	0.513	0.002	-67576.411	0.025	0.463	0.464	0.562	0.003
20	3000	50	0.006	0.163	0.082	-0.015	1.479	0.085	0.008	0.144	0.020	1192.958	0.419	0.002	-67489.537	0.049	0.369	0.322	0.516	0.011
# of replication	Sample.size	# of item	me.a	se.a	rmse.a	me.b	se.b	rmse.b	me.theta	se.theta	rmse.theta	M2	M2.p	RMSEA.2	Log-Likelihood	Eq.5	empirical-nominal alpha	-2	2	accuracy of error rate estimate
5000	2000	50	0.005	0.172	0.100	0.005	1.483	0.091	-0.002	0.062	0.018	1177.890	0.483	0.002	-45105.882	0.003	0.433	0.477	0.489	0.000
2500	2000	50	0.005	0.172	0.100	0.005	1.481	0.091	-0.002	0.062	0.018	1178.677	0.478	0.002	-45114.507	0.004	0.428	0.469	0.487	0.000
1250	2000	50	0.005	0.172	0.100	0.004	1.482	0.092	-0.002	0.061	0.018	1180.331	0.467	0.002	-45107.194	0.006	0.417	0.454	0.479	0.000
625	2000	50	0.003	0.171	0.100	0.007	1.484	0.091	-0.003	0.067	0.018	1179.662	0.476	0.002	-45083.287	0.009	0.426	0.459	0.494	0.000
312	2000	50	0.006	0.172	0.102	0.010	1.499	0.094	-0.004	0.070	0.018	1173.844	0.509	0.002	-44964.306	0.012	0.459	0.484	0.533	0.001
156	2000	50	0.008	0.172	0.100	0.003	1.479	0.090	0.000	0.057	0.018	1185.649	0.428	0.002	-45051.152	0.017	0.378	0.393	0.463	0.001
78	2000	50	0.000	0.171	0.100	-0.005	1.455	0.093	0.003	0.053	0.020	1173.769	0.492	0.002	-45380.808	0.025	0.442	0.443	0.541	0.003
20	2000	50	0.004	0.165	0.097	0.006	1.472	0.085	-0.004	0.061	0.016	1171.665	0.531	0.001	-45367.346	0.049	0.481	0.433	0.628	0.011
# of replication	Sample.size	# of item	me.a	se.a	rmse.a	me.b	se.b	rmse.b	me.theta	se.theta	rmse.theta	M2	M2.p	RMSEA.2	Log-Likelihood	Eq.5	empirical-nominal alpha	-2	2	accuracy of error rate estimate
5000	1000	50	0.010	0.198	0.143	0.004	1.491	0.131	0.000	0.146	0.025	1178.332	0.481	0.003	-22521.259	0.003	0.431	0.475	0.487	0.000
2500	1000	50	0.009	0.198	0.144	0.005	1.491	0.130	-0.001	0.145	0.025	1180.073	0.470	0.003	-22523.647	0.004	0.420	0.461	0.479	0.000
1250	1000	50	0.010	0.197	0.143	0.007	1.488	0.131	-0.002	0.149	0.025	1178.503	0.480	0.003	-22511.257	0.006	0.430	0.468	0.492	0.000
625	1000	50	0.008	0.198	0.143	0.001	1.495	0.130	0.002	0.143	0.025	1179.117	0.474	0.003	-22520.229	0.009	0.424	0.457	0.491	0.000
312	1000	50	0.010	0.199	0.144	0.005	1.495	0.133	-0.001	0.145	0.026	1171.933	0.522	0.003	-22480.886	0.012	0.472	0.497	0.547	0.001
156	1000	50	0.007	0.196	0.141	0.011	1.486	0.127	-0.003	0.146	0.022	1173.195	0.512	0.002	-22520.435	0.017	0.462	0.477	0.547	0.001
78	1000	50	0.003	0.196	0.141	-0.003	1.499	0.131	0.002	0.143	0.027	1173.711	0.510	0.003	-22478.467	0.025	0.460	0.461	0.559	0.003
20	1000	50	0.011	0.196	0.146	-0.018	1.411	0.126	0.012	0.128	0.027	1194.590	0.379	0.004	-22841.479	0.049	0.329	0.282	0.476	0.011
# of replication	Sample.size	# of item	me.a	se.a	rmse.a	me.b	se.b	rmse.b	me.theta	se.theta	rmse.theta	M2	M2.p	RMSEA.2	Log-Likelihood	Eq.5	empirical-nominal alpha	-2	2	accuracy of error rate estimate
5000	500	50	0.020	0.244	0.207	0.009	1.509	0.190	-0.001	0.087	0.036	1180.094	0.471	0.004	-11231.536	0.003	0.421	0.465	0.478	0.000
2500	500	50	0.021	0.244	0.207	0.011	1.507	0.192	-0.002	0.089	0.037	1179.945	0.471	0.004	-11228.371	0.004	0.421	0.463	0.480	0.000
1250	500	50	0.018	0.243	0.206	0.011	1.507	0.192	-0.002	0.088	0.037	1179.010	0.478	0.004	-11242.831	0.006	0.428	0.466	0.490	0.000
625	500	50	0.020	0.244	0.207	0.003	1.505	0.190	0.003	0.076	0.035	1179.340	0.477	0.004	-11244.709	0.009	0.427	0.460	0.495	0.000
312	500	50	0.020	0.244	0.206	0.004	1.489	0.187	0.002	0.082	0.036	1172.149	0.516	0.004	-11243.590	0.012	0.466	0.492	0.541	0.001
156	500	50	0.024	0.245	0.208	0.004	1.510	0.191	0.001	0.080	0.037	1173.253	0.513	0.003	-11216.526	0.017	0.463	0.478	0.548	0.001
78	500	50	0.021	0.242	0.210	0.009	1.516	0.189	-0.001	0.086	0.033	1179.763	0.476	0.004	-11173.228	0.025	0.426	0.427	0.525	0.003
20	500	50	0.003	0.248	0.215	0.037	1.529	0.199	-0.017	0.122	0.033	1186.695	0.426	0.005	-11300.049	0.049	0.376	0.329	0.524	0.011



### 7. 3. Supplementary File for 3PL Model

#of replication	sample.size	#of item	me.a	se.a	rmse.a	me.b	se.b	rmse.b	me.c	se.c	rmse.c	me.theta	se.theta	rmse.theta	M2	M2.p	RMSEA.2	Log-Likelihood	Eq. 5	empirical-nominal alpha	-2	2	accuracy of error rate estimate
5000	3000	25	0.021	0.232	0.191	-0.004	1.463	0.203	-0.002	0.083	0.077	0.000	0.056	0.014	249.045	0.512	0.002	-38094.926	0.003	0.462	0.506	0.518	0.000
2500	3000	25	0.021	0.231	0.191	-0.004	1.459	0.202	-0.002	0.082	0.077	0.000	0.055	0.015	248.907	0.516	0.002	-38086.024	0.004	0.466	0.507	0.524	0.000
1250	3000	25	0.021	0.233	0.192	-0.004	1.470	0.204	-0.002	0.082	0.076	-0.001	0.058	0.015	250.148	0.495	0.002	-38103.522	0.006	0.445	0.482	0.507	0.000
625	3000	25	0.021	0.231	0.192	-0.005	1.458	0.204	-0.001	0.082	0.076	0.000	0.055	0.015	249.078	0.510	0.002	-38167.074	0.009	0.460	0.492	0.527	0.000
312	3000	25	0.022	0.234	0.191	-0.004	1.471	0.206	-0.001	0.083	0.077	0.000	0.057	0.015	249.633	0.505	0.002	-38195.932	0.012	0.455	0.480	0.530	0.001
156	3000	25	0.021	0.233	0.190	-0.002	1.440	0.203	-0.001	0.081	0.075	-0.001	0.063	0.016	249.855	0.494	0.002	-38276.803	0.017	0.444	0.459	0.529	0.001
78	3000	25	0.021	0.225	0.184	-0.014	1.413	0.200	0.001	0.081	0.076	0.003	0.047	0.014	253.372	0.455	0.003	-37804.962	0.025	0.405	0.406	0.504	0.003
20	3000	25	0.008	0.229	0.194	-0.002	1.470	0.228	-0.007	0.082	0.076	0.004	0.035	0.012	255.648	0.421	0.003	-38432.708	0.049	0.371	0.324	0.519	0.011
#of replication	sample.size	#of item	me.a	se.a	rmse.a	me.b	se.b	rmse.b	me.c	se.c	rmse.c	me.theta	se.theta	rmse.theta	M2	M2.p	RMSEA.2	Log-Likelihood	Eq. 5	empirical-nominal alpha	-2	2	accuracy of error rate estimate
5000	2000	25	0.034	0.271	0.240	-0.009	1.475	0.251	-0.001	0.093	0.089	0.000	0.123	0.018	248.774	0.515	0.003	-25390.889	0.003	0.465	0.509	0.521	0.000
2500	2000	25	0.034	0.270	0.239	-0.008	1.477	0.250	-0.002	0.093	0.089	0.000	0.124	0.018	248.317	0.521	0.002	-25313.220	0.004	0.471	0.512	0.530	0.000
1251	2000	25	0.035	0.270	0.238	-0.011	1.470	0.250	-0.001	0.092	0.088	-0.001	0.122	0.018	248.501	0.521	0.002	-25373.037	0.006	0.471	0.509	0.534	0.000
625	2000	25	0.033	0.269	0.238	-0.009	1.473	0.249	-0.001	0.093	0.089	-0.001	0.124	0.018	248.238	0.521	0.002	-25411.342	0.009	0.471	0.503	0.538	0.000
312	2000	25	0.035	0.273	0.244	-0.010	1.487	0.255	-0.002	0.096	0.092	0.000	0.122	0.017	250.301	0.493	0.003	-25113.030	0.012	0.443	0.469	0.518	0.001
156	2000	25	0.042	0.271	0.237	-0.018	1.491	0.249	0.000	0.093	0.088	0.000	0.138	0.019	248.732	0.522	0.003	-25393.005	0.017	0.472	0.487	0.557	0.001
78	2000	25	0.040	0.281	0.248	-0.019	1.500	0.249	0.001	0.097	0.092	-0.001	0.124	0.017	246.969	0.549	0.002	-25333.664	0.025	0.499	0.499	0.598	0.003
20	2000	25	0.039	0.276	0.244	-0.014	1.419	0.267	0.004	0.093	0.087	0.001	0.092	0.010	243.860	0.593	0.001	-25785.769	0.049	0.543	0.496	0.691	0.011
#of replication	sample.size	#of item	me.a	se.a	rmse.a	me.b	se.b	rmse.b	me.c	se.c	rmse.c	me.theta	se.theta	rmse.theta	M2	M2.p	RMSEA.2	Log-Likelihood	Eq. 5	empirical-nominal alpha	-2	2	accuracy of error rate estimate
5000	1000	25	0.074	0.374	0.363	-0.024	1.506	0.362	-0.001	0.113	0.111	0.000	0.146	0.025	249.099	0.512	0.004	-12675.636	0.003	0.462	0.505	0.518	0.000
2500	1000	25	0.077	0.383	0.371	-0.026	1.516	0.370	0.001	0.113	0.111	0.000	0.147	0.025	248.589	0.520	0.004	-12675.481	0.004	0.470	0.511	0.529	0.000
1250	1000	25	0.076	0.382	0.371	-0.025	1.509	0.368	0.000	0.114	0.112	0.001	0.148	0.025	248.443	0.522	0.004	-12674.330	0.006	0.472	0.510	0.535	0.000
625	1000	25	0.073	0.372	0.359	-0.024	1.501	0.362	-0.001	0.113	0.111	0.002	0.146	0.026	248.465	0.519	0.004	-12671.896	0.009	0.469	0.501	0.536	0.000
312	1000	25	0.072	0.381	0.369	-0.023	1.533	0.365	-0.001	0.116	0.114	0.001	0.142	0.024	247.811	0.528	0.003	-12607.724	0.012	0.478	0.503	0.552	0.001
156	1000	25	0.073	0.388	0.375	-0.014	1.523	0.374	0.001	0.112	0.110	-0.001	0.147	0.025	248.867	0.507	0.004	-12735.953	0.017	0.457	0.472	0.541	0.001
78	1000	25	0.071	0.374	0.361	-0.029	1.469	0.347	0.000	0.112	0.110	0.000	0.153	0.028	250.695	0.498	0.004	-12605.100	0.025	0.448	0.449	0.548	0.003
20	1000	25	0.073	0.370	0.352	-0.031	1.469	0.390	-0.001	0.110	0.110	-0.006	0.180	0.030	243.709	0.575	0.002	-12792.127	0.049	0.525	0.477	0.672	0.011
#of replication	sample.size	#of item	me.a	se.a	rmse.a	me.b	se.b	rmse.b	me.c	se.c	rmse.c	me.theta	se.theta	rmse.theta	M2	M2.p	RMSEA.2	Log-Likelihood	Eq. 5	empirical-nominal alpha	-2	2	accuracy of error rate estimate
5000	500	25	0.179	0.631	0.649	-0.050	1.621	0.588	0.002	0.141	0.141	-0.001	0.088	0.035	247.992	0.526	0.005	-6309.552	0.003	0.476	0.520	0.532	0.000
2500	500	25	0.182	0.635	0.652	-0.052	1.622	0.589	0.002	0.140	0.140	0.000	0.085	0.035	248.248	0.520	0.005	-6309.280	0.004	0.470	0.511	0.528	0.000
1250	500	25	0.175	0.630	0.645	-0.053	1.634	0.601	0.002	0.141	0.141	0.000	0.087	0.037	246.302	0.541	0.005	-6300.432	0.006	0.491	0.528	0.553	0.000
625	500	25	0.173	0.614	0.633	-0.050	1.597	0.578	0.002	0.140	0.140	-0.001	0.086	0.035	249.147	0.508	0.005	-6318.462	0.009	0.458	0.490	0.525	0.000
312	500	25	0.193	0.626	0.647	-0.065	1.617	0.596	0.002	0.140	0.141	-0.002	0.088	0.036	249.015	0.513	0.005	-6336.257	0.012	0.463	0.488	0.537	0.001
156	500	25	0.195	0.646	0.670	-0.032	1.594	0.580	0.003	0.145	0.145	-0.004	0.096	0.038	246.692	0.542	0.005	-6249.925	0.017	0.492	0.507	0.577	0.001
78	500	25	0.199	0.694	0.714	-0.055	1.663	0.602	0.003	0.141	0.141	0.004	0.080	0.040	241.654	0.592	0.004	-6294.088	0.025	0.542	0.543	0.642	0.003
20	500	25	0.201	0.703	0.735	-0.030	1.731	0.628	-0.001	0.144	0.141	0.012	0.060	0.031	250.054	0.501	0.006	-6342.525	0.049	0.451	0.403	0.598	0.011

# of replication	sample.size	# of item	me.a	se.a	rmse.a	me.b	se.b	rmse.b	me.c	se.c	rmse.c	me.theta	se.theta	rmse.theta	M2	M2.p	RMSEA.2	Log-Likelihood	Eq. 5	empirical-nominal alpha	-2	2	accuracy of error rate estimate
5000	3000	50	0.016	0.217	0.168	-0.003	1.489	0.185	-0.002	0.078	0.071	-0.001	0.056	0.014	1122.474	0.517	0.001	-74079.570	0.003	0.467	0.510	0.523	0.000
2500	3000	50	0.016	0.218	0.168	-0.002	1.487	0.185	-0.002	0.078	0.071	0.000	0.054	0.015	1122.125	0.519	0.001	-74090.428	0.004	0.469	0.510	0.527	0.000
1250	3000	50	0.018	0.217	0.168	-0.004	1.489	0.185	-0.001	0.077	0.070	0.000	0.056	0.015	1122.678	0.517	0.001	-74100.260	0.006	0.467	0.504	0.529	0.000
625	3000	50	0.021	0.217	0.169	-0.004	1.488	0.183	-0.001	0.078	0.071	-0.001	0.058	0.015	1120.749	0.527	0.001	-73851.830	0.009	0.477	0.510	0.544	0.000
312	3000	50	0.015	0.219	0.169	-0.001	1.490	0.184	-0.002	0.079	0.071	0.000	0.053	0.014	1126.778	0.491	0.002	-74093.610	0.012	0.441	0.466	0.515	0.001
156	3000	50	0.014	0.218	0.169	-0.004	1.495	0.185	-0.002	0.081	0.074	-0.001	0.059	0.014	1120.568	0.530	0.001	-73702.999	0.017	0.480	0.495	0.565	0.001
78	3000	50	0.010	0.216	0.166	-0.004	1.466	0.184	-0.002	0.079	0.071	0.001	0.053	0.014	1125.253	0.498	0.001	-74269.253	0.025	0.448	0.448	0.547	0.003
20	3000	50	0.019	0.211	0.163	-0.005	1.450	0.172	-0.004	0.078	0.072	0.003	0.047	0.015	1134.891	0.445	0.002	-73364.730	0.049	0.395	0.348	0.543	0.011
5000	2000	50	0.029	0.250	0.211	-0.010	1.496	0.230	-0.001	0.088	0.082	0.000	0.124	0.018	1122.660	0.514	0.002	-49391.158	0.003	0.464	0.508	0.520	0.000
2500	2000	50	0.028	0.250	0.211	-0.009	1.498	0.230	-0.001	0.088	0.082	0.000	0.122	0.018	1123.245	0.510	0.002	-49394.856	0.004	0.460	0.501	0.519	0.000
1250	2000	50	0.029	0.250	0.212	-0.011	1.490	0.229	0.000	0.088	0.082	0.000	0.124	0.018	1122.263	0.518	0.002	-49411.865	0.006	0.468	0.505	0.530	0.000
625	2000	50	0.027	0.249	0.210	-0.008	1.490	0.230	0.000	0.088	0.082	-0.001	0.126	0.018	1124.352	0.500	0.002	-49411.525	0.009	0.450	0.483	0.518	0.000
312	2000	50	0.029	0.250	0.212	-0.011	1.492	0.230	0.000	0.089	0.083	-0.002	0.123	0.017	1120.916	0.533	0.002	-49393.274	0.012	0.483	0.508	0.558	0.001
156	2000	50	0.026	0.248	0.209	-0.012	1.516	0.230	-0.001	0.089	0.083	0.002	0.114	0.017	1123.579	0.511	0.002	-49213.810	0.017	0.461	0.476	0.546	0.001
78	2000	50	0.031	0.251	0.216	-0.009	1.498	0.229	-0.002	0.088	0.083	0.001	0.126	0.015	1121.690	0.499	0.002	-49426.222	0.025	0.449	0.450	0.549	0.003
20	2000	50	0.039	0.253	0.215	-0.025	1.540	0.217	0.003	0.090	0.083	0.000	0.121	0.016	1115.446	0.564	0.002	-49194.058	0.049	0.514	0.467	0.662	0.011
5000	1000	50	0.062	0.335	0.315	-0.022	1.517	0.325	0.001	0.107	0.104	0.000	0.073	0.025	1121.134	0.523	0.002	-24619.555	0.003	0.473	0.517	0.530	0.000
2500	1000	50	0.063	0.336	0.316	-0.019	1.523	0.328	0.001	0.107	0.103	-0.002	0.077	0.025	1119.738	0.530	0.002	-24626.171	0.004	0.480	0.521	0.539	0.000
1251	1000	50	0.061	0.338	0.317	-0.019	1.521	0.328	0.001	0.107	0.103	-0.001	0.147	0.025	1122.340	0.515	0.003	-24637.558	0.006	0.465	0.502	0.527	0.000
625	1000	50	0.064	0.339	0.318	-0.025	1.524	0.329	0.001	0.107	0.103	0.000	0.148	0.025	1119.339	0.531	0.002	-24603.283	0.009	0.481	0.513	0.548	0.000
312	1000	50	0.068	0.338	0.317	-0.029	1.535	0.334	0.001	0.107	0.103	0.000	0.149	0.026	1117.253	0.543	0.002	-24645.979	0.012	0.493	0.518	0.567	0.001
156	1000	50	0.066	0.330	0.309	-0.017	1.520	0.321	0.001	0.107	0.103	-0.002	0.147	0.026	1121.656	0.525	0.002	-24488.165	0.017	0.475	0.490	0.560	0.001
78	1000	50	0.059	0.329	0.310	-0.025	1.489	0.325	0.002	0.105	0.100	0.004	0.136	0.026	1121.085	0.518	0.002	-24768.173	0.025	0.468	0.469	0.568	0.003
20	1000	50	0.064	0.339	0.319	-0.027	1.506	0.340	-0.004	0.104	0.100	0.006	0.146	0.029	1119.694	0.531	0.002	-24831.287	0.049	0.481	0.434	0.629	0.011
5000	500	50	0.146	0.538	0.544	-0.047	1.602	0.519	0.004	0.133	0.131	-0.001	0.088	0.037	1119.046	0.531	0.003	-12272.765	0.003	0.481	0.524	0.537	0.000
2500	500	50	0.143	0.541	0.546	-0.047	1.598	0.521	0.004	0.133	0.131	-0.001	0.086	0.036	1119.896	0.526	0.003	-12278.759	0.004	0.476	0.517	0.535	0.000
1250	500	50	0.143	0.540	0.546	-0.045	1.612	0.528	0.003	0.132	0.130	-0.001	0.086	0.035	1122.462	0.519	0.004	-12271.311	0.006	0.469	0.507	0.532	0.000
625	500	50	0.141	0.538	0.543	-0.051	1.591	0.522	0.003	0.133	0.131	0.001	0.078	0.033	1121.053	0.521	0.003	-12277.678	0.009	0.471	0.503	0.538	0.000
312	500	50	0.139	0.523	0.525	-0.050	1.594	0.512	0.004	0.133	0.131	0.000	0.081	0.036	1118.110	0.539	0.003	-12309.872	0.012	0.489	0.515	0.564	0.001
156	500	50	0.146	0.567	0.571	-0.050	1.594	0.518	0.004	0.133	0.131	0.005	0.072	0.038	1125.730	0.501	0.004	-12288.918	0.017	0.451	0.466	0.536	0.001
78	500	50	0.139	0.524	0.528	-0.052	1.616	0.539	0.003	0.131	0.130	-0.003	0.093	0.039	1117.142	0.552	0.003	-12288.143	0.025	0.502	0.502	0.601	0.003
20	500	50	0.168	0.583	0.592	-0.046	1.668	0.553	0.001	0.134	0.134	0.007	0.071	0.038	1125.758	0.467	0.004	-12256.452	0.049	0.417	0.369	0.564	0.011



## Gathering evidence on e-rubrics: Perspectives and many facet Rasch analysis of rating behavior

Inan Deniz Erguvan <sup>1,\*</sup>, Beyza Aksu Dunya <sup>2,3</sup>

<sup>1</sup>Gulf University for Science and Technology, Block 5, Building 1, Mubarak Al-Abdullah Area, West Mishref, Kuwait

<sup>2</sup>Bartın University, Bartın, Turkey

<sup>3</sup>University of Illinois at Chicago, Chicago, IL, USA

### ARTICLE HISTORY

Received: Oct. 29, 2020

Revised: Mar. 09, 2021

Accepted: Apr. 12, 2021

### Keywords:

Freshman composition,  
Electronic rubric,  
Many Facet Rasch Model,  
Rater behavior,  
Leniency, severity.

**Abstract:** This study examined the faculty perspectives towards the use of electronic rubrics and their rating behavior in a freshman composition course. A mixed-methods approach has been employed for data collection and analysis. The data for faculty perspectives were collected from nine instructors through semi-structured interviews and for their behavior, six instructors teaching the same course in Fall 2019, shared their students' essay scores with the researchers. Many facet Rasch model (MFRM) was employed for quantitative data analysis. According to the findings of the quantitative data, the instructors differed in their degree of leniency and severity, one instructor being more lenient and one being more severe than the others. Another interesting finding was one instructor turned out to be an inconsistent user of the e-rubric. The findings of the qualitative data showed that writing faculty think e-rubrics come with great advantages such as facilitating scoring, ensuring standardization, and reducing student complaints and grade appeals. However, they view the impact of e-rubrics on student writing with cautious optimism. The findings of the qualitative and quantitative strands are overlapping, and the responses elicited from the participants seem to shed some light on the rating behavior of the writing faculty.

## 1. INTRODUCTION

The word “rubric” implies an assessment tool that describes levels of performance on a particular task and is used to assess outcomes in a variety of performance-based contexts (Hafner & Hafner, 2003). Rubrics, by this definition differ from rating scales, which have criteria but no performance level descriptions, although these may be confused with “rubrics” (Brookhart 2013). Rubrics, checklists, and rating scales all have criteria but what distinguishes them is the scale. Other than rubrics, none of the other scales offer students a description of the quality of their performance they can easily use to envision their next steps in learning (Brookhart, 2018).

---

\*CONTACT: Inan Deniz ERGUVAN ✉ [erguvan.d@gust.edu.kw](mailto:erguvan.d@gust.edu.kw) 📧 Gulf University for Science and Technology, Block 5, Building 1, Mubarak Al-Abdullah Area, West Mishref, Kuwait

There are many benefits of using rubrics; providing consistency of scoring across students, assignments, and different raters is one of the major benefits of using rubrics. Designing and using rubrics to grade assignments or tests can indeed reduce inconsistencies and make grading of the written work more objective. Subjectivity is a big concern in assessing writing, and the use of rubrics can help remove bias from grading (Allen & Tanner, 2006). Rubrics indeed provide an opportunity for reliable scoring, rather than a subjective scoring simply based on the rater's personal idiosyncrasies (Carr, 2000).

Sharing the rubric with students can have the added benefit of enhancing learning by allowing for feedback and self-assessment (Jonsson & Svingby, 2007; Reddy & Andrade, 2010). The rubric tells both teachers and students what fundamental skills teachers look for while they are assessing student performance (Arter & McTighe, 2001) because they incorporate criteria to rate the essential dimensions of performance, as well as standards of achieving those criteria (Jonsson & Svingby, 2007).

Another advantage of using rubrics is facilitating self- and peer-evaluation, both of which could be valuable avenues for providing meaningful feedback. With the development of a simple rubric, students can peer-review each other's work, thus see other examples other than their own and that of the instructor. Perhaps most importantly, sharing rubrics with students can support them in identifying where their thinking has gone wrong and promote learning (Jonsson & Svingby, 2007; Reddy & Andrade, 2010). As Broad (2000) suggested, when rubrics are utilized accurately, learners not only get feedback from the instructor, but they also receive training in self-assessment. Self-evaluation of one's work using the instructor's rubric can build meta-cognitive skills in making self-corrections. Such evaluations may provide meaningful feedback that could further enhance the learning process (Sadler & Good, 2006; Freeman & Parks, 2010). However, Brookhart and Chen (2015) underline the fact that rubrics that include descriptions of quality on criteria that reflect learning goals, rather than rubrics that focus on the requirements for an assignment and not indications of learning function as the goals toward which students can monitor their progress,

While the attitude towards using rubrics is prevalently positive, there are some negative perspectives, too. Critics complain that rubrics are rigid and even when they are modified to allow for more commentary on student strengths and weaknesses; they do a disservice to students' ability to learn. Critics also add that rubrics result in standardized measurement of standardized writing, which is not the purpose of writing instruction (Nordrum et al. 2013; Torrance, 2007).

Andrade (2000) brings up in a study that rubrics are not necessarily self-explanatory and not all students are acquainted with rubrics. Therefore, teachers must not assume that the criteria in the rubric are all clear to students. Andrade (2005) also alerts that rubrics must pass a test of quality, demonstrating that if another instructor utilizes the same rubric to review the same paper; their results should only have insignificant differences.

Some critics are concerned about the impact of using rubrics on creativity of students. Kohn (2006) argues that rubrics result in student writing with less depth of thought, therefore rubrics should not drive instruction, nor should they be shared with students. Kohn also says an excessive amount of consideration regarding the nature of work causes the students to lose enthusiasm for whatever they are doing. Another critic of rubrics, Wilson (2007), suggests that over-reliance on rubrics may result in learners stopping writing for a live audience, and beginning to write for a rubric. Wilson (2007) argues that the rubrics provide students with non-specific input and do not have much relevance with what they need to say and adds that writing offers its own set of criteria and that each piece should be examined individually.

To mitigate most of these issues, some studies emphasize the importance of involving students in developing rubrics and reduce the number of criteria incorporated so that they become easier to comprehend and apply as a learning tool (García-Ros, 2011).

### **1.1. E-rubrics**

With the rise of the digital age, Information and Communications Technologies (ICTs) have started offering new roles and resources to teachers as well as students to improve teaching and learning processes. As technology started being widely used in assessment techniques, standard rubrics have slowly been replaced by their digital companions, called electronic rubrics (e-rubrics) (Raddawi & Bilikozen, 2018). Simply put, electronic rubrics or e-rubrics are rubrics that are presented online.

E-rubrics carry out the same functionality as paper or print-out rubrics, but there is an added value of e-rubrics. According to Steffens (2014), students are increasingly working in technology enhanced learning environments. From a technical point of view, it is relatively easy to integrate e-rubrics in technology enhanced learning environments. Using e-rubrics has the advantage that feedback can be given much more quickly than in traditional learning environments with paper-and-pencil. Just like any computer-assisted system, e-rubrics make grading and assessing much simpler for instructors as they reduce the time required to grade assignments. E-rubrics facilitate more immediacy in the teacher - student communication, and frequent and quicker feedback may help students to better self-regulate their learning than in traditional learning environments (Rivasy et al., 2014), and to be active participants in the learning process.

Kirwin and DiVall (2015) express that e-rubrics offer different advantages to various groups such as students, course instructors and administrators. Students can use the feedback and comments within a rubric as well as the scores on particular items to see their strengths and weaknesses. Course instructors can use e-rubrics to see the dimensions of performance and aggregate across multiple assignments to examine learning outcome data. Program administrators can benefit from e-rubrics by aggregating student performance as an indicator of the group's competency in a particular area. With the help of learning management systems (LMSs) that are capable of aligning course outcomes with particular dimensions in e-rubrics, it will be quite practical to evaluate learning outcomes for a large number of students.

Another benefit of e-rubrics was put forward by Martinez et al. (2016). In this case study, the course professor and students generated a collaborative methodology to build a rubric with the support of educational technologies. As a result of the collaborative effort, the students and professor agreed on the criteria for assessment of student presentations. This effort eventually showed that the evaluations given by students and the course professor got closer thanks to the increased e-rubric use. E-rubrics have the advantage of facilitating collaborative rubric generation among course professors and students, which could reverse the drawback of rubrics as argued by some critics that rubrics are not always clear to students.

However, one must acknowledge the fact that writing assessment is a complex and error-prone cognitive process. Therefore, attention should also be turned to raters themselves because in the end what is central to writing performance assessment is the rater behavior. Researchers have long recognized that rater judgments have an element of subjectivity. It is inevitable that the act of rating involves rater errors or rater biases (Myford & Wolfe, 2003), and although raters are trained to use and interpret rating scales in similar ways, rater effects also need to be studied. Rater behavior must be taken into consideration in order to assess the construct in question. Among many potential rater errors, four major categories of rater errors have been given emphasis: (a) severity or leniency, (b) inconsistency (c) halo, and (d) restriction of range (Myford & Wolfe, 2003; Saal et al., 1980).

In this study, we focused on two common rater errors: Severity/leniency and inconsistency. The former is defined as the tendency of a rater to assign higher or lower ratings on average than those assigned by other raters, and it is commonly considered to be the most pervasive and detrimental rater effect in performance assessments (Dobria, 2011). Various factors contribute to a rater's severity or leniency including professional experience, and in some circumstances, the most experienced or senior rater may also be the most severe (Eckes, 2011).

Rater inconsistency is a rater's tendency to apply the rating scale inconsistently compared to the way other raters apply the same scale. The presence of rater inconsistency indicates the rater's lack of understanding of rating criteria, making the interpretation of ratings less meaningful. A rater who rates inconsistently increases the randomness in scores by assigning high ratings to those who deserve low ratings and low ratings to those who deserve high ratings. This error reduces the ability of the scores to reliably differentiate between competent and incompetent students (Iramaneerat & Yudkowsky, 2006).

Thus, the aim of this paper is double fold: to analyze the perspective of writing faculty towards using e-rubrics through interviews and to examine their rating behavior via Many Facet Rasch Model (MFRM) in a first-year composition course in a university in Kuwait. MFRM is a member of Rasch Measurement Models that is suitable to simultaneously analyze multiple facets potentially having an impact on scores (Eckes, 2011). It is an extension of the basic Rasch Model for analyzing dichotomous data and used in assessments that involve human judgment. It allows researchers to investigate potential sources of error that cause construct irrelevant variance into the ratings. The advantages of MFRM also includes that each facet's unique contribution to the scores can be partitioned out and investigated independently of other facets in the assessment (Myford & Wolfe, 2003).

The use of LMS integrated online rubrics in the institution dates to 2015, when the practice of using a common analytic rubric was adopted. A common rubric is used in all writing courses which is developed and revised by the course coordinator every year. Writing faculty upload this common rubric in Turnitin in order to check the plagiarism similarity index of the essays and also to mark and give feedback to their students' essays. However, standardization workshops conducted during the academic year indicated some discrepancy among instructors, thus, the researchers decided to conduct a study during the Fall semester of 2019-2020 to analyze whether the severity / leniency is a real problem in the department. The qualitative component was added to analyze the instructors' perspective towards the use of the e-rubric.

To this end, this paper tried to find answers to the following research questions:

1. What are some benefits of using e-rubrics in an academic writing class?
2. What are some limitations of using e-rubrics in an academic writing class?
3. How do e-rubrics compare to rubrics in assessing writing?
4. Does using e-rubrics affect students' writing performance positively?
5. Do the instructors differ in terms of their level of severity while rating the student essays with the standard e-rubric? If yes, which rater is more severe / lenient than others?
6. How consistently are the instructors able to distinguish among the students in terms of their levels of proficiency?

## 2. METHOD

This study used a mixed method research design for data collection and analysis. In the mixed method research design, both qualitative and quantitative methods are used and "mixed" for collecting and analyzing data in a single study (Creswell, 2012). In mixed method research design, the two forms of data are mixed concurrently or sequentially by giving priority to one or both forms of data (Creswell & Clark, 2011). Using multiple methods helps to provide a

more comprehensive framework of the phenomenon under investigation by enabling rich and informative data and also to validate and triangulate the data by analyzing the same issue through both quantitative and qualitative methods (Silverman, 2000).

For the qualitative study, the researchers conducted semi-structured interviews with the writing faculty. The interview questions were developed to evaluate the raters' perspective towards e-rubrics. The researchers adapted Raddawi and Bilikozen's (2018) interview questions to evaluate ELT professors' perspectives on the use of rubrics in an academic writing class in a university in the United Arab Emirates.

### **2.1. Research Population**

The data used for the qualitative strand of this study came from the interviews conducted with the writing faculty working in the university in January to March 2020. Nine instructors teaching various writing courses in the Spring 2020 academic year were interviewed for their perceptions of using e-rubrics in assessing their students' essays. Four of the faculty members were native speakers of English, whereas five of them were non-native. Out of nine, two were female, and seven were male.

The data used for the quantitative strand of the study came from six instructors teaching a particular writing course in the Fall 2019, in the same university in Kuwait. Out of six instructors, four of them were male, two were female; three were native speakers, three were nonnative. The total number of students taking this particular writing course is 442, and the number of essays that were rated by the six instructors is 424.

### **2.2. Data Collection**

The qualitative strand of the study involved interviewing the writing faculty members in the department. According to Brown (2001), interviews have a high return rate and fewer incomplete answers. They also allow researchers to ask for clarification in a participant's response to a given question (Mertler, 2009) As a result, interviews offer an advantage over surveys as researchers can get more details on vague answers.

To protect the anonymity and confidentiality of the participants, each was given a code. As for the reliability and validity of the interviews, the researcher conducted some pilot interviews with some faculty who are not in the sample to check the understandability of the questions. Due to the interactive nature of the interview and the various biases and limits that may impact human decision-making, during the interviews the interviewer did not deviate from the interview questions and kept a neutral body language with all interviewees. After the interview, the recorded voice file and the written interview text were sent to each interviewee to obtain their approval to avoid any misunderstandings.

For the quantitative strand, six instructors scored the final draft of their students' research-based essays using a common analytic e-rubric. The common e-rubric is uploaded on Turnitin and attached to the writing task; therefore, scoring takes place electronically (See [Appendix 1](#) for the e-rubric). Student essays were rated by the instructors during the first week of December 2019. No special training or a norming session was provided prior to or during the rating process. The instructors then shared their students' scores with the researcher. Also, for anchoring purposes, the researcher randomly selected two essays from the pool of essays to be rated by the six writing faculty members. The common frame of reference made it possible to compare all students and all instructors on the same scale.

### **2.3. Data Analysis**

The data analysis process that could be utilized in qualitative research can be broken down in three steps: preparing and organizing the data for analysis, reducing the data into themes through coding and condensing the codes, and finally representing the data in figures, tables,



or discussion (Creswell, 2007). In this study all interviews were audio-recorded and fully transcribed in order to prepare them for the analysis. Following Radnor's (2002) approach to analyzing semi-structured interviews in interpretive research, the data was further prepared for analysis by reading the transcribed interviews several times and noting down the topics emerging from the data. During the data analysis process, we read the transcripts carefully to draw out any implicit topics that we may have missed. We made a list of the topics and gave a code to each topic. Afterwards, we pulled out the categories within each topic and listed these categories under each topic as subheadings. We also counted the frequency of these categories in interview texts to indicate which categories are more commonly expressed by the interviewees. The percentage of agreement between the coders which represents the share of common number of codes with respect to the total number of codes was calculated as a measure of consistency for coding. As a rule of thumb, 80% agreement between coders is sufficient for ensuring intercoder reliability (Miles & Huberman, 1994). The percentage of agreement between coders was found 92% for this study. This value was well above the suggested value of 80%. The next step was reading the transcripts for content, which meant going through the text one more time to highlight and code the main quotes that go under each category. These quotes were used to illustrate the participants' voices and viewpoints more clearly in the discussion of the findings.

For the quantitative strand, we employed the MFRM (Linacre, 1989) to analyze the rater behavior by using FACETS program (Linacre, 2020). A facet is an aspect of any assessment situation that may have an influence on the outcome. A facet can be raters, performance tasks, or examinee-related characteristics such as race, gender, etc. (Myford & Dobria, 2012). In FACETS output, a column titled "measure" displays each instructor's severity measures in log-odd units. These measures are estimates of each rater's true location on the severity dimension (Eckes, 2011). MFRM can separate out each facet's unique contribution to the assessment setting and examine it independently of other facets to determine to what extent each facet is functioning. The advantage of MFRM with respect to classical approach while examining rating data is that MFRM allows an in-depth analysis of patterns in ratings even when a different set of examinees are concerned. In the classical approach, interrater reliability is reported while analyzing rating data. Interrater reliability is an informative statistic, yet it is limited in detecting possible rater effects such as severity/leniency. MFRM provides a valid account of potential irrelevant variance sources in ratings such as severity/leniency or bias. There are multiple indicators for detecting rater effects under MFRM framework which includes outfit and infit mean-square indices (Myford & Wolfe, 2003). Infit and outfit indices are used to assess randomness in the scores assigned by raters. These values are averages of squared standardized residuals and have an expected value of 1.00. Specifically, mean-square outfit which is more sensitive to outliers in the data is the non-weighted mean of the squared standardized residuals while infit is the information-weighted mean of the standardized residuals (Wolfe, 2009). In both statistics, values greater than 2.00 are accepted as indication of severe misfit that distort the measurement (Linacre, 2009). FACETS program yields individual fit values for each rater to assess rater misfit. In this part of the analysis, raters who had fit indices greater than 2.00 were flagged for further review.

### **3. RESULTS**

#### **3.1. Qualitative Study**

Nine faculty members who were interviewed come from various backgrounds and nationalities, as displayed in [Table 1](#). The table displays that although the writing faculty in the university are quite varied in their nationalities, they have similar backgrounds in education and experience in teaching. Their experience in teaching is also reflected in their experience with using rubrics, with eight of them having more than eight years of experience using rubrics in



assessing student essays. Understandably, the experience with e-rubrics is at an average of three to four years, simply because e-rubrics are new tools in assessing writing and they have been recently adopted in writing courses with the accelerated integration of internet technologies in education.

**Table 1.** *Demographics of interviewed faculty members.*

Interviewee	Nationality	Qualification	Teaching Experience	Experience with rubrics	Experience with e-rubrics
P1	American	MA in English Literature	35 years	20 years	14 years
P2	Kuwaiti	PhD in Linguistics	4 years	8 months	8 months
P3	Kuwaiti	MA in English	10 years	10 years	3 years
P4	Indian	MA in English Literature	25 years	20 years	15 years
P5	British	MA in ELT	25 years	10 years	4 years
P6	Egyptian	PhD in Composition & Rhetorics	10 years	10 years	4 years
P7	Bosnian	PhD in TESOL	21 years	10 years	6 months
P8	American	MA in TESOL	8 years	8 years	3 years
P9	New Zealander	PhD in TESOL	24 years	20 years	3 years

When asked what they used for grading and giving feedback before they started using rubrics, two faculty members said they do not remember any time when they did not use rubrics for grading. The remaining three faculty members said they graded holistically with plenty of qualitative feedback and four of them said they used a previously agreed upon checklist, scheme and some standards based on learning objectives of the course. Basically, even the instructors who used to grade holistically always referred to some standards, attended some standardization sessions or had a checklist to refer to, which means most of the writing faculty had exposure to a scorecard for grading before they fully adopted rubrics in their classes.

### 3.1.1. Research question 1

When the faculty members were asked what they see as the greatest benefit of using e-rubrics in assessing student essays, they gave a variety of responses which could be summarized in [Table 2](#).

**Table 2.** *Perceived benefits of e-rubrics.*

Perceived benefit	Frequency
Making grading objective and transparent	6
Reducing student complaints / grade appeals	4
Facilitating grading	4
Helping standardization among raters	3

[Table 2](#) shows that the most frequently mentioned benefit of using rubric is making grading transparent to students as they clearly show how their essays are graded and where their weaknesses and strengths are. Another benefit mentioned quite frequently, by four instructors, is “rubrics reduce complaints”. In fact, these two perceived benefits are quite closely connected to each other, because instructors feel less defensive when they can explain clearly where the students’ scores come from and they are able to “justify” the scores.

*P6. I would say the biggest advantage is that they break down the grade in a very quantifiable easy to explain manner so the student would see exactly how they received that grade. They*

would have a clear idea about the criteria on which they were graded. They would know exactly what their strengths and weaknesses on each criterion are or were.

Some instructors see an added benefit to making grading transparent in reducing the complaints and negotiations:

*P5. I make strenuous efforts to get the students to review the e-rubrics... as whole class activities before major essays so they are constantly aware of the requirements. That being the case, no student can make any excuses regarding being under graded or being sort of penalized in some way because they are all aware of what the requirements are, and they know that if they don't meet these requirements they have no complaints. Students have a clear template of how they got their grades. It clarifies everything for everybody. It reduces conflicts. It reduces grade appeals because, theoretically, there is nothing subjective about a rubric. This creates an objective way of grading and it roughly solves all issues and answers all questions.*

It seems student complaints due to scores are a problem in the institution and rubrics are useful tools to offer relief to instructors in this regard.

*P9. I show the whole class how I am grading... I put the rubrics here and I point out this is why he is getting 80 here for the grammar, for example, because he has made these grammar mistakes or the sentences are not complex enough, so I have chosen 80 on this rubric but they don't really read the actual all the dots on the rubric. They just want the general grade. When I grade three to five sentences that had grammar errors or such and such, they don't care about that, but they get the general idea about the grading.*

Rubrics helping standardization among raters has also been mentioned by 3 instructors. For the interviewee 7, quoted below, that was second major advantage of rubrics, after making grading transparent for students:

*P7. The biggest advantage is that ... It has actually two advantages. First, they are used to the system. Once you do this once, they get used to how they're being graded. Second, it is standardized across the university so, you know, not this professor is better than this professor because they do not grade in the same way.*

Facilitation of grading was mentioned by three instructors, and they mentioned the benefit of justification along with the ease of grading.

*P3. I am thinking about if you have large number of students in class who are trying to take the same assessment you have 60 to 90 papers to get through, it does facilitate the grading process and also facilitates the feedback process because going back to every single paper I would forget how and why I graded this paper and gave it that specific grade but going back and looking at the rubric itself and you kind of have that sliding option on Turnitin, for example, it does make easy for me to provide feedback and justify the grade as well as, when it comes to the students, it helps the students sync in, I guess.*

### **3.1.2. Research question 2**

When the faculty members were asked what they see as limitations of using e-rubrics in assessing student writing, they gave a variety of responses categorized under three themes, which is shown in [Table 3](#). As the table displays, the perceived limitations are mainly related to scoring problems and the problems the students have while using the e-rubric, as well as time and effort the construction of a successful rubric requires. Besides, two participants expressed they cannot think of any limitations related to e-rubrics, and this was also displayed in the table as a separate theme.

**Table 3.** Perceived limitations of e-rubrics.

Perceived limitation	Frequency
Scoring issues (ambiguity, rigidity)	4
Student issues (accessing / reading and understanding the e-rubric)	4
Construction of the rubric	2
No limitation	2

The perceived limitation “scoring issues” is composed of a couple issues as defined by instructors. Some have pointed out the limitations of rubrics in differentiating certain categories from one another and some have brought up the lack of flexibility when it comes to issues such as giving half points in the rigid scoring scheme of the rubric. However, these points are not necessarily related to the nature of e-rubrics in particular, but rubrics in general.

P6. *I would say the disadvantage would be that it would be very difficult to design a rubric that would cover all the possible mistakes and any potential drawbacks on any given criteria. There is no category or there is no description on the rubric so I would say that the challenge is that there is always something missing in the rubric because no matter how detailed and descriptive you are, you are always not going to be able to cover everything on a one-page rubric ...so the categories are not always clear cut in reality as they are presented on a rubric ... the criteria may fall between two categories ...*

P3. *Certain rubrics especially ones that are weighted in a specific way it would be hard to differentiate, for example, between grammar mechanics from content so if the content is preferred but the grammar is being marked down then grammar shouldn't be as important but if the grammar hinders the actual sentence and the structure and makes it completely incomprehensible then I, for example, run the risk of marking them either too harshly or too leniently.*

Three instructors mentioned the problems their students go through while accessing the e-rubrics, and the fact that they never even try to access the e-rubrics. This instructor quoted below describes the limited accessibility problem of e-rubrics for some students:

P7. *... I would say the students' ability to access the feedback. This is not related to our courses. This is rather related to their IT skills.... If it were for me, I would ask them to be trained by IT before they ever access our classes... Their access to e-rubrics is limited because of their lack of knowledge of how to navigate this technological thing, I mean the use of computers. Many of them, I discovered at the end of the year, could not access their feedback once and I had highlighted word by word almost then...in addition to the rubric. If it were a paper rubric, they would have it under their noses.*

Another student-related problem is students not reading the rubrics. This could be related to their lack of IT (information technology) skills or lack of interest in reading and or improving themselves in the course.

P9. *I honestly do not think that our students read the rubrics or the comments or anything or take notes unless you actually get them into your office and give them a lecture about what they are doing ...one to one. But I am sure they don't actually look at the rubric. They just look at the score. The grade is what gets them, and they will come after you and they will ask you “Why did you grade me like that?” They don't really read the rubric. They just want to know “Why did you give me this grade?”*

P 3. *Honestly, I had to show my students [how to use e-rubrics] multiple times. I teach them rubrics in Word document form and ... I had them click on their own assessments after I put*

*them on Turnitin and check the rubrics themselves and I did a dummy practice session in class just to have them see how much function and format were worth and how much content and grammar were worth. I think that did help them but as in the past I asked some of my students they never even saw, they never even read, they never even tried to. Honestly, I don't think they pay attention to rubrics at all until they have to.*

Students do not read the rubric and instructors spend a lot of time and effort to show them how to access the e-rubric and even give a verbal recap of the feedback to the student, so that they can take some action to improve their essay. They also spend quite an amount of time to construct a successful e-rubric to be able to communicate the course expectations to their students.

### 3.1.2. Research question 2

The instructors were asked how e-rubrics compare to traditional, paper-based rubrics. The analysis of the data regarding this research question revealed three main themes: reducing the workload in grading, providing instant feedback to students, and safe record keeping. Besides these advantages, two instructors also mentioned that paper-based rubrics are as efficient as e-rubrics, although e-rubrics' superiority in providing immediate feedback to students is undisputed, as shown in [Table 4](#) below.

**Table 4.** *E-rubrics vs. paper rubrics.*

Advantages of e-rubrics over paper rubrics	Frequency
Reducing the workload in grading	9
Making feedback instantly accessible to students	5
Safe record keeping	3

All instructors agreed that using e-rubrics reduces the time and effort spent on grading and scoring student essays. Reducing the workload was defined as making grading easier, quicker, and providing flexibility in adapting the same rubric for other assignments and courses. Direct quotes from instructors to support their viewpoints are provided below:

*P5. They are faster in the sense that if you got the essay, you got the rubric. It is part of the same sort of interface. Therefore, I think you can mark or grade the essay more quickly with more clarity because you can see the rubric and you can see the essay at the same time.*

*P8. They are much faster and much easier. I like the use of Turnitin. It makes it a lot easier. I like the fact that the rubrics are attached, and quick marks are easy to use. You can give students instant feedback on their issues and Turnitin seems to be able to track students' progress or similarities in the issues that students have. The report that Turnitin can generate would help me to assess a class's level and the issues they have and what we need to do about it and how to improve. The descriptors on the attached rubrics are very clear and easy to use, and I like the way they can change as you move across each section of the rubric.*

Providing immediate feedback to students was perceived to be another superiority of e-rubrics. Instant feedback means students are able to act on the feedback quicker and will have more time to revise their essay.

*P3. E-rubrics are much easier to get through. The weighting of each and every single category is kind of difficult because...for example, content could be scored higher than grammar and mechanics. I think e-rubrics are beneficial not just for me as a teacher to make the grading process much easier but also easier for the student to see how they can achieve let's say a higher grade in each and every single assessment and what to avoid.*

P6. *The only thing is that it is just easier to select the scale and they are more accessible to students because they would see exactly where they would fall on every category so I would say they are more accessible and they are more flexible but in terms of effectiveness or efficiency, I would say they are same as paper rubrics.*

Safer record-keeping particularly preventing papers getting lost is another advantage that the e-rubrics offer as opposed to paper rubrics. Students and instructors alike will not have to file essays and their assessment in a folder; they will always be online.

P7. *E-rubrics are easier to use, of course. You don't need to print them out beforehand and prepare them. You know all this; they are there all the time so they cannot be lost. The students can refer to them. This is a very good thing because you cannot lose this. Also, e-rubrics are easier so basically you are marking things online. You don't have to go through the hassle of printing and then writing on paper.*

P1. *Basically, paper rubrics and e-rubrics are about the same. But they, e-rubrics, have two advantages. Students can't lose them, so they are always there and they are available. For conferencing I don't have to go hunting through anything. I just open up Turnitin and find the student's name and there is the rubric attached.*

#### **3.1.4. Research question 4**

The last question in the qualitative section is what instructors think about the impact of rubrics on student learning. This question did not elicit as straightforward responses as the previous ones, because not all instructors were convinced that e-rubrics have a significant impact on students' writing performance and their responses generally ranged in a continuum of "rubrics have no impact" to "rubrics have a limited impact on student learning".

Therefore, instead of creating categories, and displaying the frequency of responses, only direct quotations were given below to reflect the participants' observations from pessimistic to optimistic in regards to the positive impact of rubrics on student learning.

P9. *So, I haven't seen any improvement or any awareness in my students. Not from the rubrics. They don't really read the rubric. They just want to know "Why did you give me this grade?"*

P1 *I have to be honest, no. Well...maybe 5 percent. The students that use them, most of them don't even bother. I would say only 5 percent of the students, but I would still argue for them because it protects us especially if a student wants to do a grade appeal.*

P4. *To an extent, yes, but not completely... This is because students won't necessarily read the whole thing, which means they won't clearly understand what is expected of them.*

P5. *Yes, for the students who care, I think I could say there has been an improvement. I wouldn't say a huge improvement, but I would say there is an improvement because they are more conscious of what they need to be putting into each paragraph and they are more conscious of the necessary structures and because they have a better awareness of the grades they will get for individual parts of an essay.*

P6. *I believe students who care do benefit from rubrics. First, because when they look at the rubric, they... before even working on their assignments... they know exactly how they are being graded and they can get an idea about what's expected for them to get an A. I think it lays out the expectations very clearly and they usually post it with the assignment sheet and they say this is what you are supposed to do to get an A and I sometimes use it for in-class activities like during the drafting stage I have students grade each other's papers using the grading rubric so that helps them see things from my perspective and understand what exactly I expect so this is very helpful in terms of working on the assignment.*

As the responses display, writing faculty think students benefit from rubrics only when they care and when the course instructors spend a substantial amount of class time or one-on-one



tutoring time to explain how the e-rubrics works. This could be interpreted as they do not expect rubrics to work a miracle and assume their students learn from e-rubrics on their own.

### 3.1.5. Research question 5

In the quantitative part of the study, MFRM was applied to the raters' data to examine if there is any potential link between their beliefs towards using e-rubric and their rating behaviour. In a rating situation, raters are expected to perform similar levels of severity/leniency which is not always possible. The severity measures by FACETS are used to analyze if the inconsistencies among raters' severity levels are significant. According to the results on the relevant output (Table 5), P7 was found the most severe in ratings with a measure of 2.62 while P6 was found the most lenient with a measure of -2.40. The other raters' severity measure varied within a small range -.43 and .65, indicating that those four instructors were not substantially different in terms of the levels of severity they exercised.

**Table 5.** Rater-related statistics.

Rater ID	Measure*	S.E.***	MnSq**
P1	.60	.07	2.77
P2	.13	.09	.61
P5	.64	.17	.51
P6	-2.40	.08	.23
P7	2.62	.12	.54
P9	-.43	.05	.72

\*Measure: Rater's severity measure in log-odd units. Higher value indicates a more severe rater and vice versa.

\*\*MnSq: Outlier sensitive fit statistics value for each rater.

\*\*\*Standard error of the estimated measure

Another statistic analyzed in this part is the rater separation index. Rater separation index value shows the number of different strata of severity in the raters. Since the raters are expected to perform a similar level of severity in theory, the expected value of this statistics is 0. In the study, the separation index value found 5.23 which indicates that within 6 raters, there were about six statistically distinct strata of severity. Reliability of rater separation is also checked to understand rater severity. It shows the degree that the raters can be differentiated in terms of their severity level. Similar to the rater separation index, a value of 0 is expected for this statistic. This value was found .29 which indicates the raters were differentiated fairly in terms of levels of severity they exercised. Lastly, fixed (all same) chi-square and its significance value which test if the raters significantly differ in their levels of severity were checked. According to the results, the chi-square value of 14.2 with a significance value smaller than .01 indicates that the severity measures for the raters were not all the same, after allowing for measurement error.

### 3.1.6. Research question 6

In addition to the severity measures that outline systematic rating behaviors, we analyzed rater fit statistics, particularly mean-square infit and outfit indices which show the degree of unexpected ratings by a rater summarized over examinees. Fit statistics allows researchers to examine if any rater effect exists in the ratings for some examinees or items/tasks. Based on the results, only P1 had a mean-square outfit value (2.77) that implies a significant inconsistency in ratings. This finding suggests that the particular rater may have adopted an *idiosyncratic rating style* (Eckes, 2011) since mean-square outfit statistics is sensitive to outliers in ratings.



#### **4. DISCUSSION**

Our findings on the first research question showed that writing teachers benefit greatly from e-rubrics. They specifically expressed that using e-rubrics makes grading transparent, reduces student complaints and grade appeals, facilitates grading and helps standardization among raters. Recently e-rubrics have also been analyzed as computer assisted grading rubrics and LMSs (such as Blackboard and Moodle) integrated rubrics, however there are not many studies conducted on teachers' perceptions towards electronic rubrics, probably due to its novelty.

In general, rubrics are perceived positively by academic faculty and teachers. Instructors have a positive attitude towards rubrics as a handy evaluative and instructional tool for nurturing students' learning, self-assessment, and self-regulation (Sharma, 2019); and most teachers agree that rubrics act as a guideline for students to know what the criteria are in order to get a good grade and that they became more consistent in grading since they started using rubrics (Qasim & Qasim, 2015). Regarding e-rubrics, Raddawi and Biliközen (2018) express that the instructors think using e-rubrics helps them with record keeping, saves time and energy in grading and giving feedback and provides objective assessment of student essays. Atkinson and Lim (2013) also suggest e-rubrics are efficient for grading and giving feedback, they provide detailed guidance to students, and promote standardization.

An interesting finding that stands out in our study is the benefit of rubrics in reducing student complaints or grade appeals. Although we have not been able to see a similar finding in other recent studies, Fulbright (2018) suggests that rubrics add transparency to grading, which is important when explaining to disgruntled students that they were not given a certain score because the instructor did not like them, but because they omitted one or more components of the required criteria of the assignment. Fulbright (2018) asserts that rubrics give faculty the needed documentation of objective assessment that is essential for grade appeals and even in court. The fact that this benefit of rubrics has been frequently expressed by the instructors in this institution may be related to the culture of the institution. If instructors are getting a high number of grade appeals and complaints in this institution, the use of rubrics may have been adopted as a defense mechanism.

The second finding was about the limitations of e-rubrics and the writing faculty expressed that they face scoring issues such as rubrics not always providing them with the clarity and flexibility they need to score papers. Also, some problems confronted by students such as not being able to access or not reading the rubric and sometimes even not understanding the rubric are frequently mentioned by the instructors interviewed.

Raddawi and Biliközen's (2018) study elicited similar responses from writing instructors with regards to the rigidity of rubrics. Lack of flexibility that they experience while grading essays with a rubric seems to be restricting teachers' freedom. In their study technical problems were expressed by the writing faculty, however in our study this was attributed as a student challenge due to their lack of IT skills. The fact that no teacher has mentioned any technical problems in regard to the use of e-rubrics is a sign that they have all mastered the use of e-rubrics in the institution.

Another similar finding comes from a study conducted on student perception on the e-rubrics. In Atkinson and Lim's study (2013) when students were asked to make suggestions to improve the rubrics their professors use, some of them said "less ambiguity and clearer requirements and relevant depth are needed. It was hard to customize the task because the structure was already defined." Some students even mentioned they would like to see more freedom included in the assignment task. A similar study analyzing university students' perceptions towards using rubrics showed that despite the many positive aspects of using rubrics, a small percentage of students indicated that the rubric lacked clarification because it contained standard feedback

for everybody (Raposo-Rivas & Gallego-Arufat, 2016) Sharma (2019) also express that especially low achieving students do not show any interest in practices like using rubrics for self-assessment unless they are specifically trained to do so, which also supports our interviewees' responses in regards to students not reading the rubrics and not accessing the rubrics (unless the professors spend a long session on how to access the rubrics online).

The third finding of the study was about how e-rubrics compare to traditional rubrics in the writing faculty's opinion. This research question revealed that although they are more or less the same in content and structure, e-rubrics are superior to paper rubrics because they reduce the workload in grading, make feedback instantly accessible to students and help instructors with record keeping.

A study comparing the use of computer-assisted grading rubrics to other grading methods and their results suggested that the computer-assisted grading rubrics were almost 200% faster than traditional hand grading without rubrics, more than 300% faster than hand grading with rubrics, and nearly 350% faster than typing the feedback into a Blackboard or Moodle (Anglin et al., 2008). Atkinson and Lim (2013) also found out that a key benefit of e-rubrics was around 40% reduction in marking time. Indeed, rubrics embedded in a learning management system (in our study, Moodle) not only make grading faster, but also record keeping much easier as they facilitate student submissions, and help faculty track details such as student names, uploaded files, similarity rates and the time of submission. Besides, automated calculations enabled by the LMSs ensure speed and accuracy. Results are available for general feedback to students, and for examination and auditing by other stakeholders.

The last finding of the qualitative analysis is about the perceptions of the writing faculty on whether rubrics have any impact on student learning. The interviewees in this study were slightly hesitant to make comments about the positive impact of rubrics on student writing; instead, they tended to see some improvement in students who really cared about self-assessment and who made the effort to go through the rubric and attended conference hours organized by the professor to get more feedback along with the extra explanation in the rubric.

This cautious optimism about rubrics could be also found in some studies done previously. Reddy and Andrade (2010) express that there is a striking difference between students' and instructors' perceptions of rubric use. College students value rubrics because they clarify the targets for their work, allow them to regulate their progress, and make grades transparent and fair. While students referred to rubrics as tools serving the purposes of learning and achievement, instructors focused almost solely on the role of a rubric as a tool to assign grades quickly, objectively, and accurately. This could be the underlying reason for the interviewees in this study to have doubts about the value of rubrics as a teaching tool because they value rubrics as something that makes their grading easier and reduces student complaints.

A rubric skeptic, Krane (2018) also suggests rubrics do not teach students how to write and foster deep learning. She conducted a study with her 88 students and 69% agreed or strongly agreed that rubrics should always be given with writing assignments; they liked discussing rubrics in their classes and referring to them when working on assignments. 86% noted rubrics helped them to understand what the professor wants. 83% noted rubrics helped them to understand assignment criteria, and 74% noted that rubrics helped them "to know what they can do to get a better grade". However, when the students were asked whether rubrics helped them to improve their writing in general, only 21% agreed.

Rubrics are assignment specific; therefore, they highlight what will be assessed in a given assignment. Therefore, they work well in the short run, especially when students need guidance and a roadmap to get a good grade. However, as the interviewees' responses suggested, this roadmap does not necessarily help them develop problem solving skills that would improve

them as writers in the long run. Not all writing instructors think rubrics work miracles for students who do not make the effort to develop their writing skills. Those who care about rubrics are probably conscientious students who would strive to develop their writing without a rubric, anyway.

Quantitative results overlapped with qualitative findings, revealed by the raters' self-report. For example, the finding related to research question five was about the raters' severity and leniency. Although there is no consensus on personal and situational determinants of rater severity effect, professional experience is commonly cited as a factor leading to severity effect (Eckes, 2011). In our study, P7 was the most severe and P6 was the most lenient rater, as per our analysis. This could partially explain the most severe rater's (P7) rating behavior who has only 6 months of experience with e-rubrics (s/he has only 6 months of experience in the institution, as well). This may be the reason why as the least experienced rater, P7 may have tended to notice even the smallest errors in the name of using e-rubric carefully and turned out to be a harsh grader.

Another interesting finding in the quantitative study was P1 turning out to be an inconsistent rater, as suggested by FACETS. Inconsistencies in grading stem from several factors related to the problem being graded, the individual grader, the time of day, the grader's level of fatigue, and the grader's overall experience. Graders are also affected by their general values and beliefs about grading, such as values of non-achievement factors, like effort, and perceptions that grades function as rewards or punishments (Hicks & Diefes-Dux, 2017).

Reddy and Andrade (2010) emphasize the striking difference between students' and instructors' perceptions of rubric use; while students referred to rubrics as a learning tool, instructors focused on the role of a rubric as a tool to assign grades quickly, objectively, and accurately. In P1's case, this purpose may have been solely to "stuff the grades" s/he is giving and justifying this grade in case s/he gets a complaint or a grade appeal. This "limited conception of the purpose of a rubric" (Reddy & Andrade, 2010) might have contributed to their unwillingness to use the rubric properly and consequently, assign their grades randomly (p.439).

## **5. CONCLUSION & RECOMMENDATIONS**

The findings of this study reveal that although freshman composition instructors seem to be enjoying the advantages of e-rubrics, they differ in their rating behavior while implementing the standard e-rubric in assessing writing. Rater severity and inconsistency may cause dissatisfaction and create a sense of unfairness among the students of such instructors; therefore, instructors should be made aware of their rating behavior, so that they can avoid repeating them. Instructors may require some clarification on how to interpret some items in the rubric and they may need to be convinced about the educational value of rubrics as well as the evaluation or justification tool as they have been using it.

The most common way of fulfilling this goal is training sessions, where instructors are introduced to a set of criteria and then they are asked to rate essays based on those criteria. The results show whether and to what extent they are on the same page as other raters and therefore interpret the rating criteria similarly. Raters who still show severity/leniency and/ or inconsistency may take additional rater training sessions to prevent such rater effects in the future. Organizing a norming session before every essay assessment is also an efficient means for departmental standardization. Consequently, the results of this study are expected to help this writing department to be more standardized in its ratings.

Writing assessment is a challenging job and writing teachers invest a lot of time and effort into helping their students improve their writing. Their rating behaviour may differ but writing scholars are interested in working collaboratively to discover the most effective method of assessing writing. Therefore, the implications of this study should not be seen as limited to the

particular institution or the region where the study was held. Even though the findings cannot be generalized, they bring out some serious considerations concerning the application of rubrics as writing assessment tools in an EFL (English as a foreign language) context. This study has implications for many rater-mediated language assessment situations, particularly in small-scale academic programs.

Last but not the least, a recommendation for researchers could be regarding the scope of such a study, which could be further enhanced by collecting data from students regarding their perceptions of e-rubric as an instruction and assessment tool and even assessing their rating behavior while they are using the rubric in peer grading.

### 5.1. Limitations

This study had several limitations. For the qualitative part, it should be noted that the faculty members probably had the standard e-rubric in mind that was in place at the time of the interview, so their perceptions and experiences were predominantly shaped by that particular e-rubric they were required to use in the department.

For the quantitative part, the limitations are the e-rubric and grading scale that are in use as well as the genre (research paper) that the students submitted as their assignment. It should also be added as a limitation that this was the second draft of the assignment, which means the students revised their first drafts based on the feedback they received from their instructor and submitted an improved version as a second draft. This may have positively skewed their grades.

### Acknowledgments

This work was funded by Gulf University for Science and Technology, Kuwait [Internal Seed Grant number 187226].

### Declaration of Conflicting Interests and Ethics


The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

### Authorship Contribution Statement

**Inan Deniz Erguvan:** Data Collection, Resources, Qualitative Analysis, and Writing the Draft.  
**Beyza Aksu Dünya:** Software, Quantitative Data Analysis and Reporting, Editing the Draft.

### ORCID

Inan Deniz ERGUVAN  <https://orcid.org/0000-0001-8713-2935>

Beyza AKSU DÜNYA  <https://orcid.org/0000-0003-4994-1429>

## 6. REFERENCES

- Allen, D., & Tanner, K. (2006). Rubrics: tools for making learning goals and evaluation criteria explicit for both teachers and learners. *CBE Life Sciences Education*, 5(3), 197-203. <https://doi.org/10.1187/cbe.06-06-0168>
- Andrade, H. G. (2000). Using rubrics to promote thinking and learning. *Educational Leadership*, 57(5), 13-18.
- Andrade, H. G. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College Teaching*, 53(1), 27. <https://doi.org/10.3200/CTCH.53.1.27-31>
- Anglin, L. Anglin, K., Schumann, P.L., & Kalinski, J. A. (2008). Improving the efficiency and effectiveness of grading through the use of computer-assisted grading rubrics. *Decision Sciences: Journal of Innovative Education*, 6(1), 51-73. <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-4609.2007.00153.x>

- Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Corwin Press.
- Atkinson, D. & Lim, S.L. (2013). Improving assessment processes in higher education: Student and teacher perceptions of the effectiveness of a rubric embedded in a LMS. *Australasian Journal of Educational Technology*, 29(5), 651-666. <https://doi.org/10.14742/ajet.526>
- Broad, B. (2000). Pulling your hair out: "Crises of standardization in communal writing assessment". *Research in the Teaching of English*, 35(2), 213-260. [www.jstor.org/stable/40171515](http://www.jstor.org/stable/40171515)
- Brookhart, S. M. (2013). *How to Create and Use Rubrics for Formative Assessment and Grading*. ASCD.
- Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education*, 3. <https://www.frontiersin.org/articles/10.3389/educ.2018.00022/full>
- Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review*, 67(3), 343-368. <https://doi.org/10.1080/00131911.2014.929565>
- Brown, J. (2001). *Using surveys in language programs*. Cambridge University Press.
- Carr, N. T. (2000). A comparison of the effects of analytic and holistic rating scale types in the context of composition tests. *Issues in Applied Linguistics*, 11(2), 207-241 <https://escholarship.org/uc/item/4dw4z8rt>
- Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches* (2nd ed.). SAGE.
- Creswell, J. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Pearson Education.
- Creswell, J. W., & Clark, V. L. (2011). Choosing a mixed methods design. In *Designing and conducting mixed methods research* (3rd ed., pp. 53, 106). SAGE.
- Dobria, L. (2011). *Longitudinal rater modeling with splines*. (Publication no. 3472389) [Doctoral dissertation, University of Illinois at Chicago]. ProQuest digital dissertations.
- Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Peter Lang.
- Freeman, S., & Parks, J. W. (2010). How accurate is peer grading? *CBE Life Sciences Education*, 9(4), 482-488. <https://doi.org/10.1187/cbe.10-03-0017>
- Fulbright, S. (2018, October 18). *Using rubrics as a defense against grade appeals*. Faculty Focus. <https://www.facultyfocus.com/articles/course-design-ideas/rubrics-as-a-defense-against-grade-appeals/>
- Garcia-Ros, R. (2011). Analysis and validation of a rubric to assess oral presentation skills in university contexts. *Electronic Journal of Research in Educational Psychology*, 9(3), 1043-1062.
- Hafner, J.C., & Hafner, P.M. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education*, 25(12), 1509-1528. <https://doi.org/10.1080/0950069022000038268>
- Hicks, N., & Diefes-Dux, H. (2017). Grader consistency in using standards-based rubrics. 2017 ASEE Annual Conference & Exposition Proceedings. <https://doi.org/10.18260/1-2--28416>
- Iramaneerat, C., & Yudkowsky, R. (2007). Rater errors in a clinical skills assessment of medical students. *Evaluation & the Health Professions*, 30(3), 266-283. <https://doi.org/10.1177/0163278707304040>
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130-144. <https://doi.org/10.1016/j.edurev.2007.05.002>



- Kirwin, J., & DiVall, M. (2015, October). Using electronic rubrics to produce actionable assessment data in a skills-based course [Conference session]. 2015 Assessment Institute in Indianapolis, Indianapolis. <https://assessmentinstitute.iupui.edu/>
- Kohn, A. (2006). Speaking my mind: The trouble with rubrics. *English Journal*, 95(4), 12-15. <https://doi.org/10.2307/30047080>
- Krane, D. (2018, August 30). *Guest post: What students see in rubrics*. Inside HigherEd. <https://www.insidehighered.com/blogs/just-visiting/guest-post-what-students-see-rubrics>
- Linacre, J. M. (2019). *A user's guide to FACETS: Rasch-model computer programs*. <https://www.winsteps.com/tutorials.htm>
- Linacre, J. M. (2020). *FACETS* (Version 3.83.2). <https://www.winsteps.com/winbuy.htm>
- Linacre, J. M. (2002c). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16, 878. <https://www.rasch.org/rmt/rmt162f.htm>
- Martínez, D., Cebrián, D., & Cebrián, M. (2016). Assessment of teaching skills with e-Rubrics in Master of Teacher Training. *Journal for Educators, Teachers and Trainers*, 7(2). 120-141. [https://jett.labosfor.com/article\\_855.html](https://jett.labosfor.com/article_855.html)
- Mertler, C. A. (2009). *Action research: Teachers as researchers in the classroom* (2nd ed.). SAGE.
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. SAGE.
- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement. *Journal of Applied Measurement*, 5(2), 189-223.
- Myford, C. M., & Dobria, L. (2012). *FACETS introductory workshop tutorial*. University of Illinois at Chicago.
- Nordrum, L., Evans, K., & Gustafsson, M. (2013). Comparing student learning experiences of in-text commentary and rubric-articulated feedback: Strategies for formative assessment. *Assessment & Evaluation in Higher Education*, 38, 919-940. <https://doi.org/10.1080/02602938.2012.758229>
- Qasim, A., & Qasim, Z. (2015). Using Rubrics to Assess Writing: Pros and Cons in Pakistani Teachers' Opinions. *Journal of Literature, Languages and Linguistics*, 16, 51-58. <https://www.researchgate.net/publication/285815750>
- Raddawi, R., & Bilikozen, N. (2018). ELT professors' perspectives on the use of E-rubrics in an academic writing class in a University in the UAE. *Assessing EFL Writing in the 21st Century Arab World*, 221-260. [https://doi.org/10.1007/978-3-319-64104-1\\_9](https://doi.org/10.1007/978-3-319-64104-1_9)
- Radnor, H. A. (2002). *Researching your professional practice: Doing interpretive research*. Open University Press.
- Raposo-Rivas, M., & Gallego-Arrufat, M.J. (2016). University students' perceptions of electronic rubric-based assessment. *Digital Education Review*, 30, 220-233. <http://greav.ub.edu/der/>
- Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, 35(4), 435-448. <https://doi.org/10.1080/02602930902862859>
- Rivasy, M. R., De La Serna, M. C., & Martínez-Figueira, E. (2014). Electronic rubrics to assess competencies in ICT subjects. *European Educational Research Journal*, 13(5), 584-594. <https://doi.org/10.2304/eeerj.2014.13.5.584>
- Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88(2), 413-428.
- Sadler P.M., & Good E. (2006). The impact of self-and peer-grading on student learning. *Educational Assessment*, 11(1), 1-31. [https://doi.org/10.1207/s15326977ea1101\\_1](https://doi.org/10.1207/s15326977ea1101_1)



- Sharma, V. (2019). Teacher perspicacity to using rubrics in students' EFL learning and assessment. *Journal of English Language Teaching and Applied Linguistics*, 1(1), 16-31. <https://www.researchgate.net/publication/337771674>
- Silverman, D. (2000). *Doing qualitative research: A practical handbook*. SAGE.
- Steffens, K. (2014). E-rubrics to facilitate self-regulated learning. *REDU.Revista de Docencia Universitaria*, 12 (1), 11-12. <https://doi.org/10.4995/redu.2014.6417>
- Torrance, H. (2007). Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning. *Assessment in Education: Principles, Policy & Practice*, 14 (3), 281-294. <https://doi.org/10.1080/09695940701591867>
- Wilson, M. (2007). Why I won't be using rubrics to respond to students' writing. *English Journal*, 96(4), 62-66. <https://doi.org/10.2307/30047167>

## 7. APPENDIX

The rubric for the research paper

	<b>Excellent 100</b>	<b>Very good 90</b>	<b>Good 80</b>	<b>Average 70</b>	<b>Inadequate 60</b>	<b>Poor 10</b>
<b>Research Elements (Sources &amp; Quotations)</b>  25%	Paper uses 6-7 quality sources (mostly scholarly), and provides quotations from a variety of sources in <u>every</u> body paragraph. The quotations support the topic sentences well. Support and evidence are expressed in the writer's words (paraphrased & summarized).	Paper uses min 5, preferably 5-6 quality sources (popular & scholarly), and provides quotations from a variety of sources in <u>every</u> body paragraph. The quotations support the topic sentences well. Support and evidence are usually expressed in the writer's words.	Paper uses 5 good sources, and provides quotations from a variety of sources in <u>most</u> body paragraphs. The quotations generally support and develop the topic sentences. Support and evidence are mostly expressed in the writer's words, but some direct quotes are unnecessary.	Paper uses 4-5 sources, but 1 -2 may be weak or not academic enough. The evidence may be irrelevant/ weak in 1-2 body paragraphs. Support and evidence are not always expressed in the writer's words. Word count may be low due to lack of sources & quotations.	Paper uses and provides quotations from less than 4 sources. The evidence is weak and irrelevant, does not develop the thesis. Support and evidence are not expressed in the writer's words. Word count is below 1000.	There is no indication of research. No sources have been used. Very low word count, or high similarity rate. Too many direct quotations.
<b>Organization &amp; Connectors</b>  25%	Introduction is interesting with detailed background and a clear thesis statement. Topic sentences introduce the arguments, body paragraphs fully explore the topic and present information in a logical order. Counter argument has a strong refutation. Conclusion restates the thesis and contains original opinions. Effective and varied transitions link all ideas.	Introduction gives good background and contains a clear thesis statement. Topic sentences introduce the arguments; body paragraphs explore the topic and present information in a logical order. Counter argument has a relevant refutation. Conclusion restates the thesis and contains opinions. Transitions link all ideas.	Introduction gives some background and contains a thesis statement. Topic sentences exist in every body paragraph. The arguments in body paragraphs explore the topic and present information in an acceptable order. There is a counter argument with some refutation. Conclusion restates the thesis and/or offers a comment. Transitions link most ideas.	There is an underlying organization, but paragraph division may not always be efficient. Introduction contains a thesis statement. The body explores the topic and presents information, but not always clear or logical. Counter argument and/or refutation is weak or refutations is nonexistent. Some transitions are used, but more are needed.	The paper is generally hard to follow. The writing lacks strong organization and it may also lack a clear thesis statement. The body presents some support, but not all relevant. Transitions may be used inconsistently or may be lacking.	There is no clear paragraphing. Introduction, body and conclusion paragraphs are not clearly divided. Lack of transitions impedes fluency.

<p><b>Grammar / Mechanics/ Spelling/</b></p> <p>25%</p>	<p>There are 1-2 minor errors in grammatical accuracy. Spelling and punctuation may contain 1-2 typos. Word choice is appropriate for an academic research paper. Complex sentences (noun, adverb, adjective clauses) are frequently used, without errors.</p>	<p>There are 3-4 minor errors in grammatical accuracy. Spelling and punctuation may contain 1-2 errors. Word choice is generally good for an academic research paper. Compound &amp; complex sentences are frequently used, with minor errors.</p>	<p>Up to 5 errors in grammatical accuracy, spelling, or punctuation may exist. The use of academic words is acceptable. Sentence variety is not as expected, complex sentences may contain max 3 errors.</p>	<p>There are (max 10) errors in grammatical accuracy; some may detract from the meaning. There is not enough evidence of academic vocabulary. Preference for simple sentences, rather than complex.</p>	<p>There are more than 10 grammar, spelling &amp; punctuation errors. Word choice is incorrect or inappropriate in most places. Writing is choppy, with many awkward or unclear passages.</p>	<p>Very poor use of English with no sense of correct grammar. Google translation or synonym finder may have been used.</p>
<p><b>APA &amp; Formatting</b></p> <p>25%</p>	<p>Consistently uses accurate in-text citations and has a flawless Reference page (alphabetical, double-spaced, in hanging indent format). All sources cited in the essay are listed on the Reference page. Entire essay is double-spaced using Times New Roman font with 1-inch margins. Student's name, instructor's name, course, date appear as per APA guidelines. 1 or 2 minor errors may exist in overall formatting. (Essays with no title cannot get excellent in this category)</p>	<p>Max 3 minor errors exist in in-text citations and/or Reference page. Essay is notably lacking in 1-2 items in the Excellent category (For example, double spacing / margins or etc.)</p>	<p>In-text citations and/or Reference page have 5 major errors. Essay is notably lacking in 3-4 items in the Excellent category (For example, double spacing / margins or header etc.).</p>	<p>The references page may lack 1-2 sources or contain formatting errors. There are 5-6 errors (page number, font, spacing) and major (in text citations &amp; references) in APA style.</p>	<p>Most formatting is incorrect or inconsistent. APA is not adhered to in in-text citations and references. There are more than 5 major errors.</p>	<p>No references page submitted or lists some sources with incorrect style. No formatting style has been followed throughout the paper. Citations don't exist.</p>

**Note:** The large gap between Inadequate (60) and Poor (10) intends to serve the purpose of differentiating D students from F students. D students are still considered passing and with a little more effort can get a C. However, F students perform quite below the expected standards, therefore they should make greater effort to pass the assignment and eventually the course. It should also be noted that as per the feedback from course instructors, rubrics are revised every year.