Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309-6575 *Güz* 2023 *Autumn* 2023 *Cilt: 14-Sayı: 3* *Volume: 14-Issue: 3*

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309 – 6575

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

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

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

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

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

Hakem Kurulu / Referee Board

Gökhan AKSU (Adnan Menderes Üni.)
Görkem CEYHAN (Muş Alparslan Üni.)
Gözde SIRGANCI (Bozok Üni.)
Gül GÜLER (İstanbul Aydın Üni.)
Gülden KAYA UYANIK (Sakarya Üni.)
Gülşen TAŞDELEN TEKER (Hacettepe Üni.)
Hakan KOĞAR (Akdeniz Üni.)
Hakan SARIÇAM (Dumlupınar Üni.)
Hakan Yavuz ATAR (Gazi Üni.)
Halil İbrahim SARI (Kilis Üni.)
Halil YURDUGÜL (Hacettepe Üni.)
Hatice KUMANDAŞ (Artvin Çoruh Üni.)
Hikmet ŞEVGİN (Van Yüzüncü Yıl Üni.)
Hülya KELECİOĞLU (Hacettepe Üni.)
Hülya YÜREKLI (Yıldız Teknik Üni.)
İbrahim Alper KÖSE (Bolu Abant İzzet Baysal Üni.)
İbrahim YILDIRIM (Gaziantep Üni.)
İbrahim UYSAL (Bolu Abant İzzet Baysal Üni.)
İlhan KOYUNCU (Adıyaman Üni.)
İlkay AŞKIN TEKKOL (Kastamonu Üni.)
İlker KALENDER (Bilkent Üni.)
İsmail KARAKAYA (Gazi Üni.)
Kübra ATALAY KABASAKAL (Hacettepe Üni.)
Levent ERTUNA (Sakarya Üni.)
Levent YAKAR (Kahramanmaraş Sütçü İmam Üni.)
Mehmet KAPLAN (MEB)
Mehmet ŞATA (Ağrı İbrahim Çeçen Üni.)
Melek Gülşah ŞAHİN (Gazi Üni.)
Meltem ACAR GÜVENDİR (Trakya Üni.)
Meltem YURTÇU (İnönü Üni.)
Metin BULUŞ (Adıyaman Üni.)
Murat Doğan ŞAHİN (Anadolu Üni.)
Mustafa ASİL (University of Otago)
Mustafa İLHAN (Dicle Üni.)
Nagihan BOZTUNÇ ÖZTÜRK (Hacettepe Üni.)
Nail YILDIRIM (Kahramanmaraş Sütçü İmam Üni.)
Neşe GÜLER (İzmir Demokrasi Üni.)
Neşe ÖZTÜRK GÜBEŞ (Mehmet Akif Ersoy Üni.)
Nuri DOĞAN (Hacettepe Üni.)
Nükhet DEMİRTAŞLI (Emekli Öğretim Üyesi)
Okan BULUT (University of Alberta)
Onur ÖZMEN (TED Üniversitesi)
Ömer KUTLU (Ankara Üni.)
Ömür Kaya KALKAN (Pamukkale Üni.)
Önder SÜNBÜL (Mersin Üni.)
Özen YILDIRIM (Pamukkale Üni.)
Özge ALTINTAS (Ankara Üni.)
Özge BIKMAZ BİLGEN (Adnan Menderes Üni.)
Özlem ULAŞ (Giresun Üni.)
Recep GÜR (Erzincan Üni.)
Ragıp TERZİ (Harran Üni.)
Sedat ŞEN (Harran Üni.)
Recep Serkan ARIK (Dumlupınar Üni.)
Safiye BİLİCAN DEMİR (Kocaeli Üni.)
Selahattin GELBAL (Hacettepe Üni.)
Seher YALÇIN (Ankara Üni.)
Selen DEMİRTAŞ ZORBAZ (Ordu Üni.)
Selma ŞENEL (Balıkesir Üni.)
Seçil ÖMÜR SÜNBÜL (Mersin Üni.)
Sait Çüm (MEB)
Sakine GÖÇER ŞAHİN (University of Wisconsin Madison)
Sema SULAK (Bartın Üni.)
Semirhan GÖKÇE (Niğde Ömer Halisdemir Üni.)
Serkan ARIKAN (Boğaziçi Üni.)
Seval KIZILDAĞ ŞAHİN (Adıyaman Üni.)
Sevda ÇETİN (Hacettepe Üni.)
Sevilay KİLMEN (Abant İzzet Baysal Üni.)
Sinem Evin AKBAY (Mersin Üni.)
Sungur GÜREL (Siirt Üni.)
Süleyman DEMİR (Sakarya Üni.)
Sümeyra SOYSAL (Necmettin Erbakan Üni.)
Şeref TAN (Gazi Üni.)
Şeyma UYAR (Mehmet Akif Ersoy Üni.)
Tahsin Oğuz BAŞOKÇU (Ege Üni.)
Terry A. ACKERMAN (University of Iowa)
Tuğba KARADAVUT (İzmir Demokrasi Üni.)
Tuncay ÖĞRETMEN (Ege Üni.)
Tülin ACAR (Parantez Eğitim)
Türkan DOĞAN (Hacettepe Üni.)
Ufuk AKBAŞ (Hasan Kalyoncu Üni.)
Wenchao MA (University of Alabama)
Yavuz AKPINAR (Boğaziçi Üni.)
Yeşim ÖZER ÖZKAN (Gaziantep Üni.)
Yusuf KARA (Southern Methodist University)
Zekeriya NARTGÜN (Bolu Abant İzzet Baysal Üni.)
Zeynep ŞEN AKÇAY (Hacettepe Üni.)

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

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(3); 171-184

Ability Estimation with Polytomous Items in Computerized Multistage Tests

Hasibe YAHŞİ SARI* Hülya KELECİOĞLU**

Abstract

This study aims to examine how individuals' ability estimations change under different conditions in tests consisting of polytomous items in a computerized multistage test environment. In this simulation study, 108 (3x3x6x2=108) conditions were examined, consisting of three categories (3, 4, and 5), three test lengths (10, 20, and 30), six-panel designs (1-2, 1-2-2, 1-3, 1-3-3, 1-4, and 1-4-4), and, two routing methods (Maximum Fisher Information (MFI) and Random). Simulations and analyses were carried out in the mstR package in the R program, with a pool of 200 items, 1000 people, and 100 replications (i.e. iterations). The mean absolute bias, RMSE, and correlation values were calculated as the research outcomes. This study discovered that as the number of categories and test lengths increase, the mean absolute bias and RMSE values decrease, while the correlation values increase. Although MFI and random methods have similar tendencies regarding routing methods, MFI provides better results. Furthermore, there is a similarity between the panel designs in terms of results.

Keywords: Computerized multistage tests, polytomous items, routing method.

Introduction

Traditional paper-and-pencil tests have been replaced by computerized adaptive tests (CAT) in educational and psychological institutions. CATs are the tests in which the abilities of individuals are estimated with a scaled item pool before the exam, which has rules of starting, progressing, and ending according to the individual's previously known or predicted ability (Weiss, 1982). There are many advantages to CATs compared to traditional paper-and-pencil applications. For instance, an advantage of CATs is the increased accurate ability estimation by using fewer items and prompt disclosure of results (Weiss, 1983). However, CAT applications also have disadvantages such as different test lengths (i.e., fixed-length is also available), different questions being asked, and the individual not being able to return to the previous question. Due to the overwhelming disadvantages of CAT, the use of computerized multistage tests (MST) is becoming widespread (Hendrickson, 2007; MacGregor et al., 2022; Zenisky et al., 2009).

MST combines the advantages of CAT and paper-pencil tests. MST achieves this by adjusting the tests based on each individual. While CATs are adapted to the individual at the item level, MSTs are adapted to the individual at the module level (Zenisky et al., 2009). Unlike CATs, MSTs consist of item groups called modules and stages. Modules consist of items; stages consist of modules; panels consist of stages. MSTs provide the opportunity to move between the items in the module and allow test preparers to better control the test content compared to CATs (Hendrickson, 2007; Sari et al., 2016).

The characteristics of the item pool are important in MSTs, as in CATs. Unlike CATs, MSTs have their own terminology including panel structure, routing method, module, and stage. A module consists of a group of items at the same or similar difficulty level. A stage consists of a different number of modules at different difficulty levels such as easy, medium or hard modules. A panel design is comprised of

* Teacher, Ministry of National Education, Kilis-Türkiye, hsbyahsi@gmail.com, ORCID ID: 0000-0002-0451-6034
** Prof.Dr., Hacettepe University, Faculty of Education, Ankara-Türkiye, hulyakelecioglu@gmail.com, ORCID ID: 0000-0002-0741-9934

To cite this article:
Yahşi Sarı, H. & Kelecioğlu ,H. (2023). Ability estimation with polytomous items in computerized multistage tests. *Journal of Measurement and Evaluation in Education and Psychology, 14*(3), 171-184. https://doi.org/10.21031/epod.1056079

Received: 10.01.2022
Accepted: 21.12.2022

different stages. Test assembly is the process of building modules, stages, and panels so it is one of the most important steps in an MST.

An MST functions as follows: Individuals take the first-stage module, called the routing module. Then, the individual is selected for the appropriate module based on the current ability level at the second stage. Finally, the exam continues until a test taker completes all required stages.

Background and Literature Review

Various past studies on routing methods generally apply Approximate Maximum Information (AMI), Defined Population Intervals (DPI), and convergent and random routing methods (Kim et al., 2010; Zenisky, 2004). Routing methods that are based on IRT are other frequently used kinds. These methods are Maximum Fisher Module Information (MFI), Maximum Likelihood Weighted Module Information (MLWMI), Maximum Posterior Weighted Module Information (MPW MI), Maximum Module Kullback-Leibler Information (MKL), Maximum Posterior Module Kullback-Leibler Information (MKLP) and random. In this study, MFI and random routing methods were used. The MFI routing method is based on the item information level. In MST, routing with MFI is made to the next stage according to the cumulative information obtained from the module items. The MFI routing method directs individuals to the module, explaining their ability levels to the maximum (Weissman et al., 2007). In the random routing method, theta estimation is made after the module is taken in the routing module. Then, the individual is randomly assigned to a module in the next stage. On the other hand, individuals are referred to any of the following stage modules with equal probability, regardless of their scores in a previous stage.

One of the conditions of MST is panel design. A panel design is formed by the combination of different numbers of modules and stages. Panel design may vary depending on the purpose of the MSTs. For example, 1-3 panel patterns consist of 2 stages and four modules. There is 1 module in the first stage (also called the routing module) and three in the 2nd stage. In a 1-3 panel design, the difficulty levels of the modules are usually determined as easy, medium, and complicated in the 2nd stage. 1-2, 1-2-2, 1-3, 1-3-3, 1-4, and 1-4-4 panel designs, which are preferred in the literature, were used in this study (Kim et al., 2010; Oztürk, 2019; Sarı & Raborn, 2018).

It is known that test length affects ability estimation in MST designs (Luecht, 2000; Sarı & Raborn, 2018). Based on the literature, while some studies use different numbers of items at all stages (Macken-Ruiz, 2008), some other studies use the same number of items (Kim et al., 2013). Kim et al. (2013) compared the MST designs that they created based on the partial credit model, using different routing methods and panel designs, in the context of the classification test. As a result, it was observed that the accuracy of the ability estimations increased as the test length increased. Previous studies using polytomous test items mainly used 9-20 items (Chen, 2010; Kim et al., 2013; Macken-Ruiz, 2008). Based on the studies examined in the literature, 10, 20, and 30 test lengths were examined in this study.

MST applications are made with dichotomous (i.e. binary) and polytomous items. Zenisky (2004) compared various panel designs with different routing methods (DPI, proximity, and random) to estimate the ability and determine its precision. The item pool was based on the three-parameter logistic IRT model. Several studies in the literature examine the ability estimations of MST designs using two-category (i.e. binary) data using different conditions and routing methods (Oztürk, 2019; Sarı & Raborn, 2018; Zenisky, 2004). Polytomous items provide more information, allowing more accurate findings in ability estimation (Donoghue, 1994). However, few research studies use different routing methods in polytomous data. Studies in the literature which use polytomous items are generally designed according to the partial credit model (Kim et al., 2010; Kim et al., 2013). Nonetheless, GPCM is used in current studies and applications such as PISA 2018 (Choi, & Asilkalkan, 2019; Ridho, 2022). Thus, in this study, we utilized GPCM when generating and analyzing polytomous items.

This study is unique because it was designed with different panel designs, routing methods, and items produced according to the generalized partial credit model. In addition, AMI, DPI, M-AMI, M-DPI, SL-DPI, and ML-DPI routing methods are frequently used in the literature (Kim et al., 2010; Kim et al.,

2013; Zenisky, 2004). Some studies use MFI, MLWMI, MPWMI, MKL, MKLP, and random routing methods with dichotomous items (Oztürk, 2019; Sari & Rabon, 2018). Also, in the new MSTGen data generator program developed by Han (2022), there are three options for the routing methods: MFI, matching b-value, and random. The MFI routing method selects the most informative item with the highest accuracy due to its formulation (Luo et al., 2016). Although MFI and random are essential methods that have been frequently used (Svetina et al., 2019), there is no study in which one performs better in polytomous items. The results of this study will provide essential contributions in terms of being a guide to the optimum conditions of real applications that are likely to be applied in the future.

In this study, we researched the answer to the following question presented: "In computerized adaptive multistage tests, in tests consisting of polytomous items (3, 4, and 5 categories), how do the ability estimations of individuals change depending on test length (10, 20, and 30), panel designs (1-2, 1-2- 2, 1-3, 1-3-3, 1-4, and 1-4-4) and routing methods (Maximum Fisher Information [MFI] and Random)?"

Methods

This research is a simulation study, and the aim of the study is to examine the effects of simulation conditions (e.g., test length, number of item categories, panel design, and routing method) on ability estimation under the context of having polytomous items. Within the scope of the research, three categories (3, 4, and 5), three test lengths (10, 20, and 30), six-panel designs (1-2, 1-2-2, 1-3, 1-3-3, 1-4, and 1-4-4) and two routing methods (Maximum Fisher Information [MFI] and Random), 108 (3x3x6x2) conditions were examined. The conditions of the study are shown in Table 1.

Table 1

Simulation Conditions

Condition	Number of Levels	Levels
Number of Category	3	3-category 4-category 5-category
Test Length	3	10 items 20 items 30 items
Panel Design	6	1-2 1-2-2 1-3 1-3-3 1-4 1-4-4
Routing Method	2	MFI Random
Total	3x3x6x2=108	

Sample size (1000), sample ability distribution [N(0,1)], item pool size (200 items), and ability estimation method (Expected a priori-EAP) were kept constant in the study. 100 iterations were run for each condition.

Three separate item pools, each consisting of 200 items in 3, 4, and 5 categories to be used in the research, were generated with the WinGen (Han, 2007) program. Item parameters were produced according to 208 items' descriptive statistics consisting of 3, 4, and 5 categories as Macken Ruiz (2008) used in his dissertation. When generating *a* and *b* parameters under different numbers of item categories (e.g., 3, 4, and 5-category), we used a uniform distribution. The parameter *a* was in the range of [0.68, 1.5] for 3-category items, [0.57, 1.01] for 4-category items, and [0.54, 1] for 5-category items. The *b* parameter was between [-2.77, 3.41] for 3-category items, [-3.01, 3.44] for 4-category items, and [-3.15,

1.68] for 5-category items. With the simulation, 200 polytomous items were produced according to the generalized partial score model (GPCM) (Muraki, 1992). The GPCM formulation is as follows (Embretson & Reise, 2013):

$$P_{ix} = \frac{\exp[\sum_{j=0}^{x} a_i\,(\theta - g_{ij})]}{\sum_{r=0}^{m} exp\,[\sum_{j=0}^{r} a_i\,(\theta - g_{ij})]} \quad (1)$$

where m is the number of categories, x is the student's score on the item, i is the item index, θ is the student's ability, a is discrimination parameter for the item j. Substituting the category information function a simplified equation for polytomous item information is calculated as (Samejima, 1969; Dodd et al., 1995):

$$I_i(\theta_j) = \sum_{x=0}^{m_i} \frac{[P'_{ix}(\theta_j)]^2}{P_{ix}(\theta_j)} \quad (2)$$

Descriptive statistics of item parameters are given in Table 2.

Table 2

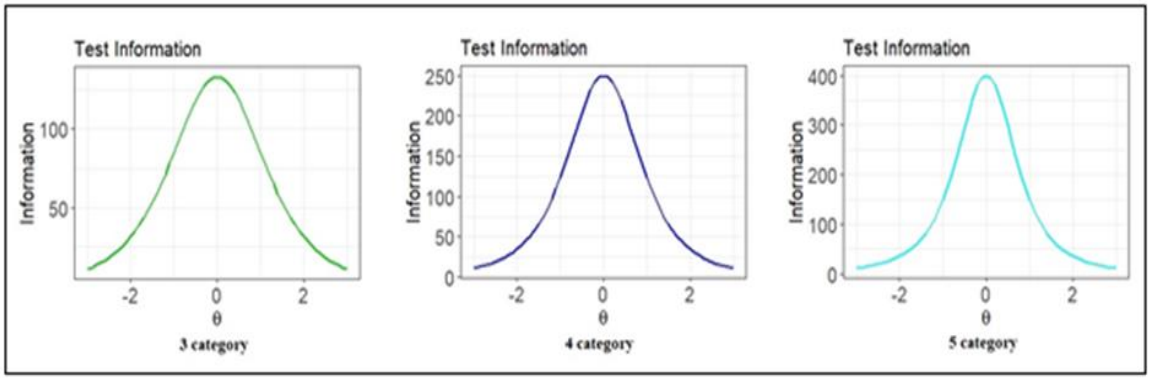
Descriptive Statistics For The Item Parameters Across The Condition

	3-Category			4-Category				5-Category				
Statistics	*a*	b_1	b_2	*a*	b_1	b_2	b_3	*a*	b_1	b_2	b_3	b_4
Min.	0.68	-2.77	-1.63	0.57	-3.01	-1.73	-0.86	0.54	-3.15	-2.31	-1.47	-0.87
Max.	1.50	0.94	3.41	1.01	-0.81	2.35	3.44	1.00	-1.05	1.45	1.68	1.03
Mean	1.09	-0.63	0.49	0.78	0.75	-0.01	0.85	0.78	0.94	-0.28	0.33	3.03

Item information functions were calculated in R program (R Development Core Team, 2018), and modules and panels were built in IBM CPLEX program (ILOG, 2006). Cplex is a mathematical modeling program that solves optimization problems consisting of linear or quadratic equations with the most precise results possible. The Cplex program selected the most appropriate items to be placed in each module from the item pool. Figure 1 shows test information function graphs of three different item pools consisting of 3, 4, and 5-category items.

Figure 1

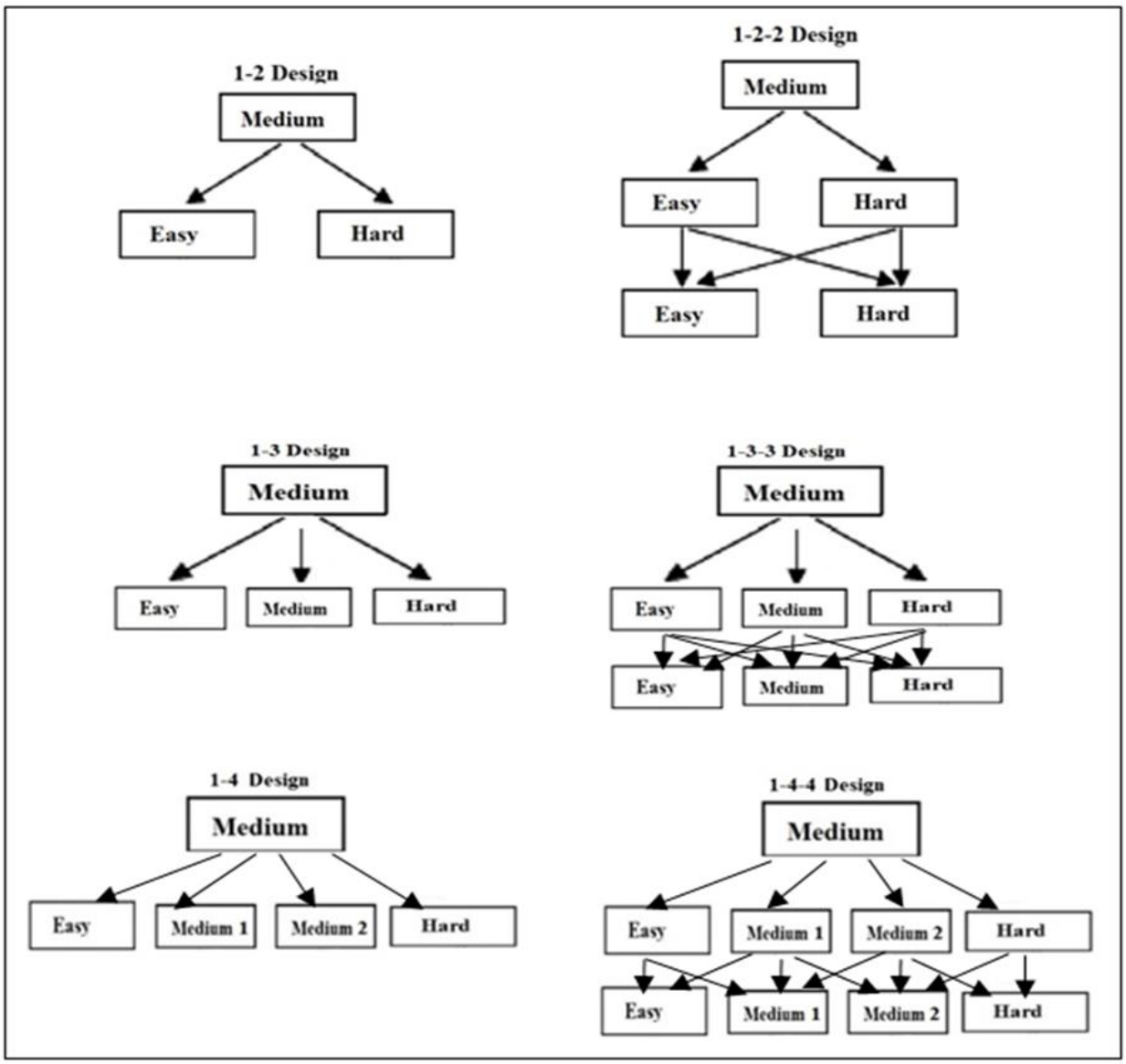
Test Information Functions of Item Pools



The routing module comprises items with medium difficulty levels. For items of medium difficulty, the total item information is maximized at theta level of 0. In 1-2, 1-3, 1-2-2, and 1-3-3 panel designs, easy modules are composed of items with easy difficulty levels meaning that module-level total item information is maximized at the theta level of -1. Lastly, hard modules are composed of items with hard difficulty levels meaning that module-level total item information is maximized at the theta level of +1. In 1-4 and 1-4-4 panel designs, the routing module comprises items with medium difficulty levels, as in the other panel designs. Easy modules are composed of items with easy difficulty levels. Lastly, hard modules are composed of items with hard difficulty levels. As the panel design implies, in 1-4 and 1-4-4 panel designs, there are four modules at different difficulty levels at other stages. These modules are easy, medium-1, medium-2, and hard. For the easy modules, module-level total item information is maximized at the theta level of -1. For the medium -1 module, module-level total item information is maximized at the theta level of -0.33. For the medium-2 modules, module-level total item information is maximized at the theta level of +0.33. For the hard modules, module-level total item information is maximized at the theta level of +1. All panel designs used in the study are shown in Figure 2.

Figure 2

All Panel Designs Used In The Study



As we mentioned above, the sample size is 1000, and there are 108 conditions in this study. The mean absolute bias, RMSE, and correlation values were obtained with a total of 10.800 iterations, 100 iterations for each condition. Four-way ANOVA was run in SPSS for the results. F values and partial η^2 statistics were used to determine the significance of the effects of the factors. Obtained results are given in the findings section.

The research conditions were determined by examining the literature, and taking into account the most frequently used conditions in simulations and real applications (see Rutkowski et al., 2022; Svetina et al., 2019). The studies in the literature related to the conditions in this study are explained in detail in the literature review section. Mean absolute bias (MAB), mean squares of error (RMSE), and correlation values were calculated to evaluate the results. These statistics were calculated from the following formulas.

The bias is the average of the difference between the actual and the predicted value. The bias ($\bar{e}$) is formulated as follows:

$$\bar{e} = \sqrt{\frac{\sum_{J=1}^{N}(\widehat{\theta_J} - \theta_J)}{N}} \tag{3}$$

The mean absolute error (MAE) derives from the unaltered magnitude (absolute value) of each difference.

$$Mean\ Absolute\ Bias = [n^{-1} \sum_{i=1}^{n} |\widehat{\theta_J} - \theta_J|] \tag{4}$$

The RMSE is the mean of the squared difference between the actual and predicted value. The mean squared error is formulated as follows.

$$RMSE = \sqrt{\frac{\sum_{J=1}^{N}(\widehat{\theta_J} - \theta_J)^2}{N}} \tag{5}$$

The correlation between actual (θ) and calculated ($\hat{\theta}$) skill levels ($p(\widehat{\theta_J}, \theta_J)$) is formulated as follows.

$$p(\widehat{\theta_J}, \theta_J) = \frac{cov(\widehat{\theta_J}, \theta_J)}{\sigma_{\widehat{\theta_J}}\, \sigma_{\theta_J}} \tag{6}$$

Results

Overall, when we analyzed the findings in terms of panel design, panel designs 1-2, 1-3, 1-4, 1-2-2, 1-3-3, and 1-4-4 produced very similar results under different conditions. However, the routing method and several item categories changed the study outcomes. Therefore, the study findings regarding routing methods and the number of categories were discussed.

Mean Absolute Bias

Table 3 shows the mean absolute bias values obtained under all simulation conditions. Regardless of panel design, number of categories, and test length, MFI gives better results than random routing

methods. Figure 3 shows the graphs of the mean absolute bias values according to the number of categories. The mean absolute bias decreased as the test length increased. Under the same conditions, as the number of categories changed from 3 to 4, there was a slight increase in the mean absolute bias values. However, MST conditions consisting of 5-category items had the lowest mean absolute bias values. The lowest mean absolute bias is seen in the MFI routing method (.149) in the 5-category, 30-item test, and 1-3-3 panel design. The highest mean absolute bias is seen in the random routing method in the tests in 4-category, 10-item, and 1-2-2 panel designs (.301). The highest score is highlighted in bold, and the lowest score is marked in bold and italic in Table 3.

Table 3

Findings of Average Absolute Bias Across All Conditions

Routing Method	Panel Design	3-Category			4-Category			5-Category		
		10	20	30	10	20	30	10	20	30
MFI	1-2	.261	.195	.164	.277	.207	.174	.242	.179	.150
	1-2-2	.259	.194	.163	.276	.205	.173	.241	.178	.150
	1-3	.260	.194	.165	.278	.207	.176	.240	.177	.150
	1-3-3	.257	.192	.164	.275	.206	.174	.239	.176	***.149***
	1-4	.259	.197	.165	.277	.208	.177	.241	.178	.153
	1-4-4	.256	.192	.165	.278	.207	.178	.237	.178	.151
Random	1-2	.295	.223	.186	.300	.225	.188	.261	.195	.162
	1-2-2	.295	.224	.186	**.301**	.226	.189	.261	.196	.163
	1-3	.292	.221	.186	.299	.224	.190	.260	.194	.162
	1-3-3	.294	.222	.188	.299	.226	.192	.259	.197	.164
	1-4	.293	.225	.189	.299	.225	.191	.262	.197	.165
	1-4-4	.296	.226	.193	.300	.230	.194	.264	.198	.165

Figure 3

Average Absolute Bias Values According to The Number of Categories

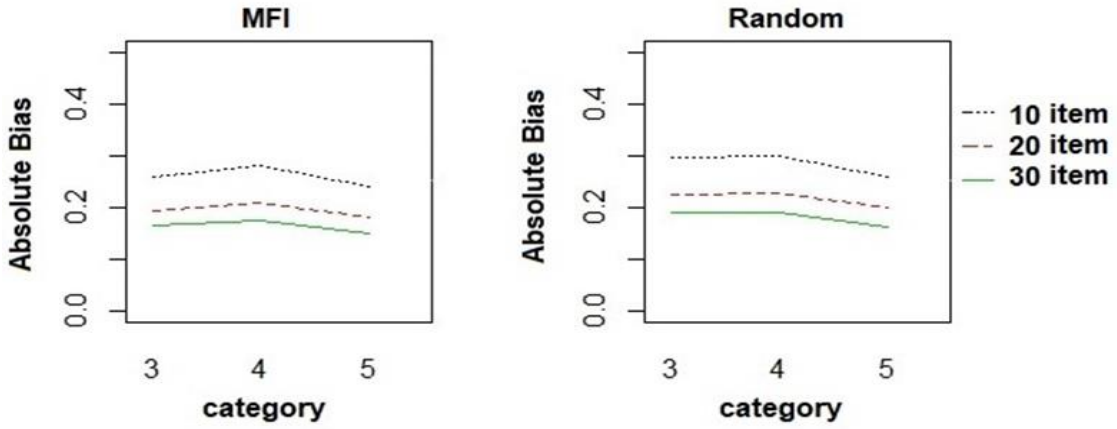


Table 4 shows that ANOVA results for mean absolute bias indicate that most interaction and main effects were significant. Four factors ANOVA was significant (η^2 =.922). However, the factors with the highest partial η^2 were the main effects of routing and test length (η^2 =.927). These effects explained about %93 of the variance in the mean absolute bias. The main effects of test length (η^2=.868) was the factor with the next largest partial η^2. The factor explained about 87% of the variance in the mean absolute bias. When category and panel design were added to routing and test length separately, the factor explained about 83% of the variance in the mean absolute bias (η^2 =.827).

Tablo 4

ANOVA Results for Grand Mean Absolute Bias

Factor	Sum of Squares	df	Mean Square	F	p	η^2_p
Routing	1.401	1	1.401	43657.236	.000	.803
Test Lenght	2.250	2	1.125	35054.378	.000	.868
Category	1.376	2	.688	21435.204	.000	.800
Panel Design	.344	5	.069	2142.806	.000	.501
Routing * Test Lenght	4.328	2	2.164	67414.127	.000	.927
Routing * Category	.937	2	.469	14599.161	.000	.732
Routing * Panel Design	.120	5	.024	746.228	.000	.259
Test Lenght* Category	.252	4	.063	1962.405	.000	.423
Test Lenght* Panel Design	.861	10	.086	2683.301	.000	.715
Category * Panel Design	.593	10	.059	1848.531	.000	.634
Routing*Test Lenght * Category	1.644	4	.411	12806.986	.000	.827
Routing * Test Lenght * Panel design	1.644	10	.164	5122.090	.000	.827
Routing * Category *Panel Design	.968	10	.097	3015.402	.000	.738
Test Lenght* Category * Panel design	1.200	20	.060	1869.315	.000	.778
Routing * Test Lenght * Category * Panel Design	4.083	20	.204	6360.937	.000	.922
Residuals	.343	10692	.000			
Total	528.233	10800				

Root Mean Square Error

Table 5 shows the RMSE values obtained under all research conditions. Figure 4 shows the graphs of RMSE values according to the number of categories. Regardless of panel pattern, number of categories, and test length, MFI gives better RMSE results than the random routing method. As the test length increased, the RMSE value decreased in both routing methods. As the number of categories increased, the RMSE value decreased. The lowest RMSE value is seen in the 5-category, 30-item test, in 1-3-3 panel design, in the MFI routing method (.190). The highest RMSE values are seen in the random routing method (.383) in the 10-item test with 3 and 4 categories. The highest RMSE value is for 4 categories in 1-2-2 panel design. Another highest RMSE value is for 3 categories in 1-4-4 panel design. The highest scores are noted in bold, and the lowest score is noted in bold and italic in Table 5.

Table 5

Findings of RMSE Across All Conditions

Routing Method	Panel Design	3-Category			4-Category			5-Category		
		10	20	30	10	20	30	10	20	30
MFI	1-2	.333	.250	.210	.352	.262	.220	.308	.228	.192
	1-2-2	.330	.247	.208	.350	.260	.219	.307	.227	.191
	1-3	.331	.248	.210	.352	.263	.223	.306	.226	.191
	1-3-3	.327	.244	.209	.350	.261	.220	.304	.224	***.190***
	1-4	.330	.251	.210	.351	.264	.224	.308	.227	.194
	1-4-4	.325	.244	.210	.351	.263	.225	.302	.226	.192
Random	1-2	.381	.292	.244	.382	.288	.241	.335	.252	.210
	1-2-2	.380	.291	.242	**.383**	.289	.243	.336	.252	.210
	1-3	.379	.288	.242	.380	.287	.243	.335	.250	.210
	1-3-3	.378	.288	.244	.380	.288	.245	.333	.253	.211
	1-4	.380	.293	.246	.381	.287	.244	.338	.253	.212
	1-4-4	**.383**	.294	.251	.382	.294	.247	.339	.255	.215

Figure 4

RMSE Values According to Category Numbers

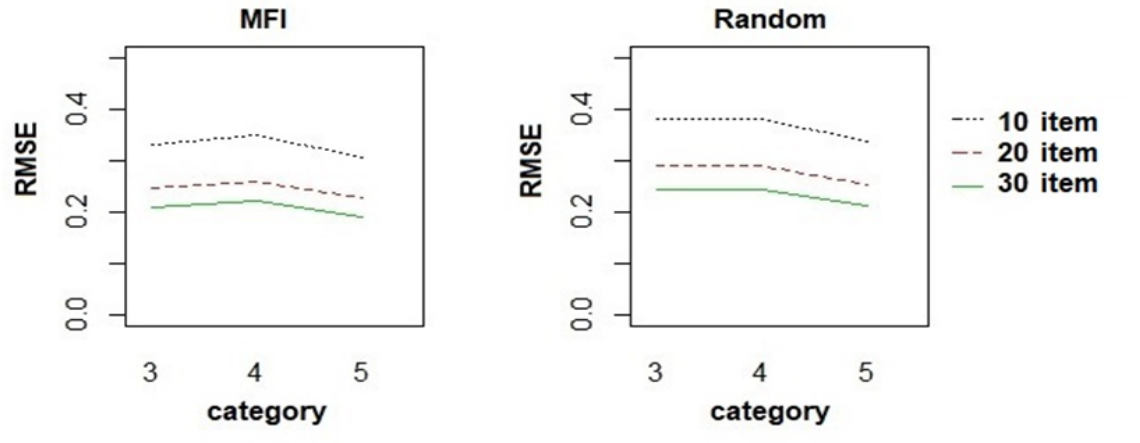


Table 6 shows that ANOVA results for grand mean RMSE indicate that most interaction and main effects were significant. Four factors ANOVA was significant (η^2 =.915). The factors with the highest partial η^2 were the main effects of routing and test length (η^2 =.920). These effects explained 92% of the variance in the mean RMSE. The main effects of test length (η^2=.855) was the factor with the next largest partial η^2. The factor explained about 86% of the variance in the mean RMSE each. Panel design and category added to routing and test length seperatly. The effect was almost the same. When panel design added to routing and test length, the factor explained about 81% of the variance in the mean RMSE (η^2 =.814). When category was added to routing and test length, the factor explained about 81% of the variance in the mean RMSE (η^2 =.810).

Tablo 6

ANOVA Results for Grand Mean RMSE

Factor	Sum of Squares	df	Mean Square	F	p	η^2_p
Routing	2.701	1	2.701	47053.219	.000	.815
Test Lenght	3.630	2	1.815	31625.578	.000	.855
Category	2.002	2	1.001	17440.144	.000	.765
Panel Design	.590	5	.118	2057.510	.000	.490
Routing * Test Lenght	7.088	2	3.544	61744.829	.000	.920
Routing * Category	1.499	2	.750	13060.690	.000	.710
Routing * Panel Design	.196	5	.039	681.434	.000	.242
Test Lenght* Category	.447	4	.112	1945.753	.000	.421
Test Lenght* Panel Design	1.420	10	.142	2474.563	.000	.698
Category * Panel Design	.983	10	.098	1713.316	.000	.616
Routing*Test Lenght * Category	2.620	4	.655	11412.098	.000	.810
Routing * Test Lenght * Panel design	2.685	10	.268	4677.228	.000	.814
Routing * Category *Panel Design	1.614	10	.161	2812.061	.000	.725
Test Lenght* Category * Panel design	1.918	20	.096	1670.814	.000	.758
Routing * Test Lenght * Category * Panel Design	6.645	20	.332	5788.613	.000	.915
Residuals	.614	10692	.000			
Total	864.884	10800				

Correlation

Table 7 shows the correlation values obtained under all research conditions. Figure 5 shows the graphs of correlation values according to the number of categories. Although the MFI routing method generally gives better results than the random routing method, it gives the same results in the 5-category, 20- and 30-item tests. Figure 5 shows the graphs of correlation values according to the number of categories. As the test length increased, the correlation value increased in both routing methods. Similarly, as the number of categories increased, the correlation value increased relatively. The lowest correlation value was found in 1-2, 1-2-2, and 1-4-4 panel designs, in the 4 categories, 10-item test, and in the 1-4-4 panel designs in the 3 category 10-item test and in the random routing method (.923). The highest correlation value was found in all panel designs except 1-4 in the 5-category 30-item test and in MFI (.981). The highest scores are highlighted in bold, and the lowest score is highlighted in bold and italic in Table 7.

Tablo 7

Findings of Correlations Across All Conditions

Routing Method	Panel Design	3-Category			4-Category			5-Category		
		10	20	30	10	20	30	10	20	30
MFI	1-2	.942	.968	.977	.935	.964	.975	.951	.973	**.981**
	1-2-2	.943	.969	.979	.936	.965	.975	.951	.973	**.981**
	1-3	.943	.968	.977	.935	.964	.974	.951	.974	**.981**
	1-3-3	.944	.969	.979	.936	.965	**.975**	.952	.974	**.981**
	1-4	.943	.967	.977	.936	.964	.974	.951	.973	.980
	1-4-4	.945	.970	.979	.936	.964	.974	.953	.973	**.981**
Random	1-2	.924	.956	.969	***.923***	.957	.970	.941	.967	.977
	1-2-2	.924	.956	.970	***.923***	.957	.970	.941	.967	.977
	1-3	.925	.957	.970	.924	.957	.969	.942	.968	.977
	1-3-3	.925	.957	.969	.924	.957	.969	.942	.967	.977
	1-4	.924	.955	.969	.924	.957	.969	.941	.967	.977
	1-4-4	***.923***	.955	.967	***.923***	.955	.968	.940	.966	.976

Figure 5

Correlation Values According to Category Numbers

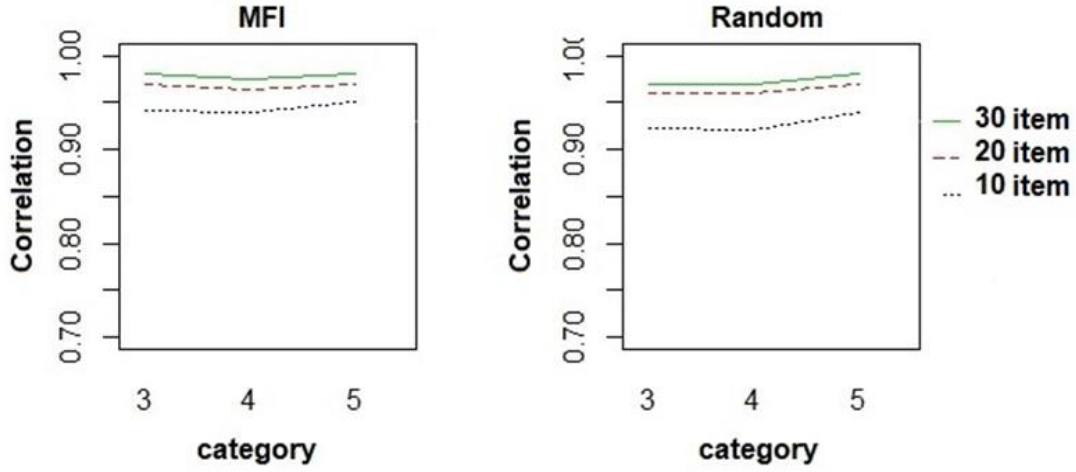


Table 8 shows that ANOVA results for correlation indicate that most interaction and main effects were significant. Four factors ANOVA was significant (η^2 =.893). The factors with the highest partial η^2 were the main effects of routing and test length (η^2 =.898). These effects explained 90% of the variance in correlation. The main effects of test length (η^2=.839) was the factor with the next largest partial η^2. The factor explained about 84% of the variance in the correlation each. When panel design added to routing and test length, the factor explained about 78% of the variance in the correlation (η^2 =.785). When category added to routing and test length, the factor explained about 77% of the variance in the correlation (η^2 =.771).

Tablo 8

ANOVA Results for correlation

Factor	Sum of Squares	df	Mean Square	F	p	η^2_p
Routing	.262	1	.262	38768.600	.000	.784
Test Lenght	.375	2	.188	27778.873	.000	.839
Category	.166	2	.083	12288.887	.000	.697
Panel Design	.057	5	.011	1699.072	.000	.443
Routing * Test Lenght	.634	2	.317	46923.297	.000	.898
Routing * Category	.152	2	.076	11248.700	.000	.678
Routing * Panel Design	.018	5	.004	535.626	.000	.200
Test Lenght* Category	.051	4	.013	1900.051	.000	.415
Test Lenght* Panel Design	.130	10	.013	1928.010	.000	.643
Category * Panel Design	.092	10	.009	1358.231	.000	.560
Routing*Test Lenght * Category	.243	4	.061	9008.331	.000	.771
Routing * Test Lenght * Panel design	.264	10	.026	3902.227	.000	.785
Routing * Category *Panel Design	.141	10	.014	2081.990	.000	.661
Test Lenght* Category * Panel design	.190	20	.009	1405.130	.000	.724
Routing * Test Lenght * Category * Panel Design	.604	20	.030	4473.880	.000	.893
Residuals	.072	10692	.000			
Total	9937.363	10800				

Discussion

Overall, this study investigated the change in the ability estimations of individuals in tests consisting of polytomous items in the computerized multistage test (MST) environment according to the routing methods based on three categories (3, 4, and 5), six-panel designs (1-2, 1-3, 1-4, 1-2-2, 1-3-3, and 1-4-4), three test lengths (10, 20, and 30-item) and two routing methods (MFI and random). The results were then analyzed for mean absolute bias, mean squares of error (RMSE), and correlation values between actual and observed ability levels.

When examining the average absolute bias, RMSE, and correlation values obtained from the item pools consisting of 3, 4, and 5 category items in terms of item categories, the values obtained from the 3 and 4-category item pools are close. Still, the mean absolute bias obtained from the item pool consisting of 4 category items (.23) and RMSE (.29) is the highest. However, the mean absolute bias (.19) and RMSE (.25) values obtained from the item pool consisting of 5-category items are lower than the other categories. In addition, the correlation value (.97) is at the highest level in 5-category items compared to other categories. According to the results obtained, as the number of categories increases, mean absolute bias and RMSE decrease, while correlation values increase.

When examined in terms of routing methods, MFI and random routing methods have similar tendencies, but MFI delivers better results. This was consistent with the previous studies. For example, Macken-Ruiz (2008) compared three routing methods with generalized partial credit model item response theory:

MI, fixed θ, and number-right routing in MST environment, and found that the best performance was observed under the maximum information routing. This was because MFI is a dynamic routing method that calculates module-level information first and, selects the best appropriate module for a test taker. However, the random routing approach, a kind of static method, does not use such an adaptation, and randomly selects the next module among the available modules. This might result in that a test taker with high ability level can receive an easier module at the next level which would inflate his/her ability estimation. Therefore, MFI yielded better results, as also found in Svetina et al. (2019).

Kim et al. (2013) observed that the accuracy of the ability estimates increased as the test length increased. Similarly, in our study mean absolute bias decreased as the test length increased. In addition, as the test length increased, the correlation values also increased. Oztürk (2019) examined how the length and feature of the routing module affect the measurement accuracy in various panel designs. In that study, with two-category items, correlation values increased as the test length increased. As the test length increased, the RMSE value decreased in both routing methods. It can be seen that when examined in terms of test length, the results obtained in our current study show similarities with studies conducted with dichotomous items in the literature (Oztürk, 2019).

Our current study examined an item pool consisting of polytomous items and different conditions, all examined panel designs (1-2, 1-2-2, 1-3, 1-3-3, 1-4, and 1-4-4) showed similar results. Kim et al. (2013) determined all routing methods classification decisions equally well in their studies where they utilized an item pool consisting of polytomous items based on partial credit model (PCM), different panel designs (1-3-3, 1-3-2, 1-2-3, and 1-2-2) and routing methods (ML- DPI, SL-DPI, and M-AMI). Zenisky (2004) did not find meaningful differences between these panel structures or routing methods. The precision of their classification decision was performed all the same. However, some studies have dichotomous items, where the mean error value decreases as we move from the two-stage panel design to the three-stage panel pattern (Sari & Raborn, 2018). Therefore, while the panel design used in MST applications in which an item pool consisting of polytomous items is used does not matter, choosing three-stage panel designs in dichotomous MST applications will provide more accurate results. However, as in the case of Sari and Raborn (2018) and Zenisky (2004), the chosen routing method severely affects the accuracy of the results.

Our study is limited to three kinds of polytomous items (3, 4, and 5 categories), six-panel designs (1-2, 1-3, 1-4, 1-2-2, 1-3-3, and 1-4-4), three test lengths (10, 20, and 30) and two routing methods (MFI and random). According to this study's results, better values were obtained as the number of categories increased. Considering the number of categories in future studies, 5-category items should be preferred. Since there is no difference between the panel designs in the current study, different applications can be made by choosing the panel design suitable for the item pool in future studies. Applications based on actual study parameters can be made with different routing methods (MLWMI, MPWMI, MKL, and MKLP). Classification precision can be examined using different test lengths and item category numbers in MST.

Declarations

Author Contribution: Hasibe Yahsi Sari-Conceptualization, methodology, analysis, writing & editing, visualization. Hülya Kelecioglu-Conceptualization, methodology, writing-review & editing, supervision.

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

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

References

Chen, L-Y. (2010). *An investigation of the optimal test design for multi-stage test using the generalized partial credit model.* [Doctoral dissertation, The University of Texas]. UT Electronic Theses and Dissertations. https://repositories.lib.utexas.edu/handle/2152/ETD-UT-2010-12-344

Choi, Y. J., & Asilkalkan, A. (2019). R packages for item response theory analysis: Descriptions and features. *Measurement: Interdisciplinary Research and Perspectives*, *17*(3), 168-175. https://doi.org/10.1080/15366367.2019.1586404

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19*(1), 5-22. https://doi.org/10.1177/014662169501900103

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement*, *31*(4), 295-311. https://doi.org/10.1111/j.1745-3984.1994.tb00448.x

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Han, K. C. T. (2022). User's Manual: MSTGen. Retrieved from https://www.umass.edu/remp/software/simcata/mstgen/MSTGen_Manual.pdf

Han, K. T. (2007). WinGen: Windows software that generates item response theory parameters and item responses. *Applied Psychological Measurement*, *31*(5), 457-459. https://doi.org/10.1177/0146621607299271

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, *26*(2), 44-52. **https://doi.org/10.1111/j.1745-3992.2007.00093.x**

ILOG. (2006). ILOG CPLEX 10.0 [User's manual]. Paris, France: ILOG S.A. Retrieved from https://www.lix.polytechnique.fr/~liberti/teaching/xct/cplex/usrcplex.pdf

Kim, J., Chung, H., & Dodd, B. G. (2010, May). *Comparing routing methods in the multistage test based on the partial credit model* [Conference presentation]. In AERA, Denver, CO.

Kim, J., Chung, H., Park, R., & Dodd, B. G. (2013). A comparison of panel designs with routing methods in the multistage test with the partial credit model. *Behavior Research Methods, 45,* 1087–1098. https://doi.org/10.3758/s13428-013-0316-3

Luecht, R. M. (2000, April). *Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high quality computer-adaptive and mastery tests*. [Conference presentation]. In NCME, New Orleans, LA. https://eric.ed.gov/?id=ED442823

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*(3), 229-249. https://doi.org/10.1111/j.1745-3984.1998.tb00537.x

Luo, F., Ding, S., Wang, X., & Xiong, J. (2016). Application study on online multistage intelligent adaptive testing for cognitive diagnosis. *Quantitative Psychology Research*, 265-275. https://doi:10.1007/978-3-319-38759-8_20

MacGregor, D., Yen, S. J., & Yu, X. (2022). Using multistage testing to enhance measurement of an english language proficiency test. *Language Assessment Quarterly*, *19*(1), 54-75. https://doi.org/10.1080/15434303.2021.1988953

Macken-Ruiz, C. L. (2008). *A comparison of multi-stage and computerized adaptive tests based on the generalized partial credit model* [Doctoral dissertation, The University of Texas]. ProQuest Dissertations Publishing. https://www.proquest.com/docview/304482829?pq-origsite=gscholar&fromopenview=true

Magis, D., Yan, D., von Davier, A., & Magis, M. D. (2018). Package 'mstR'. Retrieved from https://cran.r-project.org/web/packages/mstR/mstR.pdf

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

Oztürk, N. B. (2019). How the Length and Characteristics of Routing Module Affect Ability Estimation in ca-MST?. *Universal Journal of Educational Research*, *7*(1), 164-170. https://doi.org/10.13189/ujer.2019.070121

R Core Team. (2018). R: A language and environment for statistical computing: R foundation for statistical computing.

Ridho, A. (2022, January). Sociocultural Literacy Assessment: Validation of Multistage Generalized Partial Credit Testing Design. In *International Conference on Madrasah Reform 2021 (ICMR 2021)* (pp. 382-386). Atlantis Press. https://doi.org/10.2991/assehr.k.220104.056

Rutkowski, L., Liaw, Y. L., Svetina, D., & Rutkowski, D. (2022). Multistage testing in heterogeneous populations: Some design and implementation considerations. *Applied Psychological Measurement*, *46*(6), 494-508. https://doi.org/10.1177/01466216221108123

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34* (17). https://psycnet.apa.org/record/1972-04809-001

Sari, H. I., & Raborn, A. (2018). What information works best?: A comparison of routing methods. *Applied Psychological Measurement*, *42*(6), 499-515. https://doi.org/10.1177/0146621617752990

Sari, H.I., Yahsi Sarı, H., & Huggins Manley, A.C. (2016). Computer adaptive multistage testing: Practical issues, challenges and principles. *Journal of Measurement and Evaluation in Education and Psychology*, *7*(2), 388-406. https://doi.org/10.21031/epod.280183

Svetina, D., Liaw, Y. L., Rutkowski, L., & Rutkowski, D. (2019). Routing strategies and optimizing design for multistage testing in ınternational large-scale assessments. *Journal of Educational Measurement*, *56*(1), 192-213. https://doi.org/10.1111/jedm.12206

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*(4), 473-492. https://doi.org/10.1177/014662168200600408

Weiss, D. J. (1983). Latent trait theory and adaptive testing. In Weiss D. J. (Ed.), *New horizons in testing* (pp. 5-7). Academic Press.

Weissman, A., Belov, D. I., Armstrong, R. D. (2007). Information-based versus number-correct routing in multistage classification tests. *LSAC Research Report Series*. No. 07-05. Law School Admission Council. https://www.researchgate.net/publication/237288650_Information-Based_Versus_Number-Correct_Routing_in_Multistage_Classification_Tests

Zenisky A., Hambleton R.K.,& Luecht R.M. (2009) Multistage testing: Issues, designs, and research. In: van der Linden W., Glas C. (eds) *Elements of Adaptive Testing.* Springer.

Zenisky, A. L. (2004). *Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment* (Publication No. 5710) [Doctoral dissertation, University of Massachusetts Amherst]. UMass Amherst Libraries. https://scholarworks.umass.edu/dissertations_1/5710

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(3); 185-199

Investigation of The Measurement Invariance of Affective Characteristics Related to TIMSS 2019 Mathematics Achievement by Gender*

Mehmet ATILGAN** Kaan Zülfikar DENİZ***

Abstract

This research examines whether the affective characteristics of the TIMSS 2019 Turkey mathematics application provide measurement invariance according to gender. The research sample consists of 4048 8th-grade students participating in the TIMSS in 2019. Research data were downloaded from the international website of TIMSS. The research data collection tools are "Sense of School Belonging", "Students Confident in Mathematics", "Students Like Learning Mathematics", and "Students Value Mathematics" scales. Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) were performed in the context of validity analyses to examine measurement invariance. In terms of reliability, the Cronbach Alfa internal consistency coefficient was calculated. Accordingly, out of the four scales in the study, only "Students Confident in Mathematics" scale could not be confirmed in confirmatory factor analysis. Therefore, while "Students Confident in Mathematics" scale was not examined for measurement invariance, the other three scales were examined within the scope of measurement invariance. For measurement invariance, research data were tested with Multiple Group Confirmatory Factor Analysis (MG-CFA), one of the Structural Equation Modeling (SEM) techniques. As a result of the analyses, while the strict invariance model was provided in "Students Like Learning Mathematics" scale and "Students Value Mathematics" scale, strong invariance/scale invariance model was provided in "Sense of School Belonging" scale. It was concluded that there was no gender bias in the three scales for which MG-CFA was performed, and the mean scores were comparable according to gender. In this context, it can be said that "Sense of School Belonging", "Students Like Learning Mathematics", and "Students Value Mathematics" scales are valid in determining the differences according to gender.

Keywords: TIMSS, affective variables, measurement invariance, MG-CFA, SEM

Introduction

Raising qualified people is one of the most critical issues for countries. Education systems play a significant role in raising qualified people. States change their education policies over time and make arrangements in their education systems to train qualified people with the desired characteristics.

In Turkey, regulations have been made in the education system over time. These arrangements are made through the findings obtained from the national and international measurement and evaluation practices in which Türkiye has participated. Türkiye has been participating in international educational studies such as TIMSS (Trends in International Mathematics and Science Study), PISA (Programme for International Student Assessment), and PIRLS (Progress in International Reading Literacy Study) for many years. Türkiye participates in these studies to compare the education system of Türkiye with the education systems of others, to reveal the situation of Türkiye on an international scale, to eliminate the deficiencies in the education system based on the findings of these studies, and to make adjustments in education policies.

*This study has been produced from Master's Thesis that was conducted under the supervision of the Prof. Dr. Kaan Zülfikar Deniz and prepared by Mehmet ATILGAN.

**Research Assistant, Uşak University, Faculty of Education, Uşak-Türkiye, mehmet.atilgan@usak.edu.tr, ORCID ID: 0000-0003-1297-4630
***Prof. Dr., Ankara University, Faculty of Education, Ankara-Türkiye, zlfkrdnz@yahoo.com, ORCID ID: 0000-0003-0920-538X

To cite this article:
Atılgan, M. & Deniz, K. Z. (2023). Investigation of the measurement invariance of affective characteristics related to TIMSS 2019 mathematics achievement by gender. *Journal of Measurement and Evaluation in Education and Psychology, 14*(3), 185-199. https://doi.org/10.21031/epod.1221365

Received: 19.12.2022
Accepted: 7.09.2023

One of the most numerous international education studies that Türkiye has participated in is the TIMSS, organized by the IEA (International Association for the Evaluation of Educational Achievement). TIMSS was first implemented in 1995. Türkiye attended TIMSS for the first time in 1999 and, finally, in 2019 (Ministry of National Education-MoNE, 2020).

TIMSS is an educational study aiming to evaluate the knowledge and skills of 4th and 8th-grade students in mathematics and science. Since many variables affect students' success, detailed data about students, teachers, schools, and parents are collected through questionnaires within the scope of TIMSS (MoNE, 2016). Data on determining affective characteristics such as motivation, interest, and attitude are collected through TIMSS student questionnaires (Mullis et al., 2016).

Bloom (2012) states that affective characteristics have a 25% effect on students' academic success. In the literature, there are also studies showing that affective characteristics affect mathematics achievement (Doğan & Barış, 2010; İlhan & Öner-Sünkür, 2012; Kesici, 2018; Kesici & Aşılıoğlu, 2017; Lay et al., 2015; Mohammadpour, 2012; Ölçüoğlu & Çetin, 2016; Sarı & Ekici, 2018; Sarıer, 2020; Yücel & Koç, 2011).

Demographic variables are influential on academic achievement. Studies in the literature aim to determine at what level demographic variables such as gender, socioeconomic level, age, class, and geographical region affect success. Among these variables, studies on the gender variable attract attention. The number of studies comparing the mean scores of gender groups to examine the effect of TIMSS mathematics achievement is relatively high (Aydın, 2015; Hanci, 2015; Kilic & Askin, 2013; Louis & Mistele, 2012; Patterson et al., 2003; Sarıer, 2020; Wang et al., 2012; Webster & Fisher, 2000).

In the 8th-grade Türkiye sample of TIMSS 2019, the mean mathematics scores for male and female groups are 490 and 501, respectively. However, the difference in scores between the averages was not statistically significant (MoNE, 2020). The measurement results obtained regarding the comparison of the groups may vary depending on different characteristics of the individuals. However, the source of the differences may only sometimes be individuals. The reason can sometimes be the measurement tool itself. When comparing the measurements according to the groups, it is assumed that the measurement tool measures the same feature for all groups. In other words, measurement invariance is ensured (Başusta & Gelbal, 2015). However, it is crucial to prove that measurement invariance is ensured to conduct comparison studies with groups more validly and reliably.

Measurement invariance is defined as the same perception and interpretation of the items in the measurement tool in all groups subject to measurement (Byrne & Watkins, 2003). In the scales developed to reveal a latent structure, measurement invariance appears as one of the psychometric properties (Öncü, 2019). The measurement tool should measure the same structure in the groups to ensure measurement invariance. Factor loadings, correlations between factors, and error variances of the scale items should be equal to measure the same structure in groups (Byrn et al., 1989). There is a consensus in the literature that to compare mean scores by groups, measurement invariance should be tested, evidence of strong/scalar invariance model should be obtained, and comparison of mean scores without these conditions may not yield significant results (Başusta & Gelbal, 2015; Cheung & Rensvold, 2000; Gregorich, 2006; Kıbrıslıoğlu, 2015; Öğretmen, 2006; Salzberger et al., 1999; Vandenberg & Lance, 2000; Wicherts, 2007; Wu et al., 2007). The purpose of statistical analyses to test measurement invariance is to determine whether the established structural model is the same in subgroups and which of the parameters included in the structural model are invariant (Mulaik, 2007).

While the methods in Structural Equation Modeling (SEM) and Item Response Theory (IRT) approaches are primarily preferred in determining measurement invariance, methods based on the Latent Class Analysis (LCA) approach have also been used in recent years (Yandı et al., 2017). 80% of measurement invariance studies are conducted with approaches based on SEM (Vandenberg & Lance,2000). The MG-CFA (Multiple Group Confirmatory Factor Analysis) method is most frequently used in SEM-based approaches. Measurement invariance can be tested by examining the equality of mean covariance structures with the MG-CFA method (Yandı et al., 2017).

In TIMSS and PISA literature, measurement invariance was examined according to countries in some studies (Ercikan & Koh, 2005; Karakoc-Alatli et al., 2016; Öncü, 2019; Rutkowski & Rutkowski, 2013;

Tavlıca, 2019; Wu et al., 2007; Ma & Qin, 2021; Meng et al., 2019; Polat, 2019; Scherer et al., 2016). The subjects of these studies are mathematics and science achievement, socioeconomic level, affective variables related to mathematics and science, and using information and communication technologies. Some studies also examined measurement invariance according to gender (Ertürk & Erdinç-Akan, 2018b; Polat, 2019), geographical regions (Ölçüoğlu & Çetin, 2016; Polat, 2019), and household resources (Cakici-Eser, 2021). The subjects of these studies are affective variables related to mathematics and science, home environment, and school environment.

There are quite a lot of studies (Aydın, 2015; Kilic & Askin, 2013; Louis & Mistele, 2012; Patterson et al., 2003; Sarıer, 2020; Wang et al., 2012; Webster & Fisher, 2000) comparing gender groups on TIMSS mathematics achievement without testing measurement invariance. The literature states that measurement invariance must first be ensured. If it is not ensured, the comparisons may not yield meaningful results (Başusta & Gelbal, 2015; Cheung & Rensvold, 2000; Gregorich, 2006; Kıbrıslıoğlu, 2015; Salzberger et al., 1999; Vandenberg & Lance, 2000; Wicherts, 2007; Wu et al., 2007). For this reason, it is vital to test the measurement invariance before examining the effect of gender on mathematics achievement. Thanks to measurement invariance analysis, the way of interpreting the items of the subgroups can be determined, and it can be tested whether there is a bias of the subgroups in the scale's items (Byrne, 1998; Gregorich, 2006; Kıbrıslıoğlu, 2015; Millsap & Olivera-Aguilar, 2012). Failure to provide measurement invariance indicates that some items in the scale are biased. Some studies conducted according to gender have provided measurement invariance (Bofah & Hannula, 2015; Demir, 2020; Demir, 2017; Gungor & Atalay-Kabasakal, 2020; Jung, 2019; Polat, 2019; Uyar, 2021). However, measurement invariance cannot be achieved in some (Ertürk & Erdinç-Akan, 2018b; Gülleroğlu, 2017; Uzun & Öğretmen, 2010). Since some studies point to gender bias, this study aims to test the measurement invariance of affective characteristics related to TIMSS 2019 mathematics achievement according to gender groups. Measurement invariance studies, which indicate biases according to gender, show that the degree of accuracy of decisions taken about individuals may be inadequate (Öğretmen, 2006).

In examining the relationships between TIMSS student questionnaires and mathematics achievement, "Sense of School Belonging" (Akyüz & Pala, 2010; Akyüz & Satıcı, 2013; Işlak, 2020; Koç, 2019; Sarı et al., 2017; Sarıer, 2020), "Students Confident in Mathematics" (Akyüz-Aru, 2020; Akyüz & Pala, 2010; Atar, 2011; Aydın, 2015; Demir et al., 2010; Ertürk & Erdinç-Akan, 2018a; Işlak, 2020; Khine et al., 2015; Koç, 2019; Oral & McGivney, 2013; Usta & Demirtaşlı, 2018), "Students Like Learning Mathematics" (Erşan, 2016; Ertürk & Erdinç-Akan, 2018a; Khine et al., 2015; Koç, 2019; Oral & McGivney, 2013), and "Students Value Mathematics" (Doğan & Barış, 2010; Khine et al., 2015) scales have been used by some researchers. These affective variables are frequently used in the TIMSS literature.

In the literature, although many studies examine the effect of "Sense of School Belonging" scale on mathematics achievement (Akyüz & Pala, 2010; Akyüz ve Satıcı, 2013; Işlak, 2020; Koç, 2019; Sarı et al., 2017; Sarıer, 2020), no measurement invariance research has been found. While measurement invariance was confirmed in some studies testing the measurement invariance of "Students Confident in Mathematics" scale (Bofah & Hannula, 2015; Cakici-Eser, 2021; Polat, 2019; Uyar, 2021), measurement invariance could not be achieved in some (Ertürk & Erdinç-Akan, 2018b). In the studies that test the measurement invariance of "Students Like Learning Mathematics" (Bofah & Hannula,2015; Cakici-Eser, 2021; Ertürk & Erdinç-Akan, 2018b; Polat, 2019; Shukla & Konold, 2014) and "Students Value Mathematics" (Bofah & Hannula,2015; Polat, 2019; Uyar, 2021) scales, measurement invariance was achieved, and no research was found in which measurement invariance could not be achieved.

Accordingly, this study aimed to examine the measurement invariance of "Sense of School Belonging", "Students Confident in Mathematics", "Students Like Learning Mathematics", and "Students Value Mathematics" scales of TIMSS 2019 Turkey 8th-grade in the context of gender.

Method

This research, which aims to examine the measurement invariance of the affective characteristics of the students in the Turkish sample who participated in the TIMSS 2019 mathematics according to gender, is descriptive. Studies that aim to reveal a situation without intervening are a type of descriptive research (Fraenkel & Wallen, 2006; Karasar, 2011).

Participants

TIMSS 2019 was held with the participation of 4077 eighth-grade students from 181 schools in Turkey. However, it was determined that 29 of these students left all the scales in the student questionnaire blank. For this reason, 29 students were excluded from the analysis, and the participant group consisted of 4048 students. The descriptive statistics of the participant group are shown in Table 1.

Table 1
Descriptive Statistics of Participant Group

Gender	*f*	%	Age average
Female	2009	49.63	13.89
Male	2039	50.37	13.92

Data Collection Tools

The data obtained from "Sense of School Belonging", "Students Confident in Mathematics", "Students Like Learning Mathematics", and "Students Value Mathematics" scales in the TIMSS 2019 mathematics student questionnaire were used in this research. Data were downloaded from the TIMSS research international website (https://timss2019.org/international-database/).

"Sense of School Belonging" scale consists of 5 items and a single factor, with a 4-point Likert-type rating. The items are scored as "1= disagree a lot", "2= disagree a little", "3= agree a little", and "4= agree a lot". There is no reverse-coded item. Higher scores on the scale indicate a higher sense of belonging to the school. Regarding the validity of the TIMSS Turkey sample data set, item factor loadings for this scale ranged from 0.58 to 0.77, and the total explained variance rate was 51%. Regarding reliability, the Cronbach Alpha coefficient was 0.76 (Yin & Fishbein, 2020).

"Students Confident in Mathematics", "Students Like Learning Mathematics", and "Students Value Mathematics" scales consist of 9 items and a single factor, with a 4-point Likert-type rating. Items 2, 3, 5, 8, and 9 were reverse-coded for "Students Confident in Mathematics" scale, and items 2 and 3 were reverse-coded for "Students Like Learning Mathematics" scale. There is no reverse-coded item for "Students Value Mathematics" scale. Higher scores on these scales indicate higher self-confidence, liking, and value in mathematics. In the TIMSS Turkey sample data set, item factor loadings ranged from 0.62 to 0.80 for "Students Confident in Mathematics", 0.61 to 0.89 for "Students Like Learning Mathematics", and 0.58 to 0.81 for "Students Value Mathematics". The total explained variance was 54% for "Students Confident in Mathematics", 62% for "Students Like Learning Mathematics", and 51% for "Students Value Mathematics". The Cronbach Alpha coefficient was 0.89 for "Students Confident in Mathematics", 0.89 for "Students Like Learning Mathematics", and 0.88 for "Students Value Mathematics" (Yin & Fishbein, 2020).

Data Analysis

Data analysis was carried out in three stages. In the first stage, the processes of examining the missing data, extreme values, and normality were followed, sequentially. In the second stage, EFA (Exploratory Factor Analysis) and CFA (Confirmatory Factor Analysis) were performed to create affective trait models associated with mathematics achievement. Cronbach Alpha internal consistency coefficients for each affective trait model created for reliability were calculated. In the last stage, MG-CFA (Multiple

Group Confirmatory Factor Analysis) was performed to determine the measurement invariance according to gender groups in the validated models.

SPSS IBM 20.0 and R Studio were used to analyze the data. For MG-CFA, semTools (Jorgensen et al., 2021) and lavaan (Rosseel, 2012) packages were used.

Results

Before testing the measurement invariance for each scale, missing data, extreme values, and normality were examined in terms of the suitability of the data for analysis. As a result of missing data analysis, 47 participants for "Students Value Mathematics", 90 for "Students Like Learning Mathematics", 52 for "Students Confident in Mathematics", and 16 for "Sense of School Belonging" scales were excluded from the analysis. The extreme value analysis converted each scale's items into Z scores. No extreme values were found except -4 and +4 (Çokluk et al., 2016; Harrington, 2009; Pituch & Stevens, 2016; Tabachnick & Fidell, 2013). In the normality examination, the skewness and kurtosis values were calculated separately for all the items in each scale. It has been determined that all related scales have skewness and kurtosis values except -1 and +1 (Çokluk et al., 2016; Hair et al., 2014; Harrington, 2009; Raykov & Marcoulides, 2006). For this reason, the relevant scales did not show a normal distribution.
In the second stage, the validity and reliability of the scores collected from scales were discussed. According to the EFA, the item factor loadings for "Students Value Mathematics" ranged from 0.58 to 0.81, for "Students Like Learning Mathematics" 0.60 to 0.89, for "Students Confident in Mathematics" 0.62 to 0.79, and for "Sense of School Belonging" 0.57 to 0.77. The total explained variance was 50.39% for "Students Value Mathematics", 61.19% for "Students Like Learning Mathematics", 52.45% for "Students Confident in Mathematics", and 50.49% for "Sense of School Belonging". The DWLS (Diagonally Weighted Least Squares) method is used when the number of categories in the scoring of the items in the Likert-type graded scales at the ranking level is less than five, and the multivariate normality requirement cannot be met in the data set (Kline, 2015; Mindrila, 2010; Schumacker & Beyerlein, 2000). Therefore, the DWLS method was preferred as the estimation method in CFA. $\chi 2$, CFI, TLI, RMSEA and SRMR goodness-of-fit indices are used in this study to evaluate CFA results. The criterion values are presented in Table 2 (Çokluk et al., 2016; Harrington, 2009; Kline, 2015; Tabachnick & Fidell, 2013).

Table 2

Criterion Values in Goodness of Fit Indices

Fit Index		Good Fit			Acceptable Fit	
χ^2		$p>$	.05		$p>$	.05
CFI	0.95 ≤	CFI	≤ 1.00	0.90 ≤	CFI	≤ 0.95
TLI	0.95 ≤	TLI	≤ 1.00	0.90 ≤	TLI	≤ 0.95
RMSEA	0 ≤	RMSEA	≤ 0.05	0.05 ≤	RMSEA	≤ 0.08
SRMR	0 ≤	SRMR	≤ 0.05	0.05 ≤	SRMR	≤ 0.08

CFA was conducted for "Students Value Mathematics" scale with two modifications. Covariance was established between M3 and M4 items and M1 and M2 items with the recommendation of the R program. CFA was conducted for "Students Like Learning Mathematics" scale with a modification. Covariance was established between M2 and M3 items with the recommendation of the R program. Without modifications, CFA was conducted for "Students Confident in Mathematics" and "Sense of School Belonging" scales. After all these procedures, the CFA results for the four scales are presented in Table 3.

Table 3
CFA Results of Affective Scales

Scales	χ^2	CFI	TLI	RMSEA	SRMR
Students Value Mathematics	135.667 (p< .05)	.993	.990	.033 *(.028, .039)	.036
Students Like Learning Mathematics	116.191 (p< .05)	.998	.997	.030 *(.024, .035)	.027
Students Confident in Mathematics	1510.635 (p< .05)	.954	.938	.117 *(.112, .122)	.094
Sense of School Belonging	9.307 (p> .05)	.999	.997	.015 *(.000, .029)	.016

*Lower and upper confidence interval for RMSEA

Table 3 shows that CFI, TLI, RMSEA and SRMR values for "Students Value Mathematics" and "Students Like Learning Mathematics" scales indicate good fit, and the $\chi 2$ value is not within acceptable fit ranges. When the literature is examined, it is stated that the sample size affects the $\chi 2$ (Kline, 2015). When all goodness-of-fit indices are evaluated together, it can be said that "Students Value Mathematics" and "Students Like Learning Mathematics" scales are confirmed. CFI value for "Students Confident in Mathematics" scale indicates a good fit. TLI value is acceptable, and the $\chi 2$, RMSEA, and SRMR values are not at acceptable ranges. When all goodness-of-fit indices are evaluated together, it can be said that "Students Confident in Mathematics" scale cannot be confirmed. For "Sense of School Belonging" scale, $\chi 2$, CFI, TLI, RMSEA and SRMR values indicate a good fit. When all goodness-of-fit indices are evaluated together, it can be said that "Sense of School Belonging" scale is confirmed.

Regarding reliability, the Cronbach Alpha coefficient was calculated for three scales except for "Students Confident in Mathematics" scale because the scale was not confirmed by CFA. Cronbach Alpha coefficients were calculated as 0.87, 0.92, and 0.75 for "Students Value Mathematics", "Students Like Learning Mathematics", and "Sense of School Belonging" scales, respectively. A Cronbach Alpha coefficient of 0.70 and above indicates that the level of reliability is good, and a value between 0.60 and 0.70 indicates that the level of reliability is acceptable (Hair et al., 2014). In this context, the reliability of the three scales is reasonable.

Measurement Invariance

In the data analysis, measurement invariance according to gender was tested for three affective scales, which CFA confirmed at the last stage. MG-CFA method was used to test the measurement invariance. Measurement invariance by MG-CFA method, structural invariance, weak/metric invariance, strong/scalar invariance, and strict invariance models are examined by looking for evidence. Measurement invariance models have a 4-stage hierarchical structure (Byrne et al., 1989; Stark et al., 2006; Vandenberg & Lance, 2000; Wu et al., 2007). The model with the minor parameter constraints is the structural invariance model, and the model with the most parameter constraint is the strict invariance model. Due to the hierarchical structure of the invariance models, if there is no evidence that measurement invariance is provided for the model with fewer parameter constraints, there will be no evidence that measurement invariance is provided for the models with more parameter constraints. The goodness of fit values at that stage is considered when looking for evidence for invariance models. Then, the $\Delta\chi^2$ value between it and the previous model, which has fewer parameter limitations, is considered. Suppose the $\Delta\chi^2$ value is not statistically significant (p> .05), and the goodness-of-fit values of the model with more parameter limitations are within acceptable values. In that case, evidence of measurement invariance is obtained for the model with more parameter limitations. However, since the $\Delta\chi^2$ value is affected by the sample size, the p-value in the $\Delta\chi^2$ test tends to be significant. For this reason, it is stated by some researchers that ΔCFI (Comparative Fit Index Differences), ΔRMSEA (Root Mean Square Error of Approximation Differences), and ΔSRMR (Standardized Root Mean Square Residual Differences) values can be considered instead of $\Delta\chi^2$ (Chen, 2007; Cheung & Rensvold, 2002; French & Finch, 2006; Meade et al., 2008). When comparing the models, if the ΔCFI value is between -0.01 and +0.01, evidence is obtained that the model with more parameter constraint provides measurement

invariance (Cheung & Rensvold, 2002). Chen (2007) states that besides the 0.01 change in ΔCFI value, changes of 0.015 for the ΔRMSEA value and 0.030 for the ΔSRMR value are acceptable in the weak/metric invariance stage, while changes of 0.015 for the ΔRMSEA and ΔSRMR values are acceptable in the scalar/strong invariance and strict invariance stages. Considering all these reasons, in this study, while comparing the invariance models, ΔCFI, ΔRMSEA, and ΔSRMR values were also considered, in addition to the $\Delta\chi^2$ value. This study considered that at least two of the difference tests were within the desired criteria while deciding that models with measurement invariance were provided. MG-CFA results by gender for "Sense of School Belonging" scale are presented in Table 4.

Table 4
MG-CFA Results by Gender for "Sense of School Belonging" Scale

	Structural Invariance	Metric Invariance	Scalar Invariance	Strict Invariance
χ^2	$p > .05$	$p > .05$	$p > .05$	$p < .05$
CFI	1.000	1.000	1.000	0.991
TLI	0.999	1.000	1.000	0.992
RMSEA	0.009	0.002	0.000	0.024
	*(0.000, 0.026)	*(0.000, 0.022)	*(0.000, 0.018)	*(0.015, 0.033)
SRMR	0.015	0.017	0.018	0.039
ΔCFI	-	0.000	0.000	-0.009
ΔRMSEA	-	-0.006	-0.002	0.024
ΔSRMR	-	0.002	0.001	0.021
$\Delta\chi^2$	-	$p > .05$	$p > .05$	$p < .05$

*Lower and upper confidence interval for RMSEA

According to Table 4, all the goodness-of-fit values calculated for the structural invariance model indicate a good fit. For this reason, "Sense of School Belonging" scale provides the structural invariance model. All the goodness-of-fit values calculated for the weak/metric invariance model indicate a good fit. The $\Delta\chi^2$, ΔCFI, ΔRMSEA, and ΔSRMR values between the structural and the weak/metric invariance models are all within the benchmark values. For this reason, "Sense of School Belonging" scale provides the weak/metric invariance model. All the goodness-of-fit values calculated for the strong/scalar invariance model indicate a good fit. The $\Delta\chi^2$, ΔCFI, ΔRMSEA, and ΔSRMR values between the weak/metric and the strong/scalar invariance models are all within the benchmark values. For this reason, "Sense of School Belonging" scale provides a strong/scalar invariance model. The χ^2 value, one of the goodness-of-fit values calculated for the strict invariance model, indicates an unacceptable fit. Other goodness of fit values indicate a good fit. Only the ΔCFI value is among the benchmark values for the difference tests between the strong/scale and strict invariance models. $\Delta\chi^2$, ΔRMSEA, and ΔSRMR values are outside the criterion values. In this study, "Sense of School Belonging" scale does not provide the strict invariance model since at least two of the difference tests were determined as a prerequisite for obtaining evidence of measurement invariance. MG-CFA results by gender for "Students Like Learning Mathematics" scale are presented in Table 5.

Table 5
MG-CFA Results by Gender for "Students Like Learning Mathematics" Scale

	Structural Invariance	Metric Invariance	Scalar Invariance	Strict Invariance
χ^2	$p < .05$	$p < .05$	$p < .05$	$p < .05$
CFI	0.998	0.998	0.997	0.997
TLI	0.998	0.997	0.997	0.997
RMSEA	0.027	0.29	0.030	0.030
	*(0.021, 0.033)	*(0.023, 0.034)	*(0.025, 0.036)	*(0.025, 0.035)
SRMR	0.026	0.029	0.031	0.033
ΔCFI	-	-0.001	-0.001	-0.009
ΔRMSEA	-	0.002	0.002	0.000
ΔSRMR	-	0.003	0.003	0.001
$\Delta\chi^2$	-	$p < .05$	$p < .05$	$p < .05$

*Lower and upper confidence interval for RMSEA

According to Table 5, the χ^2 value, one of the goodness-of-fit values calculated for the structural invariance model, indicates an unacceptable fit. Other goodness of fit values indicate a good fit. For this reason, "Students Like Learning Mathematics" scale provides the structural invariance model. The χ^2 value, one of the goodness-of-fit values calculated for the weak/metric invariance model, indicates an unacceptable fit. Other goodness of fit values indicate a good fit. The ΔCFI, ΔRMSEA, and ΔSRMR values between the structural and weak/metric invariance models are among the benchmark values. Only the $\Delta\chi^2$ value was found to be statistically significant. Since three of the four difference tests are among the criteria values, "Students Like Learning Mathematics" scale provides the weak/metric invariance model. The χ^2 value, one of the goodness-of-fit values calculated for the strong/scalar invariance model, indicates an unacceptable fit. Other goodness of fit values indicate a good fit. The ΔCFI, ΔRMSEA, and ΔSRMR values between the weak/metric and the strong/scalar invariance models are among the benchmark values. Only the $\Delta\chi^2$ value was found to be statistically significant. Since three of the four difference tests are among the criteria values, "Students Like Learning Mathematics" scale provides a strong/scalar invariance model. The χ^2 value, one of the goodness-of-fit values calculated for the strict invariance model, indicates an unacceptable fit. Other goodness of fit values indicate a good fit. The ΔCFI, ΔRMSEA, and ΔSRMR values between the strong/scale and strict invariance models are within the benchmark values. Only the $\Delta\chi^2$ value was found to be statistically significant. Since three of the four difference tests are among the criteria values, "Students Like Learning Mathematics" scale provides the strict invariance model. MG-CFA results by gender for "Students Value Mathematics" scale are presented in Table 6.

Table 6

MG-CFA Results by Gender for "Students Value Mathematics" Scale

	Structural Invariance	Metric Invariance	Scalar Invariance	Strict Invariance
χ^2	p< .05	p< .05	p< .05	p< .05
CFI	0.995	0.989	0.986	0.983
TLI	0.992	0.986	0.985	0.984
RMSEA	0.030	0.040	0.041	0.042
	*(0.024, 0.036)	*(0.035, 0.045)	*(0.036, 0.046)	*(0.038, 0.047)
SRMR	0.033	0.044	0.046	0.056
ΔCFI	-	-0.006	-0.003	-0.003
ΔRMSEA	-	0.010	0.002	0.001
ΔSRMR	-	0.011	0.002	0.010
$\Delta\chi^2$	-	p< .05	p< .05	p< .05

*Lower and upper confidence interval for RMSEA

According to Table 6, the χ^2 value, one of the goodness-of-fit values calculated for the structural invariance model, indicates an unacceptable fit. Other goodness of fit values indicate a good fit. For this reason, "Students Value Mathematics" scale provides the structural invariance model. The χ^2 value, one of the goodness-of-fit values calculated for the weak/metric invariance model, indicates an unacceptable fit. Other goodness of fit values indicate a good fit. The ΔCFI, ΔRMSEA, and ΔSRMR values between the structural and the weak/metric invariance models are among the benchmark values. Only the $\Delta\chi^2$ value was found to be statistically significant. Since three of the four difference tests are among the criteria values, "Students Value Mathematics" scale provides the weak/metric invariance model. The χ^2 value, one of the goodness-of-fit values calculated for the strong/scalar invariance model, indicates an unacceptable fit. Other goodness of fit values indicate a good fit. The ΔCFI, ΔRMSEA, and ΔSRMR values between the weak/metric and the strong/scalar invariance models are among the benchmark values. Only the $\Delta\chi^2$ value was found to be statistically significant. Since three of the four difference tests are among the criterion values, it can be said that "Students Value Mathematics" scale provides a strong/scalar invariance model. The χ^2 value, one of the goodness-of-fit values calculated for the strict invariance model, indicates an unacceptable fit. SRMR value indicates an acceptable fit. Other goodness of fit values indicate a good fit. The ΔCFI, ΔRMSEA, and ΔSRMR values between the strong/scale and strict invariance models are within the benchmark values. Only the $\Delta\chi^2$ value was found to be statistically

significant. Since three of the four difference tests are among the criteria values, "Students Value Mathematics" scale provides the strict invariance model.

Discussion and Conclusion

This study aimed to examine the measurement invariance of affective scales in the TIMSS 2019 Turkey 8th-grade mathematics student questionnaire in the context of gender. For this purpose, the validity of the relevant affective structures for the Turkish sample was tested by performing CFA separately for the four scales. Then, to determine the measurement invariance, MG-CFA was performed according to gender in the scales confirmed by CFA.

"Students Confident in Mathematics" scale could not be verified by CFA. However, many studies in the literature examining the effect of self-confidence on success in mathematics (Akyüz-Aru, 2020; Akyüz & Pala, 2010; Atar, 2011; Aydın, 2015; Demir et al., 2010; Ertürk & Erdinç-Akan, 2018a; Işlak, 2020; Khine et al., 2015; Koç, 2019; Oral & McGivney, 2013; Usta & Demirtaşlı, 2018). It can be said that "Students Confident in Mathematics" scale is not valid for the mathematics data of the 8th-grade sample of TIMSS Turkey. For this reason, the validity of the results of studies in which this scale will be used in the data of the TIMSS 2019 Turkey 8th-grade mathematics sample in the future will also be low. The measurement invariance of "Students Confident in Mathematics" scale was not examined within the scope of this study since it is not statistically significant to perform MG-CFA analyses of a structure that CFA cannot verify. As a matter of fact, in some of the studies testing the measurement invariance of self-confidence in mathematics, measurement invariance is ensured (Bofah & Hannula, 2015; Cakici-Eser, 2021; Polat, 2019; Uyar, 2021), while measurement invariance cannot be achieved in some (Ertürk & Erdinç-Akan, 2018b). The finding of this study shows parallelism with studies that cannot provide measurement invariance.

"Sense of School Belonging" scale was validated by CFA. For this reason, the sense of belonging to the school is valid for the mathematics data of the 8th-grade sample of TIMSS Turkey. Although there are many studies (Akyüz & Pala, 2010; Akyüz & Satıcı, 2013; Işlak, 2020; Koç, 2019; Sarı et al., 2017; Sarıer, 2020) examining the effect of belonging to school on mathematics achievement in the literature, no study of measurement invariance of this affective variable was found. It is crucial to test the measurement invariance of belonging to the school, whose effect on mathematics achievement is the subject of research. As a result of the MG-CFA for this scale, evidence could be obtained that the scale provided strong/scalar invariance but no evidence that it provided strict invariance. Since this scale provides strong/scalar invariance, the factor score is zero in gender subgroups, while the regression constants are equal. The mean scores on the factor and observed variables are comparable. The differences between the mean scores of the subgroups arise from the latent variable (Başusta & Gelbal, 2015). As a result, it can be said that this scale provides measurement invariance according to gender, there is no bias in the items according to gender, and the mean scores are comparable according to gender.

"Students Like Learning Mathematics" scale was validated by CFA. For this reason, liking mathematics is valid for the mathematics data of the 8th-grade sample of TIMSS Turkey. In the literature, there are many studies (Erşan, 2016; Ertürk & Erdinç-Akan, 2018a; Khine et al., 2015; Koç, 2019; Oral & McGivney, 2013) examining the effect of liking mathematics on success in mathematics. It is crucial to test the measurement invariance of the affective variable of liking mathematics, whose effect on mathematics achievement is the subject of research. As a result of the MG-CFA conducted for this scale, evidence was obtained that the scale provides strict invariance. Since strict invariance is provided in this scale, it was concluded that the error variances for the measured items were equal in gender groups (Widaman & Reise, 1997). In the literature, in the studies in which the measurement invariance of the affective variable of liking mathematics was tested (Bofah & Hannula,2015; Cakici-Eser, 2021; Ertürk & Erdinç-Akan, 2018b; Polat, 2019; Shukla & Konold, 2014), evidence was obtained regarding the measurement invariance. There was no study in which measurement invariance could not be achieved in the measurement invariance studies conducted with the affective variable of liking mathematics. The finding of this study is in parallel with the studies in the literature. As a result, it can be said that this

scale provides measurement invariance according to gender, there is no bias in the items according to gender, and the mean scores are comparable according to gender.

"Students Value Mathematics" scale was validated by CFA. For this reason, valuing mathematics is valid for the mathematics data of the 8th-grade sample of TIMSS Turkey. In the literature, studies (Doğan & Barış, 2010; Khine et al., 2015) examine the effect of valuing mathematics on mathematics achievement. It is crucial to test the measurement invariance of the affective variable of valuing mathematics, whose effect on mathematics achievement is the subject of research. As a result of the MG-CFA conducted for this scale, evidence was obtained that the scale provides strict invariance. Since strict invariance is provided in this scale, it was concluded that the error variances for the measured items were equal in gender groups (Widaman & Reise, 1997). In the literature, in studies where the measurement invariance of the affective variable of valuing mathematics was tested (Bofah & Hannula, 2015; Polat, 2019; Uyar, 2021), evidence was obtained that the measurement invariance was provided. In the studies of measurement invariance conducted with the affective variable of valuing mathematics, no study was found in which measurement invariance could not be achieved. The finding of this study is in parallel with the studies in the literature. As a result, it can be said that this scale provides measurement invariance according to gender, there is no bias in the items according to gender, and the mean scores are comparable according to gender.

The study's large sample size affects the χ^2 goodness-of-fit index used in CFA and $\Delta\chi^2$ values used to compare the differences between models in MG-CFA. In future studies, the effect of sample size can be reduced by choosing a smaller sample than the entire TIMSS sample. In this study, the structure of "Students Confident in Mathematics" scale could not be confirmed by CFA. In future research, it is recommended that researchers approach this scale with caution. In this study, only measurement invariance was examined. Measurement invariance can reveal whether there are biases in terms of items in subgroups. However, it does not reveal which items have biases. In future research, it can be revealed from which items the biases originate by examining the Differential Item Functioning (DIF) based on IRT for "Students Confident in Mathematics" scale.

Declarations

Author Contribution: Mehmet Atılgan: Conceptualization, methodology, analysis, writing & editing, visualization. Kaan Zülfikar Deniz: Conceptualization, methodology, writing-review & editing, supervision.

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

Ethical Approval: The study was ethically approved by the Uşak University Social and Humanities Scientific Research and Publication Ethics Committee (decision number: 2022-66, dated 14/04/2022).

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

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

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

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

References

Akyüz-Aru, S. (2020). *4. sınıf öğrencilerinin fen ve matematik başarısına etki eden değişkenlerin incelenmesi "TIMSS 2015 durum analizi"* [Investigation of variables affecting science and mathematics success of grade 4 students "TIMSS 2015 status analysis"] [Unpublished doctoral dissertation]. Gazi Üniversitesi. https://tez.yok.gov.tr/UlusalTezMerkezi/

Akyüz, G., & Pala, N. M. (2010). PISA 2003 sonuçlarına göre öğrenci ve sınıf özelliklerinin matematik okuryazarlığına ve problem çözme becerilerine etkisi [The effect of student and class characteristics on mathematics literacy and problem solving in PISA 2003]. *İlköğretim Online, 9*(2), 668-678.

Akyüz, G., & Satıcı, K. (2013). PISA 2003 verilerine göre matematik okuryazarlığının çeşitli değişkenler açısından incelenmesi: Türkiye ve Hong Kong-Çin modelleri [Investigation of the factors affecting mathematics literacy using PISA 2003 results: Turkey and Hong Kong-China]. *Kastamonu Üniversitesi Kastamonu Eğitim Dergisi, 21*(2), 503 - 522.

Atar, B. (2011). Tanımlayıcı ve Açıklayıcı Madde Tepki Modellerinin TIMSS 2007 Türkiye Matematik Verisine Uyarlanması [An application of descriptive and explanatory ıtem response models to TIMSS 2007 Turkey mathematics data]. *Eğitim ve Bilim, 36*(159), 255 - 269.

Aydın, M. (2015). *Öğrenci ve okul kaynaklı faktörlerin TIMSS matematik başarısına etkisi* [The effects of student-level and school-level factors on middle school students' mathematics achievement] [Unpublished doctoral dissertation]. Necmettin Erbakan Üniversitesi. https://tez.yok.gov.tr/UlusalTezMerkezi/

Başusta, N. B., & Gelbal, S. (2015). Gruplar arası karşılaştırmalarda ölçme değişmezliğinin test edilmesi: PISA öğrenci anketi örneği [Examination of Measurement Invariance at Groups' Comparisons: A Study on PISA Student Questionnaire]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 30*(4), 80-90.

Bloom, B. S. (2012). *İnsan nitelikleri ve okulda öğrenme* [Human characteristics and school learning] (D. A. Özçelik, Trans.). Pegem Akademi.

Bofah, E.At., & Hannula, M.S. (2015). TIMSS data in an African comparative perspective: Investigating the factors influencing achievement in mathematics and their psychometric properties. *Large-scale Assessments in Education*, *3*(4), 1-36. https://doi.org/10.1186/s40536-015-0014-y

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS and SIMPLIS: Basic concepts, application and programming*. Lawrence Erlbaum.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Byrne, B. M. & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology, 34*(2), 155–175. https://doi.org/10.1177/0022022102250225

Cakici-Eser, D. (2021). Investigation of measurement ınvariance according to home resources: TIMSS 2015 mathematical affective characteristics questionnaire. *International Journal of Assessment Tools in Education, 8*(3), 633-648. https://doi.org/10.21449/ijate.817168

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling, 9*(2), 233-255.

Cheung, G., W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-cultural Psychology, 31*(2), 188–213. https://doi.org/10.1177/0022022100031002003

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2016). *Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL uygulamaları* (5. Baskı) [Multivariate statistics SPSS and LISREL applications for social sciences (5th ed.)]. Pegem Akademi.

Demir, E. (2017). Testing measurement invariance of the students' affective characteristics model across gender sub-groups. *Educational Sciences:Theory and Practice, 17*(1), 47–62. https://doi.org/10.12738/estp.2017.1.0223

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

Demir, M. C. (2020). *TIMSS 2015 fen duyuşsal özelliklerinin cinsiyet ve bölgelere göre incelenmesi* [An examination of TIMMS 2015 science affective factors with regard to gender and regions] [Unpublished master's dissertation]. Hacettepe Üniversitesi. https://tez.yok.gov.tr/UlusalTezMerkezi/

Doğan, N., & Barış, F. (2010). Tutum, değer ve özyeterlik değişkenlerinin TIMSS-1999 ve TIMSS-2007 Sınavlarında öğrencilerin matematik başarılarını yordama düzeyleri [Prediction levels of attitude, value and self-efficacy variables for students' mathematics achievement in TIMSS-1999 and TIMSS-2007 Exams]. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 1*(1), 44-50.

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing, 5*(1), 23-35. https://doi.org/10.1207/s15327574ijt0501_3

Erşan, Ö. (2016). *TIMSS 2011 sekizinci sınıf öğrencilerinin matematik başarılarını etkileyen faktörlerin çok düzeyli yapısal eşitlik modeliyle incelenmesi* [Investigation of the factors affecting mathematics achievement of TIMSS 2011 eighth grade students with multilevel structural equation modeling] [Unpublished master's dissertation]. Hacettepe Üniversitesi. https://tez.yok.gov.tr/UlusalTezMerkezi/

Ertürk, Z., & Erdinç-Akan, O. (2018a). TIMSS 2015 matematik başarısını etkileyen değişkenlerin yapısal eşitlik modeli ile incelenmesi [The Investigation of the Variables Effecting TIMSS 2015 Mathematics

Achievement with SEM]. *Ulusal Eğitim Akademisi Dergisi (UEAD), 2*(2), 14-34. https://doi.org/10.32960/uead.407078

Ertürk, Z., & Erdinç-Akan, O. (2018b). TIMSS 2015 matematik başarısı ile ilgili bazı değişkenlerin cinsiyete göre ölçme değişmezliğinin incelenmesi [The Investigation of Measurement Invariance of the Variables Related to TIMSS 2015 Mathematics Achievement in terms of Gender]. *Kuramsal Eğitimbilim Dergisi [Journal of Theoretical Educational Science]*, UBEK-2018, 204-226. https://doi.org/10.30831/akukeg.412604

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

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement ınvariance. *Structural Equation Modeling: A Multidisciplinary Journal, 13*(3), 378–402. https://doi.org/10.1207/s15328007sem1303_3

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

Gungor, M., & Atalay-Kabasakal, K. (2020). Investigation of measurement ınvariance of science motivation and self-efficacy model: PISA 2015 Turkey sample. *International Journal of Assessment Tools in Education, 7* (2), 207-222. https://doi.org/10.21449/ijate.730481

Gülleroğlu, H. D. (2017). PISA 2012 matematik uygulamasına katılan Türk öğrencilerin duyuşsal özeliklerinin cinsiyete göre ölçme değişmezliğinin incelenmesi [An investigation of measurement invariance by gender for the Turkish students' affective characteristics who took the PISA 2012 math test]. *Gazi Eğitim Fakültesi Dergisi, 37*(1), 151-175.

Hair Jr, J. F., Black, C. W., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis* (7th ed.). Pearson Education.

Hanci, A. (2015). *8. sınıf öğrencilerinin öğrenme stilleri ve TIMSS matematik başarılarının farklı değişkenler açısından incelenmesi: Bayburt ili örneği* [Investigation of 8th grade students' learning styles and TIMSS matematics achivements from the aspect of different variable: Bayburt sample] [Unpublished master's dissertation]. Bayburt Üniversitesi. https://tez.yok.gov.tr/UlusalTezMerkezi/

Harrington, D. (2009). *Confirmatory factor analysis*. Oxforda University Press, Inc.

Işlak, O. (2020). *TIMSS 2015 uygulamasına katılan öğrencilerin matematik başarılarının öğrenci, aile ve okul değişkenlerine göre yordanma* [Prediction of mathematics achievement of students attending TIMSS 2015 according to student, family and school variables] [Unpublished doctoral dissertation]. Burdur Mehmet Akif Ersoy Üniversitesi. https://tez.yok.gov.tr/UlusalTezMerkezi/

İlhan, M., & Öner-Sünkür, M. (2012). Matematik kaygısı ile olumlu ve olumsuz mükemmeliyetçiliğin matematik başarısını yordama gücü [The predictive power of mathematics anxiety and positive and negative perfectionism on math achievement]. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 8*(1), 178-188.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2021). semTools: Useful tools for structural equation modeling. *R package version 0.5-5*. Retrieved from https://CRAN.R-project.org/package=semTools

Jung, J. Y. (2019). *A Comparison of CFA and ESEM approaches using TIMSS science attitudes items: Evidence from factor structure and measurement invariance* [Unpublished master's dissertation]. Purdue University.

Karakoc-Alatli, B., Ayan, C., Polat-Demir, B., & Uzun, G. (2016). Examination of the TIMSS 2011 fourth grade mathematics test in terms of cross-cultural measurement invariance. *Eurasian Journal of Educational Research, 66*, 389-406. https://doi.org/10.14689/ejer.2016.66.22

Karasar, N (2011). *Bilimsel Araştırma Yöntemi* [Scientific Research Method]. Nobel Yayıncılık.

Kesici, A. (2018). Lise öğrencilerinin matematik motivasyonunun matematik başarısına etkisinin incelenmesi. *OMÜ Eğitim Fakültesi Dergisi, 37*(2), 177-194. https://doi.org/10.7822/omuefd.438550

Kesici, A., & Aşılıoğlu, B. (2017). Ortaokul öğrencilerinin matematiğe yönelik duyuşsal özellikleri ile temel eğitimden ortaöğretime geçiş (TEOG) sınavları öncesi yaşadıkları stresin matematik başarısına etkisi [The Effect of Secondary Students' Affective Features Towards Mathematics and The Stress They Experience Before The TEOG Exam (The Exam For Accessing to Various Types of High Schools) on Their Mathematical Success]. *Kırşehir Eğitim Fakültesi Dergisi, 18*(3), 394-414.

Khine, M. S., Al-Mutawah, M., & Afari, E. (2015). Determinants of affective factors in mathematics achievement: Structural equation modeling approach. *Journal of Studies in Education, 5*(2), 199–211.

Kıbrıslıoğlu, N. (2015). *PISA 2012 matematik öğrenme modelinin kültürlere ve cinsiyete göre ölçme değişmezliğinin incelenmesi: Türkiye – Çin (Şangay) – Endonezya örneği* [The investigation of measurement invariance PISA 2012 mathematics learning model according to culture and gender: Turkey - China (Shangai) - Indonesia] [Unpublished master's dissertation]. Hacettepe Üniversitesi. https://tez.yok.gov.tr/UlusalTezMerkezi/

Kilic, S. & Askin, Ö. E. (2013). Parental influence on students' mathematics achievement: the comparative study of Turkey and best performer countries in TIMSS 2011. *Procedia - Social and Behavioral Sciences, 106*, 2000-2007. https://doi.org/10.1016/j.sbspro.2013.12.228

Kline, R. B. (2015). *Principles and practices of structural equation modeling* (4th ed.). The Guilford Press.

Koç, O. (2019). *4. ve 8. sınıf öğrencilerinin TIMSS 2015 matematik başarısını yordayan değişkenlerin belirlenmesi* [Determination of predictive variables of 4th and 8th grade students' on TIMSS 2015 mathematics achievement] [Unpublished master's dissertation]. Akdeniz Üniversitesi. https://tez.yok.gov.tr/UlusalTezMerkezi/

Lay, Y. F., Ng, K. T., & Chong, P. S. (2015). Analyzing affective factors related to eighth grade learners' science and mathematics achievement in TIMSS 2007. *Asia-Pacific Education Researcher, 24*(1), 103–110. https://doi.org/10.1007/s40299-013-0163-0

Louis, R.A., & Mistele, J.M. (2012). The differences in scores and self-efficacy by student gender in mathematics and science. *International Journal of Science and Mathematics Education, 10*, 1163–1190. https://doi.org/10.1007/s10763-011-9325-9

Ma, Y., & Qin, X. (2021). Measurement invariance of information, communication and technology (ICT) engagement and its relationship with student academic literacy: Evidence from PISA 2018. *Studies in Educational Evaluation, 68*, 1–15. https://doi.org/10.1016/j.stueduc.2021.100982

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568–592. https://doi.org/10.1037/0021-9010.93.3.568

Ministry of National Education (MoNE) (2020). *TIMSS 2019 Türkiye ön raporu* [TIMSS 2019 Turkey preliminary report]. Ankara: Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü.

Ministry of National Education (MoNE). (2016). *TIMSS 2015 ulusal matematik ve fen ön raporu: 4. ve 8. sınıflar* [TIMSS 2015 national math and science preliminary report: 4th and 8th grades]. Ankara.

Meng, L., Qiu, C., & Boyd-Wilson, B. (2019). Measurement invariance of the ICT engagement construct and its association with students' performance in China and Germany: Evidence from PISA 2015 data. *British Journal of Educational Technology, 50*(6), 3233–3251. https://doi.org/10.1111/bjet.12729

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

Mindrila, D. (2010). Maximum likelihood (ML) and diagonally weighted least squares (DWLS) estimation procedures: A comparison of estimation bias with ordinal and multivariate non-normal data. *International Journal of Digital Society, 1*(1), 60–66. https://doi.org/10.20533/ijds.2040.2570.2010.0010

Mohammadpour, E. (2012). Factors accounting for mathematics achievement of singaporean eighth-graders. *The Asia-Pacific Education Researcher, 21*(3), 507-518.

Mulaik, S. A. (2007). There is a place for approximate fit in structural equation modelling. *Personality and Individual Differences, 42*(5), 883-891. https://doi.org/10.1016/j.paid.2006.10.024

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. Boston, US.

Oral, I., & McGivney, E. (2013). *Türkiye'de matematik ve fen bilimleri alanlarında öğrenci performansı ve başarının belirleyicileri: TIMSS 2011 analizi* [Student performance and determinants of success in mathematics and science in Turkey: TIMSS 2011 analysis]. İstanbul: Eğitim Reformu Girişimi Raporu.

Öğretmen, T. (2006). *Uluslararası okuma becerilerinde gelişim projesi (PIRLS) 2001 testinin psikometrik özelliklerinin incelenmesi: Türkiye-Amerika Birleşik Devletleri örneği* [The investigation of psychometric properties of the test of progress in international reading literacy (PIRLS) 2001: The model of Turkey-United States of America] [Unpublished doctoral dissertation]. Hacettepe Üniversitesi. https://tez.yok.gov.tr/UlusalTezMerkezi/

Ölçüoğlu, R., & Çetin, S. (2016). TIMSS 2011 sekizinci sınıf öğrencilerinin matematik başarısını etkileyen değişkenlerinin bölgelere göre incelenmesi [The investigation of the variables that affecting eight grade students' TIMSS 2011 math achievement according to regions]. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 7*(1), 202-220. https://doi.org/10.21031/epod.34424

Öncü, Ö. (2019). *TIMSS 2015 sekizinci sınıf matematik başarı testinin oecd ülkelerine göre ölçme değişmezliğinin incelenmesi* [An investication into the measurement invariance according to OECD countries of TIMSS 2015 eight grade math achievement test] [Unpublished master's dissertation]. Akdeniz Üniversitesi. https://tez.yok.gov.tr/UlusalTezMerkezi/

Patterson, M., Perry, E., Decker, C., Eckert, R., Klaus, S., Wendling, L., & Papanastasiou, E. (2003). Factors associated with high school mathematics performance in the United States. *Studies in Educational Evaluation, 29*(2), 91-108. https://doi.org/10.1016/S0191-491X(03)00017-8

Pituch, K. A., & Stevens, J. P. (2016). *Applied multivariate statistics for the social sciences: Analyses with SAS and IBM's SPSS* (6th Ed.). Routledge.

Polat, M. (2019). *TIMSS-2015 matematik ve fen duyuşsal özellik modellerinin kültürlere, cinsiyete ve bölgelere göre ölçme değişmezliğinin incelenmesi* [The investigation of measurement invariance of TIMSS-2015 mathematics and science affective characteristics models according to culture, gender and statistical region] [Unpublished master's dissertation]. Hacettepe Üniversitesi. https://tez.yok.gov.tr/UlusalTezMerkezi/

Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling* (2nd ed.). Lawrence Erlbaum Associates, Publishers.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rutkowski, D., & Rutkowski, L. (2013). Measuring socioeconomic background in PISA: One size might not fit all. *Research in Comparative and International Education, 8*(3), 259–278. https://doi.org/10.2304/rcie.2013.8.3.259

Salzberger, T., Sinkovics, R., R., & Schlegelmilch, B. B. (1999). Data equivalence in cross-cultural research: a comparison of classical test theory and latent trait theory based approaches. *Australasian Marketing Journal, 7*(2), 23-38. https://doi.org/10.1016/S1441-3582(99)70213-2

Sarı, M. H., & Ekici, G. (2018). İlkokul 4. sınıf öğrencilerinin matematik başarıları ile aritmetik performanslarını etkileyen duyuşsal değişkenlerin belirlenmesi [Determination of affective variables affecting mathematical achievement and arithmetic performance of primary school 4th grade students]. *OPUS International Journal of Society Researches, 8*(15), 1562-1594. https://doi.org/10.26466/opus.451025

Sarı, M. H., Arıkan, S., & Yıldızlı, H. (2017). 8. sınıf matematik akademik başarısını yordayan faktörler-TIMSS 2015 [Factors predicting mathematics achievement of 8th graders in TIMSS 2015]. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 8*(3), 246-265.

Sarıer, Y. (2020). TIMSS uygulamalarında Türkiye'nin performansı ve akademik başarıyı yordayan değişkenler [Turkey's performance in TIMSS applications and variables predicting academic achievement]. *Temel Eğitim Dergisi, 2*(2), 6-27.

Scherer, R., Nilsen, T., & Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality: An investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Frontiers in Psychology, 7*(110), 1-16. https://doi.org/10.3389/fpsyg.2016.00110

Schumacker, R. E., & Beyerlein, S. T. (2000). Confirmatory factor analysis with different correlation types and estimation methods. *Structural Equation Modeling, 7*(4), 629–636. https://doi.org/10.1207/S15328007SEM0704_6

Shukla, K., & Konold, T. (2014). *Fondness of math and science as measured by the TIMSS student questionnaire: Invariance across U.S. ethnic groups.* Paper presented at the 2014 Annual Meeting of the American Educational Research Association, Philadelphia, USA

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with comfirmatory factor analysis and ıtem response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292–1306. https://doi.org/10.1037/0021-9010.91.6.1292

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson Education.

Tavlıca, A. (2019). *TIMSS 2015 dördüncü sınıf matematik testinin ölçme değişmezliğinin ülkelere göre incelemesi* [An investigation of measurement invariance for TIMSS 2015 fourth grade mathematics test according to countries] [Unpublished master's dissertation]. Akdeniz Üniversitesi. https://tez.yok.gov.tr/UlusalTezMerkezi/

Usta, H. G., & Demirtaşlı, R. N. (2018). PISA 2012 matematik okuryazarlığı üzerine uluslararası bir karşılaştırma: Türkiye ve Finlandiya [An international comparison according to PISA 2012 mathematical literacy Turkey and Finland]. *Turkish Studies (Elektronik), 13*(11), 1389 - 1420. https://doi.org/10.7827/TurkishStudies.13377

Uyar, S. (2021). Factor structure and measurement invariance of the TIMSS 2015 mathematics attitude questionnaire: Exploratory structural equation modelling approach. *International Journal of Assessment Tools in Education, 8*(4), 855–871. https://doi.org/10.21449/ijate.796862

Uzun, B., & Öğretmen, T. (2010). Fen başarısı ile ilgili bazı değişkenlerin TIMSS-R Türkiye örnekleminde cinsiyete göre ölçme değişmezliğinin değerlendirilmesi [Assessing the measurement ınvariance of factors that are related to students' science achievement across gender in TIMSS-R Turkey sample]. *Eğitim ve Bilim, 35*(155), 26–35.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the MI literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–69. https://doi.org/10.1177/109442810031002

Wang, Z., Osterlind, S., & Bergin, D. A. (2012). Building mathematics achievement models in four countries using TIMSS 2003. *International Journal of Science and Mathematics Education, 10*(5), 1-28. https://doi.org/10.1007/s10763-011-9328-6

Webster, B. J., & Fisher, D. L. (2000). Accounting for variation in science and mathematics achievement: A multilevel analysis of Australian data third international mathematics and science study (TIMSS). *School Effectiveness and School Improvement, 11*(3), 339–360. https://doi.org/10.1076/0924-3453(200009)11:3;1-G;FT339

Wicherts, J. M. (2007). *Group differences in intelligence test performance* [Unpublished doctoral dissertation]. University of Amsterdam.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). American Psychological Association.

Wu, A.D., Li, Z., & Zumbo, B.D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research and Evaluation, 12*, 1-26. https://doi.org/10.7275/mhqa-cd89

Yandı, A., Köse, İ. A., & Uysal, Ö. (2017). Farklı yöntemlerle ölçme değişmezliğinin incelenmesi: PISA 2012 örneği [Examining measurement invariance with different methods: Example of PISA 2012]. *Mersin Üniversitesi Eğitim Bilimleri Dergisi, 13*(1), 243-253. https://doi.org/10.17860/mersinefd.305952

Yin, L., & Fishbein, B. (2020). Creating and interpreting the TIMSS 2019 context questionnaire scales. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report* (pp. 16.1-16.331). Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods/chapter-16.html

Yücel, Z., & Koç, M. (2011). İlköğretim öğrencilerinin matematik dersine karşı tutumlarının başarı düzeylerini yordama gücü ile cinsiyet arasındaki ilişki [The relationship between the prediction level of elementary school students' math achievement by their math attitudes and gender]. *İlköğretim Online, 10*(1), 133-143.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(3); 200-221

Investigation of Differential Item and Step Functioning Procedures in Polytomus Items*

Yasemin KUZU** Selahattin GELBAL***

Abstract

This study aimed to compare differential item functioning (DIF) and differential step function (DSF) detection methods in polytomous items under various conditions. In this context, the study examined Kazakhstan and Turkey data obtained from the ICT Familiarity Questionnaire in PISA 2018. Mantel test, Liu-Agresti statistics, Cox β, and poly-SIBTEST methods were used for polytomous DIF analysis while Adjacent Category Logistic Regression Model and Cumulative Category Log Odds Ratio methods were used for DSF analysis. This study was carried out by using "differential category combining, focus group sample size, focus group: reference group sample ratio and DIF/DSF detection method". SAS and R software were utilized in the creation of conditions; SIBTEST was used for poly-SIBTEST analysis and DIFAS programs were used for the other methods. Analyses demonstrated that the number of items with large DIF was higher in the small sample according to the polytomous DIF detecting methods. Likewise, the number of steps with large DSF is higher in large samples according to the DSF methods. However, it was found that the methods give more consistent results in large samples. During the steps, the DIF value was lower in the items containing DSF with the opposite sign; therefore, not performing DSF analysis on an item with no DIF may yield erroneous results. Although the differential category combining conditions created within the scope of the research did not have a systematic effect on the results, it was suggested to examine this situation in future studies, considering that the frequency of marking the combined categories differentiated the results.

Keywords: polytomous differential item function, differential step function, adjacent approach, cumulative approach, AC-LOR, CU-LOR

Introduction

Valid measures are needed for test scores to reflect individuals' real scores and for interpretations to display the correct results. Validity, which is an aspect of theory and evidence (American Educational Research Association et al., 2014) that supports interpretations or decisions made based on test scores, is one of the most important features that must exist in measurement tools. Tests should measure all individuals with the same accuracy, regardless of variables unrelated to the measured construct (Sireci & Rios, 2013). It would be misleading to compare different countries or groups with a test that does not mean the same thing for everyone, in other words, when the degree of serving its purpose varies according to groups or countries. In this respect, the property measured by the test items should be invariant according to individuals, groups, and countries. The invariance of the items means that the response probabilities of the items do not change according to the groups with the same characteristics. Item and test bias are the most important threats to validity (Clauser & Mazor, 1998).

*This study is a part of the doctoral thesis prepared by the first author and conducted under the supervision of the second author.

** Research Assistant Dr., Kırşehir Ahi Evran University, Faculty of Education, Kırşehir -Türkiye, yaseminkuzu@yandex.com, ORCID ID: 0000-0003-4301-2645
*** Prof.Dr., Hacettepe University, Faculty of Education, Ankara-Türkiye, sgelbal@gmail.com, ORCID ID: 0000-0001-5181-7262

To cite this article:
Kuzu, Y. & Gelbal, S. (2023). Investigation of differential item and step functioning procedures in polytomus items. *Journal of Measurement and Evaluation in Education and Psychology*, *14*(3), 200-221. https://doi.org/10.21031/epod.1221823
Received: 20.12.2022
Accepted: 9.09.2023

Differential Item Functioning

Detecting the biased items in a test should include, first of all, determining whether the items have a DIF. DIF refers to the fact that the probability of answering an item correctly differs between individuals with the same ability level in different subgroups (Embretson & Reise, 2000; Hambleton et al., 1991). Examination of DIF studies in the literature shows that while dichotomous (two-category) items were studied first, in recent years, detecting DIF has been more common on polytomous items as well as dichotomous items with the widespread use of performance-based evaluation. Unlike dichotomous items, DIF can take different forms in polytomous items due to the number of response categories.

Various DIF detection methods are cited in the literature, and these methods are classified in different ways in different sources (Camilli & Shepard, 1994; Zumbo, 2007; Ellis & Raju, 2003). DIF detection in polytomous items is more complex than DIF detection in dichotomous items. Based on the invariance in polytomous items, the form of invariance may differ in score levels. So that, while invariance cannot be achieved at one score level, it can be achieved at other score levels and in cases where invariance cannot be achieved in the item, DIF can be observed in favor of the reference group at one score level and in favor of the focus group at another score level (Penfield et al., 2008).

Mantel test

The Mantel test statistic, an extension of the Mantel-Haenszel (MH) test, was developed to determine the relationship between matched groups on variables at the ordinal scale level (Mantel, 1963). DIF analysis with the Mantel test includes testing the null hypothesis with statistics on the chi-square distribution at one degree of freedom. In this context, equation of the Mantel test analysis is as follows (Zwick et al., 1993):

$$Mantel\ \chi^2 = \frac{(\sum_k F_k - \sum_k E(F_k))^2}{\sum_k Var(F_k)}$$

The Mantel statistic has a chi-square distribution with one degree of freedom. The rejection of the null hypothesis as a result of this test indicates that the item contains DIF.

Liu Agresti estimator

Although the Liu Agresti estimator is not as common as other MH based methods, it is a recommended method for DIF analysis for polytomous items (Penfield & Algina, 2003). Odds ratios are used in the Liu Agresti estimation.

Cox's β statistic

Cox's β statistic is a mathematically equal but conceptually a different approach to the Mantel test (Cox, 1958) and it assumes that the data come from a decentralized multivariate hypergeometric distribution with β parameter. The β value is calculated as follows (Camilli & Congdon, 1999).

$$\hat{\beta} = \frac{\sum_k \sum_J J(n_{RJk} - \tau_{Jk})}{\sum_k \zeta_k^2}$$

A significant difference in β value from zero means that the item contains DIF.

Poly-SIBTEST

The poly-SIBTEST statistic used for DIF detection in polytomous items is an extension of SIBTEST used in dichotomous items and is a non-parametric model (Chang et al., 1996).

The SIBTEST method presents an effect size (β) that indicates the DIF values as well as the presence of DIF in the item. The estimation of the β effect size, which is defined as the expected group difference in the item thought to have DIF at each valid subtest score level, is defined as follows:

$$\hat{\beta} = \sum_{k=0}^{n_m} p_k(\bar{Y}_{Rk} - \bar{Y}_{Fk})$$

The β effect size index proposed by Roussos and Stout (1996) as the SIBTEST effect size is also used to interpret the poly-SIBTEST DIF index in dichotomous and polytomous items (Henderson, 2001).

DIF detection methods, which are widely used in polytomous items, are based on examining the invariance at the item level (Penfield & Lam, 2000). In approaches such as Mantel's chi-square statistic (Mantel, 1963) and the Generalized Mantel Haenszel (GMH) statistic (Somes, 1986); a single DIF index is given because the general invariance collected at all score levels is measured. In this case, it cannot be determined from which score level DIF originates. Therefore, efforts to identify possible causes of DIF and the item revision process with the contribution of experts are less efficient after the DIF analyses, making it more challenging both in terms of time and economy.

Differential Step Functioning

Differential step functioning is a comprehensive approach used to describe the "between-group difference" in measured properties in a particular step of a polytomous item (Penfield, 2007). Unlike DIF analyses, which give a single statistic for the item, DSF analyses yield as many statistics as the number of steps in the item. So, the differential step functioning can be viewed as a subset of the differential item functioning that focuses on DIF effects in the item.

Evaluation of DSF in a polytomous item begins by dividing the item into J= r-1 step function (where r is the number of score levels in the item). Each step function defines the probability of progressing, or "stepping through", from each score level to a successively higher score level. If there is a difference between the groups in one or more of the step functions of the item, it is concluded that the item exhibits DSF. DSF analysis can be performed by using different approaches. Logistic regression (French & Miller, 1996) and IRT-based approaches such as Graded Response Model (GRM) (Cohen et al., 1993) and the Partial Credit Model (PCM) (Penfield et al., 2008) can be given as examples. In this study, DSF detection was done with the most common DSF methods used in the literature: Adjacent Category Logistic Regression Model (AC-LOR) and Cumulative Category Log Odds Ratio (CU-LOR) methods.

For this purpose, Penfield's (2008) probability ratio approach was used which compared the probability of success of the focus and reference group members with the same observed score at step j. Accordingly, the test takers are divided into score groups according to the raw total scores of a test with possible score values k = 1, 2, 3,…., K. In this context, the ratio of the probability of success of the reference group at step j to the probability of success of the focus group is calculated as follows:

$$\hat{\alpha}_j = \frac{\sum_{k=1}^{K} A_{jk} D_{jk} \,/\, N_{jk}}{\sum_{k=1}^{K} B_{jk} C_{jk} \,/\, N_{jk}}$$

A_{jk}: Number of reference group members who succeeded in step j.

B_{jk}: Number of reference group members who failed in step j.

C_{jk}: Number of focus group members who succeeded in step j.

D_{jk}: Number of focus group members who failed in step j.

N_{jk}: $A_{jk} + B_{jk} + C_{jk} + D_{jk}$

This value is equivalent to the Mantel-Haenszel probability ratio for dichotomous items and each step is considered as a dichotomous item (Gattamorta & Penfield, 2012). The natural logarithm of $\hat{\alpha}_j$is denoted by $\hat{\lambda}_j$. $\hat{\lambda}_j$ with a value of zero means no DSF, a negative $\hat{\lambda}_j$ value means DSF in favor of the focus group, and a positive $\hat{\lambda}_j$ value means that DSF exists in favor of the reference group.

Adjacent Category Approach

When performing DSF analysis on polytomous items, each of the J step functions is defined using the adjacent category approach, which is consistent with Generalized Partial Credit Model (GPCM). Under this approach, j. step function expresses the probability of successfully progressing from the j-1 score level to the j score level.

Cumulative Category Approach

When performing DSF analysis on polytomous items, each of the J step functions is defined using the cumulative category approach, which is consistent with GRM. Under this approach, j. step function indicates the probability of successfully progressing from 0, 1, …, j-1 score level to j, …, J score level. Therefore, in the DSF analysis under the cumulative approach, all scores are taken into account in total, unlike the adjacent category approach. Therefore, it is very important to know the approach used to define the step function in the interpretation of step level parameters.

DSF analyses are an important component of a comprehensive DIF analysis for polytomous items. In recent years, researchers have argued by citing many reasons that each score level should be taken into account instead of a single total score level while examining the invariance form in polytomous items (Gattamorta & Penfield, 2012). One of these reasons is that many omnibus DIF methods such as the poly-SIBTEST and the Standard Mean Difference (SMD) show relatively low power when the DSF effect changes in sign or values in steps of a polytomous item (Penfield & Algina, 2003; Wang & Su, 2004). The second reason is related to the fact that the omnibus DIF methods give a value representative of the DSF aggregated across all steps, and thus large values of DSF at certain steps may be missed if only one step has a large amount of DSF or if the DSF is of opposite sign across the steps. Therefore, calculating the DSF for each step will allow important information to be noticed and taken into account. Finally, with such an approach, it will be possible to understand which score levels are responsible for the violation of invariance, and thus, information about the possible causes of DIF will be obtained.

Examination of the studies in which DIF and DSF analyses are performed in conjunction shows that they are rather limited. The statistics in the studies were undertaken mostly on simulation data, and when real data were used, the focus was usually on the current situation (Akour et al., 2015; Ayodele, 2017; Benítez et al., 2015; Gattamorta & Penfield, 2012; Miller et al., 2010; Penfield, 2007; Penfield et al., 2008; Penfield, 2008; Penfield, 2010). This study aimed to compare the DIF and DSF detection methods in polytomous items by manipulating the conditions on real data, and in line with this purpose, answers were sought to the following questions.

1. Do the DIF values obtained by polytomous DIF methods change based on differential category combining and focus group:reference group (F:R) sample ratios when the focus group sample size is 200 (small)?
2. Do the DIF values obtained by polytomous DIF methods change based on differential category combining and F:R sample ratios when the focus group sample size is 1000 (large)?
3. Do the DSF values obtained by DSF methods change based on differential category combining and F:R sample ratios when the focus group sample size is 200 (small)?
4. Do the DSF values obtained by DSF methods change based on differential category combining and F:R sample ratios when the focus group sample size is 1000 (large)?
5. Do the DIF values obtained by polytomous DIF methods with differential category combination rule and F:R sample ratios differ according to sample size?
6. Do the DSF values obtained by DSF methods with differential category combination rule and F:R sample ratios differ according to sample size?
7. In terms of DSF, how are the similarity rates in classifying the item steps of the methods according to the sample sizes of the focal group?

Methods

This research conducted with correlational survey model compared the polytomous DIF/ DSF detection methods on the items taken from PISA 2018 under various conditions.

Study Group

The sample of the research included items related to the frequency of digital device use at school (IC011) within the scope of the "ICT Familiarity Questionnaire" in PISA 2018, which was used for students from Kazakhstan, Turkey and the United States of America (USA). In selecting the countries, firstly, the country rankings in the field of reading skills (weighted area) were examined according to the results of PISA 2018, and the countries were divided into three groups as low, medium, and high level. Considering the fact that the relevant survey was not applied to all of the countries participating in PISA 2018, two conditions (success and economic level) were taken into account in addition to answering this survey in selecting the countries. Therefore, Kazakhstan (69th), a non-OECD country, was selected from the low-level group, Turkey (40th), an OECD country, was selected from the middle-level group and the USA (13th), an OECD country, was selected from the high-level group. This study included the results obtained from comparing Turkey-Kazakhstan, which better reflect the results, in order to ensure that the text would be concise and more precise. The results of the Turkey-USA comparison are included in Kuzu (2021).

Data Collection

The research data were obtained from the official internet address of the OECD (https://www.oecd.org/pisa/data/2018database) where the PISA 2018 data were announced. In this context, the data of Kazakhstan and Turkey, for which the "ICT Familiarity Questionnaire" was answered within the scope of PISA 2018, was studied. The questionnaire includes items related to digital media and digital devices such as desktop computers, laptops, smartphones. The questionnaire consists of different sections, such as the possibility of accessing digital tools at home/school or the time allotted to digital devices. In this study, 10 items -5-point Likert type- related to the frequency of use of digital devices in school (IC011) were examined. As a result of expert opinions, it was decided that the items measured the same dimension and could be summed. The scores obtained from the questionnaire varied

between 10 and 50; high scores meant that the frequency of using digital devices at school was high while low scores meant that the frequency of using digital devices at school was low. Table 1 presents the descriptive statistics and score category distributions for each item on the basis of countries.

Table 1

Descriptive Statistics and Score Category Distributions for the Items in the Data Collection Tool

							Score Category Distributions (%)				
Item	Country	$\bar{\chi}$	Sd	Kurtosis	Skewness	Item-Total Correlation	1	2	3	4	5
I1	KAZ	2.71	1.43	-1.30	.20	.63	29.7	17.1	20.0	19.0	14.2
	TUR	1.78	1.16	.41	1.27	.43	61.9	13.2	13.1	8.6	3.3
I2	KAZ	2.38	1.31	-.92	.52	.79	35.8	20.6	21.6	13.7	8.2
	TUR	1.39	.74	2.19	1.79	.58	74.5	13.9	9.9	1.8	
I3	KAZ	2.79	1.30	-1.06	.11	.79	21.9	19.5	27.6	19.3	11.6
	TUR	2.20	1.15	-.64	.57	.52	37.3	22.8	25.9	10.3	3.6
I4	KAZ	2.54	1.32	-1.06	.33	.84	30.3	20.1	24.0	16.4	9.2
	TUR	1.56	.88	1.27	1.47	.65	66.0	16.9	13.1	3.5	.5
I5	KAZ	2.22	1.29	-.73	.69	.78	42.6	18.9	19.7	12.0	6.8
	TUR	1.36	.73	2.82	1.97	.64	77.5	11.3	9.4	1.9	
I6	KAZ	2.17	1.27	-.66	.72	.77	43.5	19.1	19.9	11.5	6.0
	TUR	1.34	.69	2.87	1.97	.60	77.4	12.5	8.9	1.3	
I7	KAZ	2.62	1.27	-.99	.24	.82	25.5	21.3	27.2	17.1	8.8
	TUR	1.96	1.13	-.23	.89	.49	48.8	19.9	20.5	8.0	2.8
I8	KAZ	2.49	1.30	-1.00	.38	.84	31.4	20.7	24.0	15.5	8.5
	TUR	1.52	.86	1.98	1.63	.59	67.9	16.7	12.0	2.7	.7
I9	KAZ	2.54	1.30	-1.03	.34	.84	29.6	20.9	24.4	16.2	8.8
	TUR	1.51	.86	2.31	1.69	.64	68.1	17.0	11.4	2.6	.9
I10	KAZ	2.55	1.32	-1.06	.33	.84	29.7	20.7	23.7	16.4	9.5
	TUR	1.78	1.04	.66	1.21	.61	55.4	20.8	16,5	5,0	2,3

According to Table 1, the highest mean for all countries was obtained in item 3 ($\bar{\chi}_{KAZ} = 2.79$, $\bar{\chi}_{TUR} = 2.20$) and the lowest mean for all countries was obtained in item 6 ($\bar{\chi}_{KAZ} = 2.17$, $\bar{\chi}_{TUR} = 1.34$) in the "ICT Familiarity Questionnaire" in the items related to the frequency of using digital devices at school. However, it was found that the item means were mostly above 2 for the Kazakhstan data and below 2 for the Turkey data. In this case, it can be argued that the students who participated in PISA 2018 from Turkey had a low level of digital device use at school. On the other hand, examination of the score category distributions of the items shows that more than half of the data for Turkey was concentrated in the 1st category in the majority of the items, whereas specifically the 4th and 5th categories were marked less. It is noteworthy that, the 5th category was not marked at all in items 2, 5, and 6 and the ratio of students who marked the 5th category in items 4, 7, 8, and 9 was below 1%. When the Kazakhstan data

was examined, it was found that the distribution spread to all category levels. For both countries, the 1st category was marked the most and the 5th category the least.

Dimensionality

Exploratory factor analysis was performed to examine the dimensionality of the scale. The sample size of the country data was determined by the Kaiser-Meyer-Olkin (KMO) coefficient and the distribution of the data was checked with the Bartlett Test of Sphericity. The KMO coefficients for country data ranged between .87-.94. According to Kaiser (1970), the value of KMO takes a value between 0 and 1, and when this value approaches 1, it means that the sample size is suitable for factor analysis. On the other hand, when the results of the Bartlett Test of Sphericity were examined, the chi-square value was found to be statistically significant for all two countries ($\chi^2_{KAZ(45)} = 101454.496$, $\chi^2_{TUR(45)} = 16161.504$; p<,01) and therefore, the data were suitable for exploratory factor analysis. In this context, the results of the exploratory factor analysis are presented in Table 2.

Table 2

Factors Obtained as a result of Exploratory Factor Analysis and Amount of Variance Explained

			Eigenvalue	% of variance
IC011	KAZ	Factor 1	7.01	70.09
	TUR	Factor 1	6.279	62.791

The result of the exploratory factor analysis demonstrated a single component with an eigenvalue above 1 for the Kazakhstan and Turkey data, therefore it was unidimensional. Table 3 presents the results regarding the factor loadings of the items.

Table 3

Factor Loading Values for Items Found via Exploratory Factor Analysis

	Item	Factor Loading	
		KAZ	TUR
IC011	I1	.69	.66
	I2	.83	.79
	I3	.82	.70
	I4	.88	.85
	I5	.83	.86
	I6	.82	.83
	I7	.86	.72
	I8	.88	.84
	I9	.88	.85
	I10	.88	.80

Table 3 presents the factor loading values obtained for the items as a result of the exploratory factor analysis. In general, factor loading values varied between .66 and .88.

Items examined within the scope of the research: Ayodele (2017) developed a 20-item test and analyzed the research questions by manipulating 2 items. In this study, three items were chosen to be interpreted due to the high number of research conditions. Psychometric properties were taken into account in the

selection of the items, and the items with the highest item-total correlation for the two countries were selected because they had the highest representative power in the scale. Table 4 shows that the items with the highest item-total test correlations for both countries were Items 4, 9, and 10. In this context, polytomous DIF and DSF analysis results for Item 4, Item 9, and Item 10 were reported and interpreted. The results of the research are limited to the data, methods and conditions used in the research.

Conditions that were examined in this study

This section presents the conditions manipulated in the research.

Category combining rule. First of all, items that were currently coded in the 5-point scale type (1-5) were coded as (0-4) in accordance with the working principles of the DIFAS 5.0 program (Penfield, 2013). Since the aim was to change the number of item categories, afterwards, the categories were combined. All possible combinations in category combination were taken into account, paying attention to the fact that the combined categories were adjacent (Gelin & Zumbo, 2003; Göçer-Şahin et al., 2016). Table 4 presents the category combining conditions created for the purpose of this research.

Table 4

Category Combination Conditions Created within the Scope of the Research Goal

		Before Recoding	New categories	Explanation
three-category	1st condition (C1)	(1,2)	0	(1 and 2) and (4 and 5) merged.
		3	1	
		(4,5)	2	
	2nd condition (C2)	1	0	(2 and 3) and (4 and 5) merged.
		(2,3)	1	
		(4,5)	2	
	3rd condition (C3)	(1,2)	0	(1 and 2) and (3 and 4) merged.
		(3,4)	1	
		5	2	
four-category	4th condition (C4)	(1,2)	0	(1 and 2) merged.
		3	1	
		4	2	
		5	3	
	5th condition (C5)	1	0	(2 and 3) merged.
		(2,3)	1	
		4	2	
		5	3	
	6th condition (C6)	1	0	(4 and 5) merged.
		2	1	
		3	2	
		(4,5)	3	
	7th condition (C7)	1	0	(3 and 4) merged.
		2	1	
		(3,4)	2	
		5	3	
five-category	8th condition (C8)	1	0	It has been recoded due to the conditions of the DIFAS 5.0 program.
		2	1	
		3	2	
		4	3	
		5	4	

According to Table 4, a total of eight category combination conditions were obtained in the analysis of the data: three for three-category data, four for four-category data, and one for five-category data.

Sample size and focus group-reference group sample ratio. Another condition examined in the study was the focus group sample size. Sample size is very important in DIF studies. If the sample size is too small, it leads to poor parameter estimation, thus no DIF and if the sample is too large, it may cause hypersensitivity in DIF detection (Ayodele, 2017). For this reason, this study aimed to make the right decision via working with different sample sizes. Examination of the studies conducted with polytomous items demonstrated that the studies were performed with data from at least 440 individuals (40 focus group-400 reference group) while the common approach was to use data of 100 to 2000 people (Ankenmann et al., 1999; Elosua & Wells, 2013; Gonzalez-Roma et al., 2006; Meade & Lautenschlager, 2004; Wood, 2011). The focus group sample in this study was addressed had two different sizes: 200 (small) and 1000 (large). However, (focus group): (reference group) sample ratios were examined in three conditions as 2:1, 1:1, and 1:3. In this case, while the sample size of focus group was 200, the sample size of reference group was 100, 200 and 600; while the sample size of focus group was 1000, the sample size of reference group was 500, 1000 and 3000.

Polytomous DIF/ DSF detection methods. Mantel test, Liu Agresti, Cox's β, and poly- SIBTEST were used to determine DIF while AC-LOR and CU-LOR analyses were performed as DSF detection methods.

Data Analysis

Polytomous DIF Analyses

DIFAS 5.0 program (Penfield, 2013) was used for the Mantel test, Liu Agresti estimation, and Cox's β statistics from among polytomous DIF detection methods. First of all, the data were re-coded to start from 0 as the smallest value in accordance with the operating principles of the relevant program (1=0, 2=1, 3=2, 4=3, 5=4) and the total score was used as the matching variable in the analyses. A research design with 8*2*3*4 = 192 cells was created for polytomous DIF analysis including category combining rules (8), focus group sample size (2), focus group: reference group sample ratio (3), and DIF detection method (4). For interpretation of the Mantel test results, the critical value for Type I error probability at the .01 level was accepted as 6.63. On the other hand, while interpreting the Liu Aggresti statistic, the standardized Liu Aggresti Cumulative Common Log-Odds Ratio (LOR Z) value in the analysis outputs was used. If this value is greater than 2 or less than -2, DIF is present in the item. A positive Liu Agresti statistic points to the existence of DIF in favor of the reference group while a negative statistic points to the existence of DIF in favor of the focus group. Another statistic obtained from DIFAS program outputs in this study was Cox's β statistics. If the Cox Z value, which is obtained by dividing the Cox's β table value by its standard error, is greater than 2 or less than -2, DIF is present in the item. If this value is positive, the existence of DIF works in favor of the reference group, and if it is negative, the existence of DIF works in favor of the focus group (Penfield, 2013).

The last DIF analysis was performed with the poly-SIBTEST method for polytomous items. While interpreting the results of the analysis conducted by using the SIBTEST program, the β value was taken into consideration and the values $|\beta| \geq 0.088$ were marked as DIF (C level) (Roussos & Stout, 1996).

DSF Analyses

DIFAS 5.0 program was used for CU-LOR and AC-LOR statistics to determine whether the items had DSF. The calculated DSF values for each step of each item were examined. In this context, the $\hat{\lambda}_j$ values obtained from the analysis outputs were interpreted and the items showing large DSF in the steps were marked separately for both methods. The $|\hat{\lambda}_j| > 0.64$ criterion was taken into account for marking items with large DSF (Penfield, 2007; Penfield et al., 2008).

The findings section presents the results of the polytomous DIF and DSF analyses for the selected items with the help of graphics. Critical values of each method are indicated with dashed lines to facilitate the

interpretation of the graphs. In this context, in addition to the critical values indicated with dashed lines in Cox's β, the Liu Agresti, and poly-SIBTEST, the values above the line in the Mantel test statistic point to large DIF. Similarly, the values outside the critical values presented by the dashed lines for the DSF analyses indicate that the item step exhibits large DSF under the relevant conditions. In the DIF and DSF graphs, the DIF/ DSF level increases as you move away from the critical values.

Results

Findings Related to Research Question 1

Figure 1 presents the results obtained according to polytomous DIF methods (The Mantel test, the Liu-Agresti statistics, Cox's β, poly-SIBTEST) under varying conditions when the focus group sample size was 200 (small).

Figure 1

The change in the DIF values in the items when the focus group sample size was 200.

The examination of the change in the DIF values in the items in Figure 1 showed that Item 4 did not show large DIF for almost all sample size ratios and under all conditions according to the Mantel test, the Liu-Agresti and Cox β methods and DIF values were below critical values. According to the poly-SIBTEST method, while the first two sample size ratios showed negative large DIF in the last conditions, large DIF was not observed in other conditions and with 1:3 sample size ratio. The examination of Item 9 showed that values close to the critical value were obtained in the first two sample size ratios when

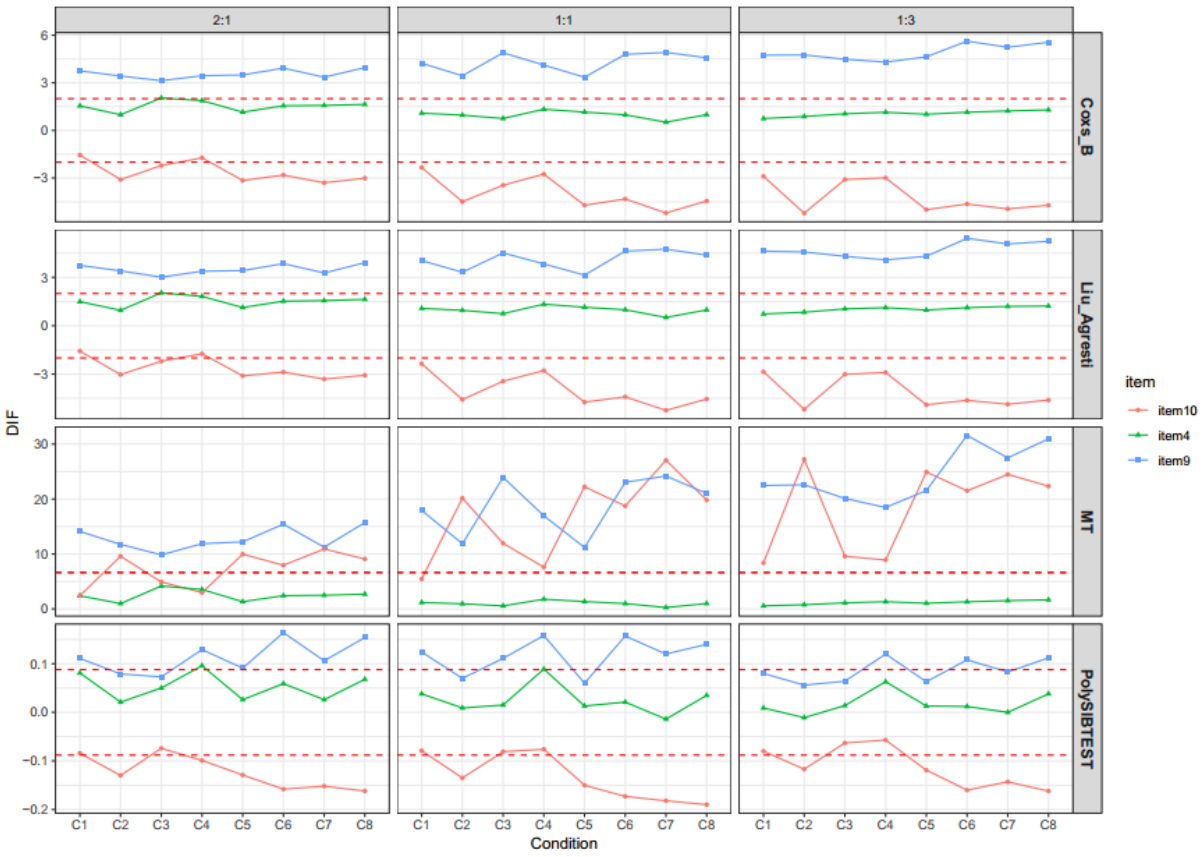
methods other than the poly-SIBTEST method were used. While the sample size ratio was 1:3, there was an increase in the DIF values calculated for the item and a positive large DIF was obtained in all methods and conditions. Finally, the examination of Item 10 showed that the DIF values obtained in the Cox's β and Liu Agresti methods were around the critical value and higher DIF values were obtained at a sample size of 1:3. In the Mantel test and poly-SIBTEST methods, the sudden increase in C5, C6, C7 and C8 conditions was remarkable, especially at the sample size ratio of 1:1. However, these changes did not show a systematic pattern on the basis of conditions.

Findings Related to Research Question 2

Figure 2 presents the results obtained according to polytomous DIF methods (the Mantel test, the Liu-Agresti statistics, Cox β, poly-SIBTEST) under varying conditions when the focus group sample size was 1000 (large).

Figure 2

The change in the DIF values in the items when the focus group sample size was 1000.



The change in the DIF values in the items in Figure 2 was examined. It was observed that Item 4 did not indicate large DIF almost with all sample size ratios and under all conditions for all methods, and DIF values were found to be below the critical values. The examination of Item 9 showed that the first two sample size ratios exhibited a very large DIF in all methods, except for the poly-SIBTEST method. The DIF values obtained were found to increase positively as the sample size ratio increased. In the poly-

SIBTEST method, there were conditions below the critical value as well as large DIF values. Finally, the examination of Item 10 demonstrated that the DIF values obtained in the Cox's β and the Liu Agresti methods were below the critical value in the C1 and C4 conditions at a sample size ratio of 2:1, but exhibited large DIF in all other conditions, with the highest DIF values at the sample size of 1:3. In the poly-SIBTEST method, while large DIF was obtained in some conditions, the values obtained under some conditions were below the critical value. Similar situations were obtained in general based on the sample size ratios.

Findings Related to Sub-Problem 3

Figure 3 presents the results obtained according to the DSF methods (AC-LOR, CU-LOR) under varying conditions when the focus group sample size was 200 (small).

Figure 3

The change in the DSF values in the item steps when the focus group sample size was 200.

When the change in the DSF values in the item steps in Figure 3 was examined, it was seen that large DSF values were obtained in the positive direction in the 1st step of Item 4 under some conditions. The values of DSF obtained from the AC-LOR method were mostly higher than the values obtained from the CU-LOR method. Large DSF values were observed in the negative direction in the other steps of Item 4. Large DSF values were obtained for all conditions and sample ratios in Step 1 of Item 9. The examination of Step 2 showed that the DSF values were below the critical value in all sample ratios and almost all conditions based on the AC-LOR method while positive large DSF was obtained especially in C6, C7, and C8 conditions in the CU-LOR method. While the sample size ratio was 1:2 in Step 3,

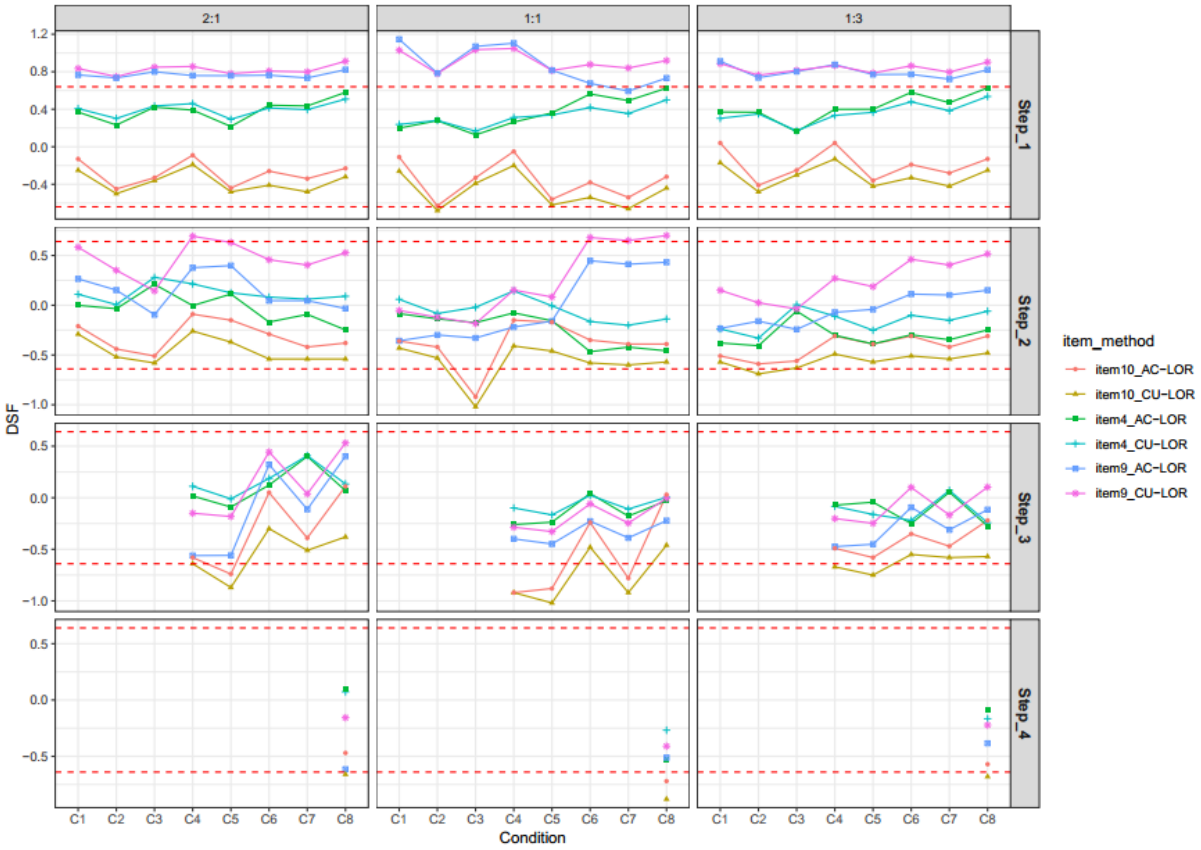
negative DSF was observed in some conditions, while these values remained below the critical value in other sample size ratios. When Step 4 was examined, it was seen that it exhibited higher DSF in the negative direction compared to the AC-LOR method. Finally, the examination of the Steps in Item 10 showed that large DSF was not obtained in the Step 1, except for some conditions where the sample size ratio was 1:2. In Steps 2 and 3, large DSF was obtained mostly in the negative direction. The values of DSF obtained from the CU-LOR method were higher than the values of DSF obtained from the AC-LOR method. Large DSF was observed in all sample size ratios according to the CU-LOR method in Step 4. Large DSF was not obtained with sample size ratios 1:1 and 1:3 with the AC-LOR method.

Findings Related to Research Question 4

Figure 4 presents the results obtained according to DSF methods (AC-LOR, CU-LOR) under varying conditions when the focus group sample size was 1000 (large).

Figure 4

The change in the DSF values in the item steps when the focus group sample size was 1000.



The change in the DSF values in the item steps in Figure 4 was examined and it was seen that large DSF was not observed in the 1st Step of Item 4, except for the C8 condition. DSF values obtained from the AC-LOR method were higher in some conditions while DSF values obtained from the CU-LOR method were higher in other conditions. The DSF values calculated in the other item steps were below the critical value. In Step 1 of Item 9, large DSF values were obtained in almost all conditions and sample ratios. The examination of Step 2 showed that DSF values were below the critical value in all sample ratios and under all conditions in the AC-LOR method; on the other hand, large DSF was obtained in the

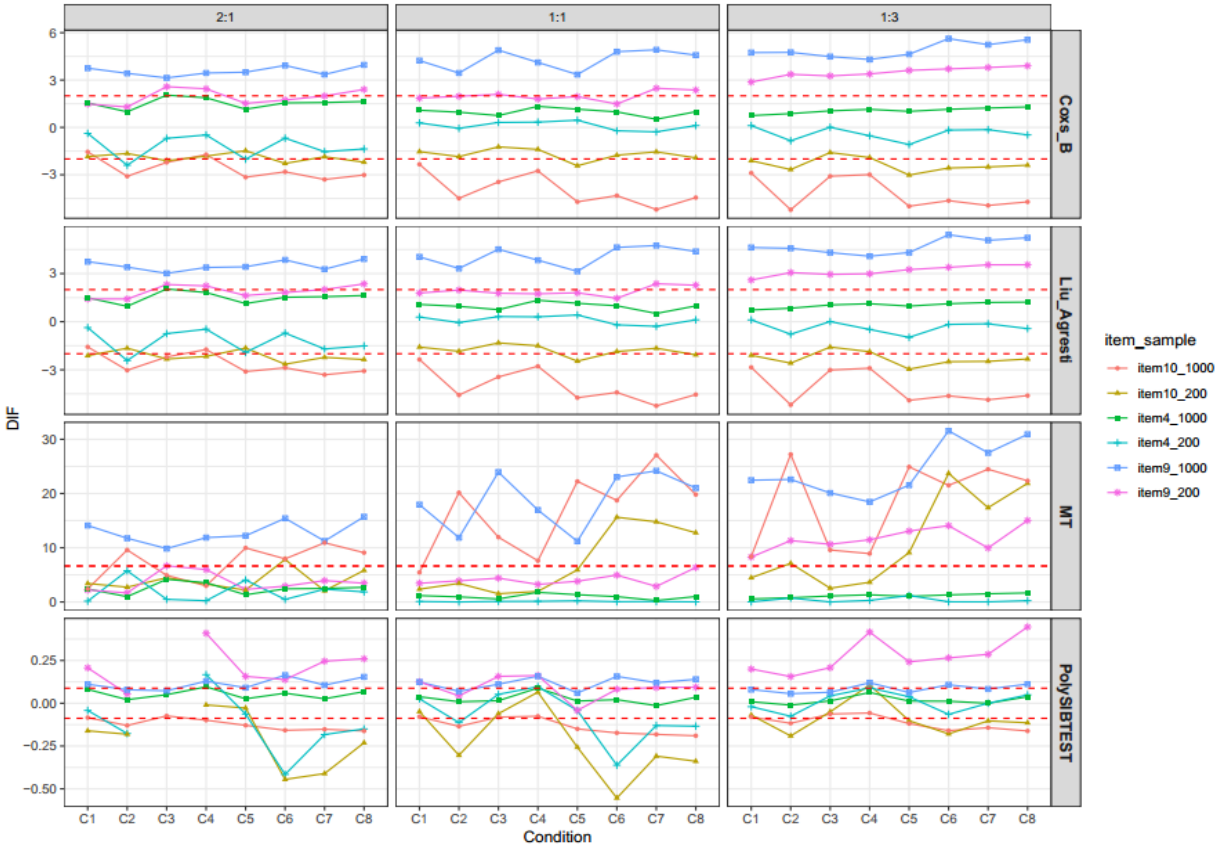
positive direction in some conditions in the CU-LOR method. The DSF values calculated in all conditions and sample size ratios in Steps 3 and 4 were below the critical values. Finally, the examination of the steps in Item 10 demonstrated that large DSF values were not obtained in the 1st Step in general. In Steps 2 and 3, large DSF was obtained in the negative direction under some conditions. The values of DSF obtained from the CU-LOR method were higher than the values of DSF obtained from the AC-LOR method. In step 4, large DSF was observed in all sample size ratios according to the CU-LOR method. Large DSF was not obtained in the AC-LOR method when the sample size ratio was 2:1 and 1:3.

Findings Related to Research Question 5

Figure 5 presents the DIF values obtained by the polytomous DIF methods with differential category combination rule and F: R sample ratios based on the focus group sample size.

Figure 5

The change in the DIF values in items based on focus group sample size



The examination of the change in the DIF values in the items according to the focus group sample size in Figure 5 showed that the DIF values obtained from the large and small samples differed in the opposite direction in item 4, especially according to the Cox's β and the Liu Agresti methods. Accordingly, Item 4 tended to exhibit negative DIF in the small sample, while it exhibited positive DIF in the large sample. When the DIF values related to Item 9 were examined, it was found that the DIF values calculated for both sample sizes were positive, and the DIF values calculated in the large sample were generally large. Unlike other methods, higher DIF values were obtained in the small sample in the poly-SIBTEST method. The DIF values calculated in the small sample at 2:1 and 1:1 sample size ratios were mostly

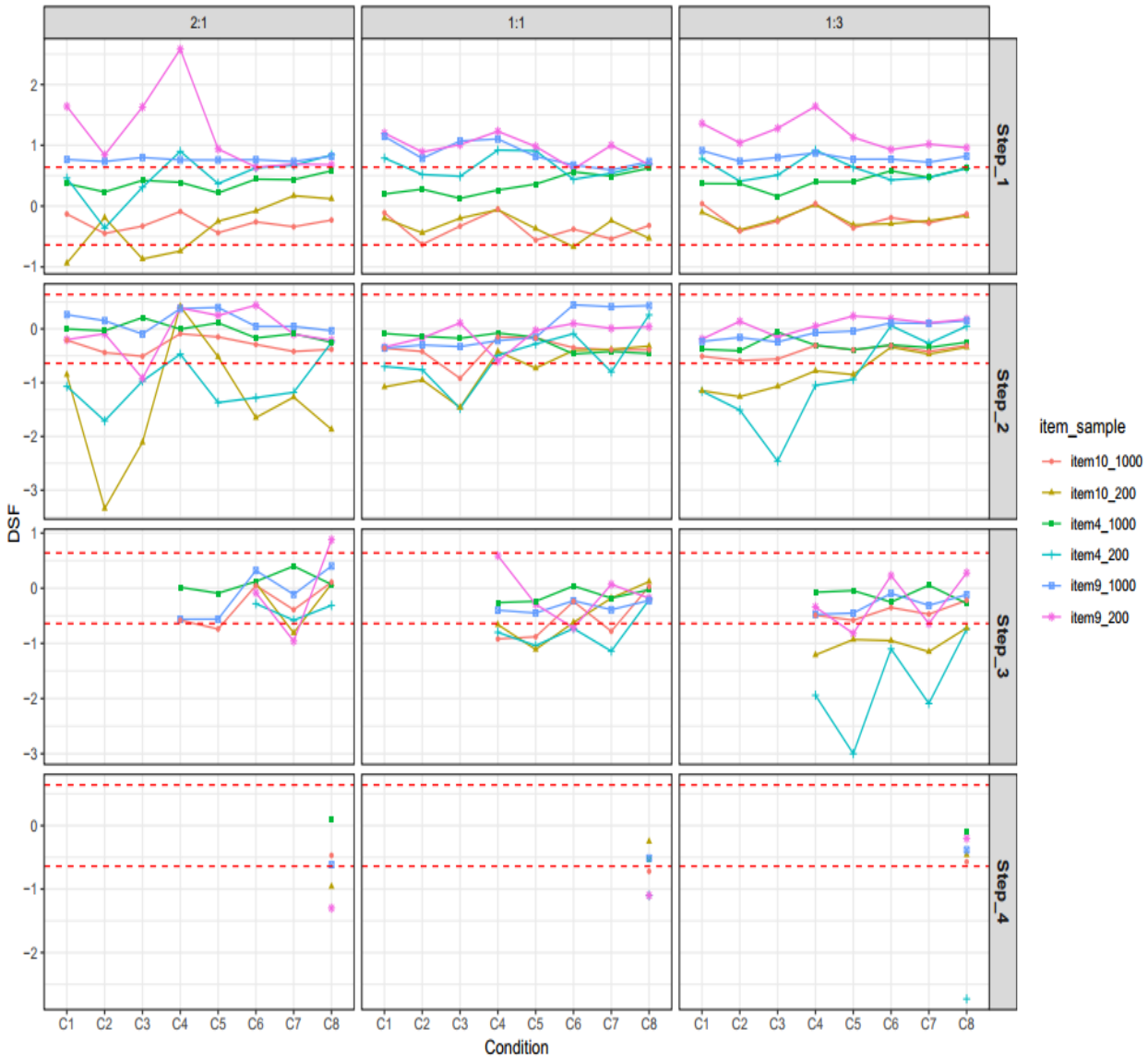
below the critical value in the Cox's β, the Liu Agresti and the MT methods. DIF values calculated on the basis of the conditions did not show a systematic pattern.

Findings Related to Research Question 6

Figure 6(a) presents the DSF values obtained by the AC-LOR method with differential category combination rule and F:R sample ratios based on the focus group sample size.

Figure 6(a)

The change in the DSF values in item steps based on focus group sample size (AC-LOR)

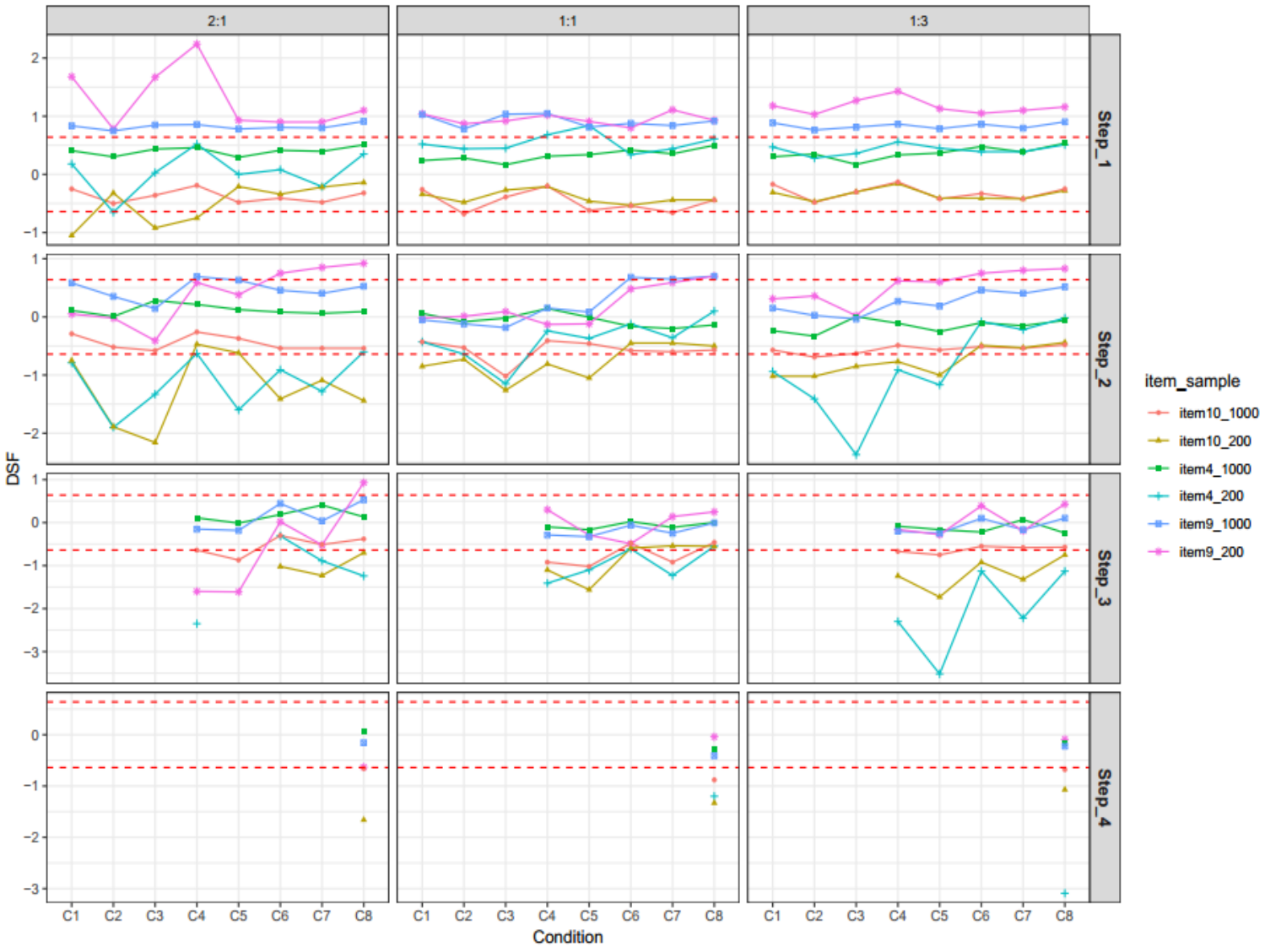


In Figure 6(a), the examination of the change in the DSF values in the item steps according to the focus group sample size based on the AC-LOR method showed that higher DSF was obtained in cases where the focus group sample size was small in Step 1 of Item 4. Similarly, in other steps, large DSF values were obtained when the focus group sample was small. In the steps of Item 9 and Item 10, large DSF was observed in the small sample in general while DSF values were below critical values in the large sample. There was no systematic pattern on the basis of the conditions.

Figure 6(b) presents the DSF values obtained by the CU-LOR method with differential category combination rule and F:R sample ratios based on the focus group sample size.

Figure 6(b)

The change in the DSF values in item steps based on focus group sample size (CU-LOR)



In Figure 6(b), the examination of the change in the DSF values in the item steps according to the focus group sample size based on the CU-LOR method showed that the DSF values calculated in the small and large samples in the 1st and 4th Steps of Item 4 were mostly below the critical value. In the other steps, while large DSF was mostly not observed in the small sample; quite large DSF values in the large sample drew attention. Similar results were obtained in the large and small samples in the steps of Item 9 and Item 10, but slightly higher DSF values were obtained in the small sample. A systematic pattern was not obtained on the basis of the conditions.

Findings Related to Research Question 7

Table 5 presents the similarity ratios of the methods in classifying the item steps in terms of DSF according to the sample sizes.

Table 5

The Similarity Ratios of the Methods in Classifying the Item Steps in Terms of DSF

			Amount of DSF (CU-LOR)							
			Sample size of focal group: 200				Sample size of focal group: 1000			
	Item		Large DSF	Other	Total	Similarity (%)	Large DSF	Other	Total	Similarity (%)
Amount of DSF (AC-LOR)	I4	Large DSF	32	13	45	79.55	-	1	1	98.86
		Other	5	38	43		-	87	87	
		Total	37	51	88		-	88	88	
	I9	Large DSF	32	8	40	77.27	31	-	31	94.32
		Other	12	36	48		5	52	57	
		Total	44	44	88		36	52	88	
	I10	Large DSF	36	1	37	86.36	5	-	5	88.64
		Other	11	40	51		10	73	83	
		Total	47	41	88		15	73	88	

Table 5 provides the number of steps that were marked/unmarked by the AC-LOR and CU-LOR methods as exhibiting large DSF based on sample size. When Turkey-Kazakhstan comparison was evaluated for all items and conditions, it was found that 136 (25.76%) steps showed large DSF based on both methods and it was determined that there was no large DSF compared to both methods in 326 (61.74%) steps. However, in 43 (8.14%) steps, large DSF was detected compared to the CU-LOR method although the AC-LOR method did not mark these steps as large DSF. Likewise, large DSF was calculated according to the AC-LOR method in 23 steps (4.36%) which were marked as without large DSF by the CU-LOR method. The similarity rates in classifying the item steps of the methods in terms of DSF in this comparison changed from 77% to 86% for the focus group with sample size of 200, while they ranged from 89% to 99% for the focus group with sample size of 1000. Therefore, it can be argued that the similarity rates in classifying the item steps of the methods in terms of DSF were higher in the large sample.

Discussion and Conclusion

In this section, the results pertaining to the research problems were discussed in conjunction with the related literature.

Examination of polytomous DIF detection methods (Cox's β, Liu Agresti, MT, and poly-SIBTEST) based on sample size and conditions

The examination of the results obtained from the DIF detection methods shows that the DIF values obtained from Cox's β, the Liu Agresti and MT methods were quite similar to each other in the small sample, while the DIF values obtained from the poly-SIBTEST method differed from the other methods. Among these methods, the poly-SIBTEST helped to detect the highest number of conditions that exhibited large DIF. Although compatible with other methods, the poly-SIBTEST method was found to be the method to detect the items that exhibited the most DIF and provided more sensitive results compared to other methods (Henderson, 2001; Mellor, 1995). It can be argued that the results obtained from the four methods were closest to each other when the sample size ratio was 1:3. However, it was stated that the DIF determination power of the methods tended to decrease with the increase in the sample size of the reference group versus the sample size of the focus group. And It was stated that Type

I error tends to increase in cases where the sample sizes of the reference and focus groups are equal (Wang & Su, 2004; Zwick, 2012)

When the results obtained from DIF detection methods were analyzed in terms of focus group sample size, it was found that all methods provided parallel results when the sample size increased. The highest DIF values and the variability in these values on the basis of the conditions were obtained when the sample size ratios were 1:2 and 1:3. There are studies reporting that the statistical power ratios of the tests are highly affected by the sample size (Bolt, 2002; Kristjansson et al., 2005). Accordingly, it is stated that the methods have a higher statistical power ratio as the sample size increases (Yandı, 2017). When the DIF values obtained from the methods in this study were examined, it was found that the amount of large DIF was higher in the large sample, while the DIF values of the items in the small sample were mostly below the critical values. However, it was observed that the methods provided more consistent results in a large sample.

Examination of DSF detection methods (AC-LOR and CU-LOR) according to sample size and conditions. A comparison of the AC-LOR and CU-LOR methods demonstrated that the DSF values obtained from the AC-LOR method in Steps 1 and 2 for Item 4 were higher than the DSF values obtained from the CU-LOR method. In the other steps of Item 4, the results obtained from the CU-LOR method were found to be higher. On the other hand, the examination of Item 9 demonstrated that the results obtained from the AC-LOR method in some conditions and the CU-LOR method in some conditions were higher in the first two steps, so there was no significant difference between the methods on the basis of the conditions. However, the values of DSF obtained from the CU-LOR method were higher in the other steps of Item 9 and all steps of Item 10. In their study comparing these two methods, Gattamorta and Penfield (2012) stated that there are more steps that exhibit medium to large DSF only according to the effect size in the AC-LOR method used in the adjacent categories approach. When analyzed according to both effect size and significance tests, it was seen that the number of steps exhibiting significant DSF was higher than the CU-LOR method used under the cumulative approach. Due to the smaller standard errors obtained with the CU-LOR method, it was stated that the results were more likely to be statistically significant compared to the AC-LOR method. On the other hand, due to the use of responses from all steps in the cumulative approach, the CU-LOR statistic has higher power than the AC-LOR statistic, which only uses responses in adjacent categories (Ayodele, 2017).

When the DSF detection methods were examined according to sample sizes, it was seen that the DSF values obtained from both methods were higher when the sample was small compared to the large sample. While the same items (Item 4, Item 10) contained half and half DSF in the small sample; they exhibited almost no large DSF in the large sample. On the other hand, the similarity rates in the classification of the item steps of the methods in terms of DSF were higher in the large sample. It clearly shows the importance of the methods used, especially in small samples, when interpreting the invariance and ultimately deciding on the revision or removal of the item.

When the classifications were examined regarding whether the item steps contained large DSF on the basis of methods, it was quite remarkable to note that the similarity rates of the methods were much higher in the large sample. Especially when the sample size was 1000, the percentages of agreement of the methods in the DSF classification made with the CU-LOR and AC-LOR methods of Item 4 and Item 9 were quite high (99% and 94%). Therefore, it can be argued that the methods generated very consistent results, especially in the large sample, in classifying the items in terms of DSF. Parallel to this result, it has been stated in the literature that although the AC-LOR method provides higher DSF values in other DSF classifications, except for small DSF, both methods mostly generate consistent results (Gattamorta, 2009).

When the results of the methods were analyzed on the basis of sample size ratios and conditions, an increase was observed in the DSF values for some items at the same sample size, while a decrease was observed in the DSF values for some items. Therefore, it can be argued that sample size ratios did not have a significant effect on the results of DSF. On the other hand, although the examined conditions did not significantly affect the results, there were fluctuations in the results obtained from the AC-LOR method as the conditions changed. The results show parallelism on the basis of conditions in the CU-

LOR method. In the literature, it is stated that the DSF values estimated under the cumulative approach are more stable than the DSF values estimated under the adjacent categories approach (Gattamorta & Penfield, 2012; Penfield, 2008). It was found that the pattern of the number of steps on the DSF results was not systematic in both methods, whether stable or not. Ayodele (2017) reached similar results and stated that the sample size ratio and the number of steps did not have a statistical and practical significance on the DSF values. Therefore, if the data is polytomous, using the data in its raw form without any changes in the data will produce more valid results. However, if category combining will be used for various reasons, it is recommended to combine categories in accordance with the nature of the research and the data, as which adjacent categories will be combined has no effect.

When the frequency of marking the score categories related to the items was examined, it was observed that approximately half of the individuals concentrated on the first two options in Item 4, Item 9, and Item 10. However, the fact that more than half of the individuals in Turkey data marked the first option made the distribution of categories more skewed. When the creation of the conditions was examined in this context, it was seen that the 1st and 2nd most marked options were combined in conditions 1 and 3 for three-category data and were combined in condition 4 for four-category data. The 4th and 5th least marked options were combined in conditions 1 and 2 for three-category data and in condition 6 for four-category data. DIF analyses showed that the highest DIF values were mostly obtained in condition 2 among conditions 1, 2, and 3 generated for the three-category data. When the four-category data (conditions 4, 5, 6 and 7) were evaluated among themselves, it can be argued that although there was no systematic pattern, more DIF was obtained in condition 6 compared to condition 4. The results of the DSF analysis demonstrated that the results of conditions 1 and 3, in which the first two options were combined in Step 1, differed from the results of condition 2. This differentiation was not systematic and the results of condition 2 were large in some items and small in some others. On the other hand, it can be argued that the DSF values obtained in Step 1 under the conditions created for the four-category data differed between condition 4 and the others. The direction of this differentiation was not standard, while the largest DSF value was obtained in condition 4 for some items, the smallest DSF amount was obtained for some others in condition 4.

Examination of the results obtained from Polytomous DIF and DSF detection methods together

The examination of the studies on DIF and DSF shows that there are studies in which DIF/DSF analyses are performed simultaneously (Akour et al., 2015) or DIF analysis is performed first and then DSF analysis is performed only on DIF-containing items (Miller et al., 2010). Akour et al. (2015) stated that items that do not exhibit large DSF in any of their steps also do not exhibit DIF. However, it has been observed that Type I error is high in some methods that determine DIF when there is no DSF in the item steps (Ayodele, 2017). In other words, although it is rare, cases where a non-DIF-containing item was marked as DIF were encountered in some of the methods. When the results obtained from this study were examined, it was found that Item 4, which did not exhibit DSF at any step in the large sample, was below the critical values of the DIF analysis results, that is, it did not exhibit DIF. On the other hand, when the DSF results for Item 9 were examined when the sample size ratio was 1:1 in the small sample, the DSF values obtained in Steps 1 and 4 were found to be high and with opposite signs. When the DIF results of the related item were examined, it was determined that the item was not DIF according to most of the methods at the same sample size. This may be due to the fact that the DSF values with opposite signs observed in the steps reduce the DIF effect to almost zero. If DSF analysis is not performed on items that do not exhibit DIF, information about the DSF values of the steps cannot be obtained. Therefore, it should be kept in mind that important information about the steps may be overlooked if you first perform the DIF analysis and then perform the DSF analysis only on the DIF-containing items. As a matter of fact, many DIF detection methods have been reported to show relatively low power when the DSF values change in sign and size across steps (Ankenmann et al., 1999; Chang et al., 1996; Penfield & Algina, 2003; Wang & Su, 2004). Therefore, while making decisions for item revision or item removal, it is recommended to perform a DSF analysis on all items, not only on the items with DIF.

When the DIF and DSF analyses were examined together, it was found that in cases where the DIF amount was the highest, the DSF values obtained from the steps of the relevant items varied, but the signs stayed the same.

Declarations

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

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

References

Akour, M., Sabah, S., & Hammouri, H. (2015). Net and global differential item functioning in pisa polytomously scored science items. *Journal of Psychoeducational Assessment, 33*(2), 166–176. https://doi.org/10.1177/0734282914541337

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

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement, 36*(4), 277–300. https://doi.org/10.1111/j.1745-3984.1999.tb00558.x

Ayodele, A.N. (2017). *Examining power and type 1 error for step and item level tests of invariance: Investigating the effect of the number of item score levels* (Doctoral dissertation). University of Minnesota, USA.

Benítez, I., Padilla, J.L., Hidalgo Montesinos, M. D., & Sireci, S. G. (2015). Using mixed methods to interpret differential item functioning. *Applied Measurement in Education*, 29(1), 1–16. https://doi.org/10.1080/08957347.2015.1102915

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

Camilli, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioral Statistics*, 24(4), 323–341. https://www.jstor.org/stable/pdf/1165366.pdf

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

Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomous items: An adaptation of the SIBTEST procedure. *Journal of educational measurement, 33*(3), 333–353. https://doi.org/10.1111/j.1745-3984.1996.tb00496.x

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. An NCME instructional module. *Educational Measurement: Issues and Practice*, 17(1), 31–44. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-3992.1998.tb00619.x

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

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological), 20*(2), 215–232. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x

Ellis, B. B., & Raju, N. S. (2003). Test and Item Bias: What they are, what they aren't, and how to detect them. In J. E. Wall & G. R. Walz (Eds.*), Measuring up: Assessment issues for teachers, counselors, and administrators*. CAPS Press.

Elosua, P., & Wells, C. S. (2013). Detecting DIF in polytomous items using MACS, IRT and ordinal logistic regression. *Psicológica, 34*(2), 327–342. https://www.redalyc.org/pdf/169/16929535011.pdf

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*(3), 315–332. https://doi.org/10.1111/j.1745-3984.1996.tb00495.x

Gattamorta, K. A. (2009). *A comparison of adjacent categories and cumulative DSF effect estimators* [Doctoral dissertation]. University of Miami, Florida.

Gattamorta, K. A., & Penfield, R. D. (2012). A comparison of adjacent categories and cumulative differential step functioning effect estimators. *Applied Measurement in Education, 25*(2), 142–161. https://doi.org/10.1080/08957347.2012.660387

Gelin, M. N., & Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the center for epidemiologic studies depression scale. *Educational and Psychological Measurement, 63*(1), 65–74. https://doi.org/10.1177/0013164402239317

Gonzalez-Roma, V., Hernandez, A., & Gomez-Benito, J. (2006). Power and Type I error of the mean and covariance structure analysis model for detecting differential item functioning in graded response items. *Multivariate Behavioral Research, 41*(1), 29–53. https://doi.org/10.1207/s15327906mbr4101_3

Göçer-Şahin, S., Gelbal, S., & Walker, C. M. (2016, October). *Impact of decreasing category number of polytomous items on DIF* [Conference presentation]. 15th International Mineral Processing Symposium (IMPS 2016), USA.

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

Henderson, D. L. (2001, April 10-14). *Prevalence of gender DIF in mixed format high school exit examinations*. American Educational Research Association 2001 Annual Meeting, USA. https://files.eric.ed.gov/fulltext/ED458284.pdf

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika, 35*(4), 401–415. https://doi.org/10.1007/BF02291817

Kristjansson, E., Aylesworth, R., Mcdowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*(6), 935–953. https://doi.org/10.1177/0013164405275668

Kuzu, Y. (2021). *Investigation of Differential Item and Step Functioning Procedures in Polytomously Scored Items* [Doctoral dissertation]. Hacettepe University, Ankara.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, 58(303), 690–700. https://www.jstor.org/stable/pdf/2282717.pdf

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, 7(4), 361–388. https://tarjomefa.com/wp-content/uploads/2019/10/F1430-TarjomeFa-English.pdf

Mellor, T. L. (1995). *A comparison of four differantial item functioning methods for polytomously scored items* [Unpublished doctoral dissertation]. The University of Texas, Austin.

Miller, T., Chahine, S., & Childs, R. A. (2010). Detecting differential item functioning and differential step functioning due to differences that should matter. *Practical Assessment, Research, and Evaluation, 15*(10), 1–13. https://doi.org/10.7275/dzm4-q558

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of educational measurement, 44*(3), 187–210. https://doi.org/10.1111/j.1745-3984.2007.00034.x

Penfield, R. D. (2008). Three classes of nonparametric differential step functioning effect estimators. *Applied Psychological Measurement, 32*(6), 480–501. https://doi.org/10.1177/0146621607305399

Penfield, R. D. (2010). Distinguishing between net and global DIF in polytomous items. *Journal of Educational Measurement, 47*(2), 129–149. https://doi.org/10.1111/j.1745-3984.2010.00105.x

Penfield, R. D. (2013). DIFAS 5.0 differential item functioning analysis system user's manual. https://soe.uncg.edu/wp-content/uploads/2015/12/DIFASManual_V5.pdf

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement, 40*(4), 353–370. https://doi.org/10.1111/j.1745-3984.2003.tb01151.x

Penfield, R. D., Alvarez, K., & Lee, O. (2008). Using a taxonomy of differential step functioning to improve the interpretation of DIF in polytomous items: An illustration. *Applied Measurement in Education, 22*(1), 61–78. https://doi.org/10.1080/08957340802558367

Penfield, R. D., & Lam, T. C. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice*, 19(3), 5–15. https://onlinelibrary.wiley.com/doi/pdfdirect/10.1111/j.1745-3992.2000.tb00033.x

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://www.jstor.org/stable/pdf/1435184.pdf

Sireci, S.G., & Rios, J.A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, 19(2-3), 170-187. https://doi.org/10.1080/13803611.2013.767621

Somes, G. W. (1986). The generalized Mantel–Haenszel statistic. *The American Statistician*, 40(2), 106–108. https://www.jstor.org/stable/pdf/2684866.pdf

Wang, W. C., & Su, Y. H. (2004). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement*, 28(6), 450–480. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=96cc44755a12838b2cde4401a0635aaa6b075768

Wood, S. W. (2011). *Differential item functioning procedures for polytomous items when examinee sample sizes are small* [Unpublished doctoral thesis]. The University of Iowa, USA.

Yandı, A. (2017). *Comparison of the methods of examining measurement equivalence under different conditions in terms of statistical power ratios* [Unpublished doctoral thesis]. Ankara University, Ankara.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly*, 4(2), 223–233. https://doi.org/10.1080/15434300701375832

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, 2012(1), i-30. https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/j.2333-8504.2012.tb02290.x

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of educational measurement*, 30(3), 233–251. https://doi.org/10.1111/j.1745-3984.1993.tb00425.x

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(3); 222-234

Rubrics in Terms of Development Processes and Misconceptions*

Fuat ELKONCA** Görkem CEYHAN*** Mehmet ŞATA****

Abstract

The present study aimed to examine the development process of rubrics in theses indexed in the national thesis database and to identify any misconceptions presented in these rubrics. A qualitative research approach utilizing document analysis was employed. The sample of theses was selected based on a literature review and criteria established by expert opinions, resulting in a total of 395 theses being included in the study using criterion sampling. Data were collected through a "thesis review form" developed by the researchers. Descriptive analysis was employed for data analysis. Findings indicated that approximately 27% of the 395 theses contained misconceptions, with a disproportionate percentage of these misconceptions (The rating scale was called rubric and the checklist was called rubric) being found in master's theses. Regarding the field of the thesis, the highest rate of misconceptions was observed in health, social sciences, special education, and fine arts, while the lowest rate was found in education and linguistics. Additionally, theses with misconceptions tended to possess a lower degree of validity and reliability evidence compared to those without misconceptions. This difference was found to be statistically significant for both validity evidence and reliability evidence. In theses without misconceptions, the most frequently presented validity evidence was expert opinion, while the reliability evidence was found to be the percentage of agreement. The findings were discussed in relation to the existing literature, and recommendations were proposed.

Keywords: rubric, document analysis, misconception, reliability, validity.

Introduction

In the field of social and educational sciences, the use of appropriate measurement tools and methods is crucial to ensure the consistency and accuracy of decisions made about test takers. These characteristics are often intangible and exist only through indirect measurement. Therefore, it is important to provide evidence of the reliability and validity of the measurements obtained from these tools. There are various classifications for measurement tools, but they can generally be divided into traditional and complementary/versatile categories. The shift towards a constructivist approach in education since 2005-2006 has led to increased use of complementary measurement tools.

Rubrics, a type of complementary measurement tool, have gained widespread use in education and training activities (Brookhart, 2018). This trend is largely attributed to the flexibility and appropriateness of rubrics in assessing 21st-century skills, which are higher-order cognitive abilities (Dochy et al., 2006). Rubrics must be designed with clear and well-defined criteria and performance level definitions to measure these skills effectively (Brookhart & Chen, 2015; Lane & Tierney, 2008). One of the main reasons for the popularity of rubrics in education and training is their high level of reliability and validity in measurement (Jonsson & Svingby, 2007). Several studies have explored the use of rubrics in education and have discussed the reliability and validity issues surrounding their use (Brookhart, 2018; Brookhart & Chen, 2015; Jonsson & Svingby, 2007; Panadero & Jonsson, 2013; Reddy & Andrade, 2010). These studies suggest that the development of rubrics should be approached in a systematic manner, with a focus on collecting evidence for their reliability and validity (Moskal, 2000; Moskal & Leydens, 2000).

---

* A part of this study was presented at 8th International Congress on Measurement and Evaluation in Education and Psychology. Ege University, İzmir, Turkey.

** Asst. Prof. Dr., Muş Alparslan University, Muş- Türkiye, f.elkonca@alparslan.edu.tr, ORCID ID: 0000-0002-2733-8891
*** Asst. Prof. Dr., Muş Alparslan University, Muş- Türkiye, g.ceyhan@alparslan.edu.tr, ORCID ID: 0000-0001-9342-6876
**** Assoc. Prof.Dr., Van Yüzüncü Yıl University, Faculty of Education, Van-Türkiye, mehmetwsata@gmail.com, ORCID ID: 0000-0003-2683-4997

To cite this article:
Elkonca, F., Ceyhan, G. & Şata, M. (2023). Rubrics in terms of development processes and misconceptions. *Journal of Measurement and Evaluation in Education and Psychology, 14*(3), 222-234. https://doi.org/10.21031/epod.1251470
Received: 15.02.2023
Accepted: 13.09.2023

Unlike checklists and rating scales, rubrics provide a clear definition for each performance level, which is essential for ensuring the validity of measurements. In the process of developing rubrics, it is crucial to seek input from experts in the field to ensure that the definitions accurately represent the relevant features being measured (Moskal & Leydens, 2000). Rubrics are widely used in both educational research and classroom evaluation practices, as they also measure psychological constructs. Therefore, evidence of construct validity is crucial for making accurate inferences. According to the literature, rubrics have several benefits, including higher rater reliability, improved measurement of complex performance tasks, and increased individual reasoning skills (Jonsson & Svingby, 2007; Morrison & Ross, 1998; Wiggins, 1998). These benefits can be realized by ensuring reliability and validity in the development process of rubrics.

It is evident in the national literature that the concept of rubrics is utilized in a variety of different concepts and meanings, indicating the presence of misconceptions. Misconceptions, defined as perceptions or understandings that deviate from the expert consensus (Zembat, 2010), are not solely indicative of errors or lack of knowledge, but rather emerge as a result of faulty cognitive structures. As misconceptions correspond to situations in which cognitive perception leads to systematic errors, individuals who hold misconceptions often exhibit resistance and are unwilling to accept their existence (Yenilmez & Yaşa, 2008). The literature is limited in terms of studies that specifically investigate misconceptions related to rubrics in detail (Brookhart, 2013; Brookhart, 2018; Reynolds-Keefer, 2010). Brookhart (2013) has highlighted that the most prevalent misconceptions include the belief that rubrics are solely used as assessment tools for products and that they serve to quantitatively measure student learning, as well as the conflation of rubrics with rating scale tools. These misconceptions limit the purpose of using rubrics and hinder the full realization of student learning. Therefore, it is crucial to identify and address misconceptions surrounding rubrics. In the present study, the prevalence of misconceptions surrounding rubrics is considered to be of equal importance to the development processes of rubrics.

Numerous studies in the academic literature have examined the use and analysis of rubrics. A commonality among these studies is the emphasis on the presentation of reliability and validity evidence in the development and utilization of rubrics. The present research endeavors to not only investigate this aspect but also to determine if misconceptions exist concerning the utilization of rubrics in master's theses and dissertations (hereafter theses). The evaluation of both the development processes and correct use of rubrics, which are frequently employed in the precise and consistent assessment of 21st-century skills, highlights the significance of this study. Furthermore, while international literature offers a plethora of studies examining rubrics across various levels of education and educational research, the dearth of such studies in the national literature underscores the importance of this research. Additionally, the study aims to examine the rubrics used in theses conducted between 2005, when the constructivist approach was incorporated into the education system, and 2022.

In this study, the primary objective was to examine the development process of rubrics utilized in theses and to investigate any misconceptions surrounding their use. To achieve this goal, the research sought to address the following questions:

1. What is the distribution of rubrics used in theses according to the type and field of theses?
2. Are there misconceptions in the process of developing and using rubrics used in theses, and if so, what types of misconceptions exist?
3. Is there a difference in theses with and without misconceptions in relation to the field and type of theses?
4. Is there a difference in terms of presenting the validity and reliability evidence of theses with and without misconceptions?
5. What is the distribution of theses without misconceptions (the rating scale was called rubric and the checklist was called rubric)?
6. Is there a difference in the validity evidence and reliability evidence of theses without misconceptions according to the field of theses?

Method

Research Design

The study employs document analysis, a qualitative research method, to examine the development processes of rubrics used in theses and associated misconceptions. Document analysis is a systematic approach to evaluate both electronic and printed sources (Corbin & Strauss, 2008; Koyuncu, et al., 2018). While various processes have been reported in the literature, this study adhered to the five stages proposed by Forster (1995), namely, (1) accessing the documents, (2) verifying their authenticity, (3) comprehending the content of the documents, (4) conducting data analysis, and (5) utilizing the obtained data (as cited in Yıldırım & Şimşek, 2011).

Population and Sample

The population of theses was all dissertations and theses scanned in the YÖK (Council of Higher Education) thesis system. However, the criterion sampling method was used, and all theses included in the YÖK thesis system between January 1, 2005 and March 1, 2022, were selected as a sample. This selection was influenced by the fact that constructivist education and complementary measurement and evaluation approaches were commonly used after 2005. The search words such as rubric, rating scale, and checklist were used in the YÖK thesis system to identify relevant theses. A total of 512 theses and dissertations were found as a result of the search with the criteria of year and searching words, but 38 theses with duplicate ID numbers were removed, resulting in 474 theses being included in the examination. Of these, 79 theses were excluded from the study because they only mentioned the name of the rubric and did not use it, leaving a total of 395 theses examined.

Data Collection Tool

The thesis review form developed by the researchers was used as a data collection tool. This tool was created through an analysis of relevant literature and the development of a list of criteria that align with the characteristics and processes that rubrics should possess. Initially, a total of 15 criteria were established.

Validity and Reliability Evidence for the Data Collection Tool

The researchers collected evidence to establish the reliability and validity of measurements obtained based on the checklist developed in their study. To assess content validity, the researchers employed Lawshe's (1975) approach and solicited the opinions of eight experts in the field of Measurement and Evaluation in Education to determine the appropriateness and content validity of the criteria. The content validity ratio limit value for eight experts was set at .69, and one criteria that fell below this threshold were removed (Wilson et al., 2012). One criterion was also revised, resulting in a final data collection tool comprising 14 criteria. This criteria; type of thesis and dissertations, sample group, field of the thesis and dissertations, status of having misconceptions, type of misconception, validity evidence, reliability evidence, rubric type, originality, sample size, guided theory, number of rating scale levels, weighting and scoring.

Considering that the checklists and rating scales mentioned by Brookhart (2018) as misconception types are often referred to as rubrics, these misconceptions were expected to emerge.

To establish the reliability of the measurements obtained from the measurement tool, three experts independently coded 10 randomly selected theses and evaluated each one according to the 13 different criteria. Krippendorff's Alpha reliability coefficient was calculated to determine inter-coder agreement, yielding a coefficient of .93.

Data analysis

In studies conducted based on a qualitative research approach, there are two basic analysis processes: content and descriptive analysis. In this study, a qualitative research approach was adopted, and the method of descriptive analysis was selected as the primary technique for data analysis. The choice of this method was based on the pre-determined features of the rubric, which were established through a thorough examination of existing literature. Furthermore, the chi-square analysis was applied to investigate the incidence of misconceptions, while the z ratio test was utilized to ascertain the presence of significant differences between the categories of the criteria. All data analysis procedures were conducted with a significance level of 0.05.

Findings

The findings were presented according to the order of the research questions. Thus, Table 1 presented information about the thesis type, the field of the thesis, and the sample group of the documents analyzed.

Table 1

Distribution of theses according to their type and field, and sample group

Criterion	Category	f	%
Type of thesis	Master's thesis	241	61.0
	Dissertation	154	39.0
Sample group	Primary school	55	13.9
	Middle school	124	31.4
	High school	35	8.9
	Associate degree	2	0.5
	Undergraduate	120	30.4
	Teacher	29	7.3
	Other	30	7.6
Field of the thesis	Educational sciences	98	24.8
	Basic education	45	11.4
	Special education	3	0.8
	Science and math education	99	25.1
	Turkish and social education	67	17.0
	Science	4	1.0
	Health sciences	4	1.0
	Social sciences	5	1.3
	Fine arts	41	10.4
	Linguistics	29	7.3
Total		395	100

Regarding Table 1, most of the theses utilizing rubrics were master's theses. Furthermore, the primary sample population for these theses was composed of individuals at the secondary school and undergraduate levels. Upon examination of the distribution of theses by field, the majority were in the fields of science and mathematics education and educational sciences. Following the analysis of the distribution of rubrics according to the type and field of the thesis, Table 2 presented an examination of the prevalence of misconceptions and, if present, identified the specific misconceptions.

Table 2

Distribution of the presence of misconceptions in rubrics and identification of specific misconceptions

Misconception		f	%
Status of having misconceptions	Yes	104	26.3
	None	291	73.7
	Total	395	100.0
Type of misconception	The rating scale was called rubric	88	85.0
	The checklist was called rubric	16	15.0
	Total	104	100.0

Based on the distributions in Table 2, 104 rubrics had misconceptions while 291 (73.7%) did not. It is seen that in 88 (85.0%) of the theses with misconceptions, the rating scale was called as rubric, and the checklist was called as rubric in 16 (15.0%) of the theses with misconceptions. The comparison of field of the thesis in terms of having misconceptions was presented in Table 3.

Table 3

Comparison of the theses with and without misconceptions according to the field of the thesis

Category	Misconception				Chi-square			
	No		Yes				Compare Column Proportions	
	f	%	f	%	χ^2	p	Misconception (No)	Misconception (Yes)
Educational sciences (A)	79	80.6	19	19.4	17.01	.009*	A-F (p = .047) A-G (p = .031)	
Basic education (B)	31	68.9	14	31.1				
Science and math education (C)	78	78.8	21	21.2			C-G (p = .032)	
Turkish and social education (D)	49	73.1	18	26.9				
Linguistics (E)	23	79.3	6	20.7				
Fine arts (F)	24	58.5	17	41.5				F-A (p = .047)
Other (G)	7	43.8	9	56.3				G-A (p = .031) G-C (p = .032)
Total	291	73.7	104	26.3				

*p< .05

As demonstrated in Table 3, a significant difference (χ^2= 17.01; p < .05) was observed in the prevalence of misconceptions in the rubrics of the theses analyzed within the scope of the study, based on the fields of the theses. The lowest prevalence of misconceptions was found in the fields of Educational Sciences (80.6%), Linguistics (79.3%), and Science and Math Education (78.8%), while the highest prevalence of misconceptions was found in the fields of Fine Arts (41.5%) and other fields (Health, Social Sciences, Special Education, Science) (56.3%). In order to determine the source of the difference, column ratios were compared (z-test) and it was concluded that theses written in the fields of Educational Sciences

and Science and Math Education contained fewer misconceptions than theses written in Fine Arts and other fields (health, social sciences, special education, science). The findings related to the comparison of the rubrics with and without misconceptions according to thesis type were presented in Table 4.

Table 4

Comparison of the theses with and without misconceptions according to the thesis type

Category	Misconception				Chi-square			
	No		Yes		χ^2	p	Compare Column Proportions	
	f	%	f	%			Misconception (No)	Misconception (Yes)
Master's thesis	174	72.2	67	27.8	0.69	.406	---	---
Dissertation	117	76.0	37	24.0			---	---
Total	291	73.7	104	26.3				

As exhibited in Table 4, an analysis was conducted to investigate the prevalence of misconceptions in the rubrics, based on the level of degree (master's thesis or dissertation). Results revealed that there was no statistically significant difference in the prevalence of misconceptions between the two groups (χ^2= 0.69; p >. 05). Specifically, it was found that 27.8% of the master's theses and 24% of the dissertations contained misconceptions, with similar ratios observed in both groups.

Table 5

Comparison of theses with and without misconceptions regarding validity and reliability evidence

Variable	Category	Misconception				Chi-square			
		No		Yes		χ^2	p	Compare Column Proportions	
		f	%	f	%			Misconception (No)	Misconception (Yes)
Validity evidence	No (A)	102	65.4	54	34.6	9.13	.003*	---	A-B (p = .003)
	Yes (B)	189	79.1	50	20.9			B-A (p = .003)	---
	Total	291	73.7	104	26.3				
Reliability evidence	No (A)	141	64.7	77	35.3	20.28	.000*	---	A-B (p = .000)
	Yes (B)	150	84.7	27	15.3			B-A (p = .000)	---
	Total	291	73.7	104	26.3				

*p < .05

Table 5 presented the results of a chi-square analysis comparing the presence of misconceptions in the rubrics of the theses within the scope of the research, in terms of the inclusion of validity and reliability evidence. The results indicated a statistically significant difference between the two groups (χ^2= 9.13; p < .05). The findings revealed that the proportion of theses containing misconceptions was higher among the group without validity evidence (34.6%) compared to the group with validity evidence (20.9%). The same pattern was observed when examining the presence of misconceptions in relation to reliability evidence, with 35% of theses without reliability evidence containing misconceptions, compared to 15% of theses with reliability evidence (χ^2= 20.28; p < .05).

Table 6

Distribution of the rubrics used in theses without misconceptions according to various characteristics

Criterion	Category	f	%
Rubric type	Analytic	248	85.2
	Holistic	43	14.8
Originality	Developed	227	78.0
	Adapted	13	4.5
	Original	51	17.5
Sample size	0-30 sample size	102	35.1
	31-100	115	39.5
	101-200	37	12.7
	201 and above	37	12.7
Guided Theory	No	135	46.4
	Classical test theory (CTT)	141	48.5
	Generalizability theory	6	2.1
	More than 1 theory	9	3.1
Number of rating scale levels	Three-level	74	25.4
	Four-level	121	41.6
	Five-level	59	20.3
	Six-level	11	3.8
	Seven-level and above	6	2.1
	Multiple different levels	17	5.8
	No level	3	1.0
Weighting	Criteria were weighted the same	255	87.6
	Criteria weighted differently	34	11.7
	Criteria were not scored	2	0.7
Scoring	Total score	239	82.1
	Median	1	0.3
	Mean	36	12.4
	Percentage	15	5.2

In the analysis of Table 6, 248 (85.2%) rubrics used were analytical, while 43 (14.8%) were holistic. 227 (78%) rubrics were created by the researchers themselves, 13 (4.5%) were adapted, and 51 (17.5%) were taken from another study. In terms of sample sizes, 102 (35.1%) of the rubrics used 0-30 samples, 115 (39.5%) used 31-100, 37 (12.7%) used 101-200, and 37 (12.7%) used 201 or more. 135 (46.4%) of the rubrics lacked theory-based steps, 141 (48.5%) included classical test theory, 6 (2.1%) included generalizability theory, and 9 (3.1%) included more than one theory. Considering the findings on how many levels the criteria of the DPAs were graded, 74 (25.4%) were graded in threes, 121 (41.6%) in fours, 59 (20.3%) in fives, 11 (3.8%) in sixes and 6 (2.1%) in sevens and above. In addition, the criteria were scored differently in 17 rubrics (5.8%), and 3 rubrics were not scored. Considering the different weighting of the criteria, equal weighting was used in the majority of the rubrics (f = 255; 87.6%) while 34 (11.7%) criteria were weighted differently, and 2 (0.7%) rubrics were not rated. Considering the methods used in the interpretation of the scores obtained from rubrics, 239 rubrics (82.0%) were interpreted by taking the total score, 36 (12.4%) by taking the mean score, 15 (5.2%) by taking the percentage, and 1 by taking the median score. Whether the rubrics used in theses without misconceptions contain validity evidence was compared according to their fields, and the findings were presented in Table 7.

Table 7

Comparison of the validity evidence of the rubrics without misconceptions according to the thesis fields

Category	Validity Evidence				Chi-square			
	No		Yes				Compare Column Proportions	
	f	%	f	%	χ^2	p	Validity Evidence (No)	Validity Evidence (Yes)
Educational sciences (A)	17	21.5	62	78.5				A-B (p = .031) A-C (p = .000)
Basic education (B)	13	41.9	18	58.1			B-A (p = .031)	
Science and math education (C)	38	48.7	40	51.3			C-A (p = .000) C-F (p = .040)	
Turkish and social education (D)	17	34.7	32	65.3				
Linguistics (E)	8	34.8	15	65.2	14.66	.023*		
Fine arts (F)	6	25.0	18	75.0				F-C (p = .040)
Other (G)	3	42.9	4	57.1				
Total	102	35.1	189	64.9				

*p < .05

As can be seen in Table 7, the presence or absence of validity evidence in the rubrics without misconceptions in the theses analyzed within the scope of the research was compared according to the thesis fields and a statistically significant difference was obtained (χ^2= 14.66; p < .05). Based on the findings, in the process of developing or using rubrics, the most validity evidence was presented in the fields of Educational Sciences (78.5%) and Fine Arts (75%), respectively. In addition, the least validity evidence was in the fields of Science and Mathematics education (51.3%), Other fields (57.1%) and Basic Education (58.1%). In order to determine the source of the difference, column ratios were compared (z-test). The rate of having validity evidence of rubrics in theses written in the fields of Educational Sciences and Fine Arts Education was significantly higher than Basic Education and Science and Math fields. The types of validity evidence presented for the rubrics used in theses without misconceptions were also analyzed and their distributions were presented in Table 8.

Table 8

Distribution of the types of validity evidence presented in the rubrics used in theses without misconceptions

Types of Validity Evidence		Yes		No	
		f	%	f	%
Validity Evidence		189	64.9	102	35.1
Factor Analysis		9	3.1	282	96.9
Content Validity		186	63.9	105	36.1
	Expert Opinion Only	178	61.2	113	38.8
	Lawshe-Davis	7	2.4	284	97.6
	Table of specification	5	1.7	286	98.3
Criterion Validity		2	0.7	289	99.3

According to the findings, validity evidence was reported in a total of 189 (64.9%) theses. The striking result of the study was that the evidence presented for content validity (f = 186; 63.9%) was quite high, but it was concluded that most of this evidence relied on expert opinion only (f = 178; 61.2%). For content validity, statistical analyses such as Lawshe-Davis (f=7; 2.4%) and table of specification (f=5; 1.7%) were involved in a minimal number of theses. Similarly, it was concluded that the evidence

presented for factor analysis (f=9; 3.1%) and criterion validity (f=2; 0.7%) were very few. Within the scope of the research, whether the DPAs used in theses without misconceptions contain reliability evidence was compared according to the fields in which the theses were written and the findings obtained are given in Table 9.

Table 9

Comparison of the reliability evidence of the rubrics without misconceptions according to the thesis fields

Category	Reliability Evidence				Chi-square			
	No		Yes				Compare Column Proportions	
	f	%	f	%	χ^2	p	Reliability Evidence (No)	Reliability Evidence (Yes)
Educational sciences (A)	31	39.2	48	60.8			---	---
Basic education (B)	16	51.6	15	48.4			---	---
Science and math education (C)	43	55.1	35	44.9			---	---
Turkish and social education (D)	23	46.9	26	53.1	6.77	.343	---	---
Linguistics (E)	13	56.5	10	43.5			---	---
Fine arts (F)	10	41.7	14	58.3			---	---
Other (G)	5	71.4	2	28.6			---	---
Total	141	48.5	150	51.5				

As seen in Table 9, the presence or absence of reliability evidence in the rubrics without misconceptions in the theses was compared according to the thesis fields, and no statistically significant difference was found (χ^2= 6.77; p > .05). In general, 51.5% of the theses had reliability evidence, while 48.5% did not. Although, similar to the validity results, more reliability evidence was reported in the rubrics used in theses in the fields of educational sciences (60.8%) and fine arts (58.3%), this difference was not statistically significant. Within the scope of the research, the types of reliability evidence presented for the rubrics used in the theses without misconceptions were also analyzed and their distributions were given in Table 10.

Table 10

Reliability evidence presented in the rubrics used in theses without misconceptions

Types of Reliability Evidence		Yes		No	
		f	%	f	%
Reliability Evidence		150	51.5	141	48.5
Item Analysis (Difficulty, discrimination, t-test)		7	2.4	284	97.6
Test-retest		5	1.7	286	98.3
Cronbach Alpha		25	8.6	266	91.4
Inter-Rater Reliability		138	47.4	153	52.6
	Percentage agreement	53	18.2	238	81.8
	Intraclass correlation	44	15.1	247	84.9
	Cohen Kappa	31	10.7	260	89.3
	Kendall Tau	17	5.9	274	94.1
	Krippendorff's Alpha	7	2.4	284	97.6
	G study (generalizability)	4	1.4	287	98.6
	Rasch	3	1.1	288	98.9

According to Table 10, a total of 150 (51.5%) theses reported reliability evidence. The evidence presented for rater reliability was generally high (f = 138; 47.4%). Considering the types of rater

reliability, reliability coefficient was reported using Percentage agreement in 53 (18.2%) theses, intraclass correlation coefficient in 44 (15.1%) theses, Cohen kappa in 31 (10.7%) theses, Kendall Tau in 17 (5.9%) theses, Krippendoff's Alpha in 7 (2.4%) theses, G coefficient in 4 (1.4%) theses, and Rasch method in 3 (1.1%) theses. In addition to these results, 7 (2.4%) theses reported evidence for item analysis, 5 (1.7%) theses reported test-retest and 25 (8.6%) theses reported Cronbach's Alpha reliability coefficient.

Discussion, Conclusion and Recommendations

This study aimed to examine the development process of rubrics used in theses and the misconceptions about use and consruction of rubrics in this process. Findings were discussed according to the research questions.

Most postgraduate theses that used rubrics as data collection tools were at the master's level, with the sample mostly from the secondary school and undergraduate levels. Most were used in science and math education and educational sciences. Document analysis studies showed similar results (Brookhart, 2018; Çolak-Ayyıldız, 2022; Ocak & Yeter, 2018; Reddy & Andrade, 2010). Brookhart (2018) examined the articles published between 2005-2017 and found that most rubrics were based on undergraduate students.

Regarding the findings related to the misconceptions and misconception types, it was found that one-fourth of the theses contained misconceptions. The majority of misconceptions were caused by the use of a rating scale as a rubric. Only a small number of theses used checklists as rubrics. In a similar study, Brookhart (2018) found that checklists were used as rubrics in only 7 of 51 articles. This misconception is present in both national and international literature but is more prevalent in national literature. This highlights a deficiency in the knowledge of researchers in national literature. The lack of addressing this issue in the literature presents a significant problem in practice.

The analysis of misconceptions according to discipline area revealed that the lowest number of misconceptions were in educational sciences, science and math education, and linguistics, while the highest number of misconceptions were in fine arts, which was found to be statistically significant. This may suggest lower reliability and validity of scores obtained through the use of rubrics in fine arts, compared to higher reliability evidence presented in educational sciences theses, which may be due to courses on scale development in postgraduate education. No significant difference was found in misconceptions according to thesis type (dissertation or master's). This indicates that misconceptions are similar in both levels, with 25% of theses having misconceptions, pointing to a high level of misconceptions. Despite regular monitoring of dissertations, this situation highlights a significant deficiency in practice and evaluation.

An analysis was conducted to differentiate the validity and reliability evidence of theses using rubrics as a data collection tool between those with and without misconceptions. Results showed a statistically significant difference between the two groups, with theses without misconceptions having greater validity and reliability evidence. Studies in the literature (Jonsson & Svingby, 2007; Panadero & Jonsson, 2013; Reddy & Andrade, 2010; Rezaei & Lovorn, 2010) showed that reliability and validity evidence for measurements obtained from rubrics were presented. The validity evidence presented in theses without misconceptions was found to mostly be based on expert opinion (content validity), a non-statistical process. Review studies in the literature (Brookhart, 2018; Jonsson & Svingby, 2007; Panadero & Jonsson, 2013) reported similar results. (Brookhart, 2018; Jonsson & Svingby, 2007; Panadero & Jonsson, 2013). In Brookhart (2018), it was found that expert opinion (content validity) was the main form of validity evidence presented. Jonsson & Svingby (2007) found a lower frequency of content validity as validity evidence. Rater reliability was the most commonly reported form of reliability evidence when using rubrics without misconceptions (Brookhart, 2018; Jonsson & Svingby, 2007; Panadero & Jonsson, 2013; Reddy & Andrade, 2010; Rezaei & Lovorn, 2010). Jonsson and Svingby (2007) reported that over half of 76 articles used rater reliability. Rezaei and Lovorn (2010)

similarly found that rater reliability was commonly reported. Brookhart (2018) argued that rater reliability was generally reported in studies using rubrics. This shows that studies using rubrics in the literature tend to present inter-rater reliability as evidence of reliability.

This study focused on theses in which misconception-free rubrics were used since it examined the properties of rubrics. It was found that analytical rubrics were the most commonly used type, primarily developed by researchers rather than adapting pre-existing rubrics (similar to findings in Brookhart, 2018). The widespread use of analytical rubrics in academic studies can be attributed to the specificity of such rubrics. In the development processes of the rubrics, generally a sample size of 100 or less was used, and CTT-based analyses were conducted. In the study conducted by Brookhart (2018), small samples were used more. The prevalence of analytical rubrics in academic studies is largely due to their specificity and the demands of the evaluation process. Analytical rubrics necessitate a more extended evaluation time and are geared towards specific goals and in-class evaluations rather than broader assessments. The use of small sample sizes, as seen in the examination of theses in relevant research, reflects these factors. The rubrics used in these studies were typically assigned levels of four, three, and five, with criteria often having equal weight and total scores being the predominant scoring method. The utilization of mean and median scores was limited.

The results of this research were summarized as follows:

- An analysis of theses utilizing rubrics as data collection tools showed that a majority of the publications were from educational sciences, science and mathematics education, and secondary and higher education. Master's theses made up the majority of the sample.
- The study found that 25% of the theses containing rubrics had misconceptions, and the rating scale was the most commonly used rubric type.
- The least number of misconceptions was found in educational sciences, science and mathematics education, and linguistics, while fine arts showed the highest number of misconceptions. Master's theses and dissertations had similar levels of misconceptions.
- The reliability and validity evidence of the theses with misconceptions were less than those without, and this difference was statistically significant.
- Validity evidence was reported more in theses without misconceptions, especially in theses in the field of Educational Sciences, compared to theses written in other fields.
- The most common validity evidence presented in theses without misconceptions is expert opinion, and the majority of these do not include statistics based on methods such as Lawshe/Davis.
- Percentage agreement was used as reliability evidence, and the use of methods such as Krippendorff's Alpha, generalizability and Rasch was very limited.
- The rubrics used in the theses mainly were equally weighted, analytical, and total score-based.

It should be noted that the results of this research are limited to theses published between 2005 and 2022 and do not encompass other forms of publication. Hence, the findings are restricted to the analysis of theses and may not be representative of the broader literature in the field.

The research highlights the need for increased training and education on rubric development, with a focus on their general features and reliability and validity evidence. It is suggested that experts with experience in scale development be included in thesis committees. It is recommended that, in order to mitigate the identified limitations and misconceptions in the use of rubrics in theses, thesis supervisors should encourage and recommend courses on scale development and adaptation for students working on projects involving measurement tools. The language barriers and resulting translation misconceptions

can be addressed by establishing a common vocabulary or dictionary for concepts in the field of measurement and evaluation.

Declarations

Author Contribution: Fuat Elkonca: Methodology, analysis, discussion, writing & editing, visualization. Görkem Ceyhan: Analysis, discussion, writing & editing, visualization. Mehmet Şata: Introduction, discussion, writing & editing.

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

Ethical Approval: The research data was obtained from theses scanned in the YÖK (Council of Higher Education) thesis system. Therefore, ethical approval is not required.

References

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. *In Frontiers in Education, 3*(22), 1-12. https://doi.org/10.3389/feduc.2018.00022

Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educ. Rev, 67*(3), 343–368. https://doi.org/10.1080/00131911.2014.929565

Corbin, J. & Strauss, A. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (3rd ed.). Thousand Oaks, CA: Sage.

Çolak-Ayyıldız, A. (2022). Alternatif eğitim konusunda yapılmış lisansüstü eğitim tezlerinin incelenmesi. *Gümüşhane Üniversitesi Sosyal Bilimler Dergisi, 13*(3), 877-886.

Dochy, F., Gijbels, D., & Segers, M. (2006). Learning and the emerging new assessment culture. In L. Verschaffel, F. Dochy, M. Boekaerts, & S. Vosniadou (Eds.), Instructional psychology: *Past, present and future trends*. Elsevier.

Forster, N. (1995). *The analysis of company documentation.* C. Cassell ve G. Symon (Eds.), Qualitative medhods in organizational research: A practical guide. Sage.

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

Koyuncu, M. S., Şata, M. & Karakaya, İ. (2018). Eğitimde ölçme ve değerlendirme kongrelerinde sunulan bildirilerin doküman analizi yöntemi ile incelenmesi. *Journal of Measurement and Evaluation in Education and Psychology, 9*(2), 216-238. https://doi.org/10.21031/epod.334292

Lane, S., & Tierney S. T., (2008). Performance Assessment. Thomas L. G, (Ed), In *21st century education: A reference handbook* (Vol. 1), SAGE.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel psychology, 28*(4), 563-575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Morrison, G. R., & Ross, S. M. (1998). Evaluating technology-based processes and products. *New Directions for Teaching and Learning, 74*, 69-77. https://doi.org/10.1002/tl.7407

Moskal, B. M. (2000). Scoring rubrics: What, when and how*?. Practical Assessment, Research, and Evaluation, 7*(3), 1-5. https://doi.org/10.7275/a5vq-7q66

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research, and Evaluation, 7*(10), 1-6. https://doi.org/10.7275/q7rm-gg74

Ocak, İ., & Yeter, F. (2018). Investigation of national theses and articles on "the nature of science" between 2006-2016 years. *Journal of Theoretical Educational Science*, *11*(3), 522-543. https://doi.org/10.30831/akukeg.344726

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational research review, 9*(1), 129-144. https://doi.org/10.1016/j.edurev.2013.01.002

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

Reynolds-Keefer, L. (2010). Rubric-referenced assessment in teacher preparation: An opportunity to learn by using. *Practical Assessment, Research, and Evaluation, 15*(8), 1-9. https://doi.org/10.7275/psk5-mf68

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing writing*, *15*(1), 18-39. https://doi.org/10.1016/j.asw.2010.01.003

Wiggins, G. (1998). *Educative assessment*. Jossey-Bass.

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development, 45*(3), 197-210. https://doi.org/10.1177/0748175612440286

Yenilmez, K., & Yaşa, E. (2008). İlköğretim öğrencilerinin geometrideki kavram yanılgıları. *Uludağ Üniversitesi Eğitim Fakültesi Dergisi, 21*(2), 461-483.

Zembat, İ, Ö. (2010). Kavram yanılgısı nedir?. MF. Özmantar, E. Bingölbali & H. Akkoç (Eds.), *Matematiksel kavram yanılgıları ve çözüm önerileri içinde*. Pegem Akademi.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(3); 235-248

A Bibliometric Analysis of Power Analysis Studies

Gül GÜLER*

Abstract

The primary purpose of this study was to establish a theoretical framework for studies on power analysis conducted in the fields of education, psychology, and statistics for researchers. It also determined which concepts were associated with power analysis over the years and the authors and countries that contributed to the advancement of research regarding this concept. Therefore, the bibliometric characteristics of publications related to power analysis in the Web of Science database were analyzed using the Biblioshiny interface in the R programming language. Our investigation encompassed 515 studies selected based on specific criteria. Data revealed that from 1970 to 2023, these studies originated from 183 sources and involved 1246 authors. Among them, 98 studies were single-authored, and the average number of co-authors per paper stood at 2.88. According to Bradford's Law, Behavior Research Methods, Psychological Methods, and Multivariate Behavioral Research were the most productive journals concerning power analysis, taking up a larger proportion within the core sources compared to other journals. These journals were among the top three in terms of the number of publications, h-index, total number of citations, and publication rankings. These journals were followed by Structural Equation Modeling-A Multidisciplinary Journal, Frontiers in Psychology, and Educational and Psychological Measurement. An examination of studies on power analysis in education, psychology, and statistics according to Lotka's Law indicated that the relevant literature is insufficient and needs further development.

Keywords: Power analysis, bibliometric analysis, Biblioshiny, WOS

Introduction

One of the factors determining the quality of studies in a scientific research process is how the steps of the research are carried out. In this context, the sample representing the population of the research becomes crucial as much as identifying the research problem (Güler, 2022). In research, when considering factors such as accessibility, cost, and time, studies are generally conducted on a sample that represents the relevant population. In this context, the sample size representing the population is also important for the accuracy of statistical decisions. Indeed, applying the same method with different sample sizes in two separate studies can lead to different statistical decisions. Working with excessively large or small samples can lead to specific challenges. As the sample size increases, even a small difference can become significant. Considering the clinical research, working with an excessive number of patients can bring along not only financial challenges but also ethical concerns and various risks (Cohen et al., 2003). Additionally, testing the efficacy of a drug with an insufficient number of patients can result in erroneous conclusions. In many comparative studies, the accuracy of the H1 alternative hypothesis statement—which posits a difference between the compared conditions—should mirror reality. Consequently, the power of these tests is vital in research (Tabachnick & Fidell, 2013; Stevens, 2009). The purpose of many inferential statistics is to test specific hypotheses about potential group differences or correlations between variables (Cohen et al., 2018; Rossi, 2012; Sink & Mvududu, 2010). Statistical power refers to the probability of rejecting the false null hypothesis (H_0; Cohen, 1988). The probability of revealing the desired true effect in the population of a research study is higher with more powerful statistical tests, leading to a more robust outcome. In other words, statistical power is a factor that influences the validity of the decisions made based on the statistical tests used for testing a hypothesis established in a research study. For instance, in a study comparing a characteristic of two or

* Assist. Prof. Dr., Trakya University, Faculty of Education, Edirne-Türkiye, gulyuce2010@gmail.com, ORCID ID: 0000-0001-8626-4901

To cite this article:
Güler, G. (2023). A bibliometric analysis of power analysis studies. *Journal of Measurement and Evaluation in Education and Psychology, 14*(3), 235-248. https://doi.org/10.21031/epod.1343984

Received: 15.08.2023
Accepted: 26.09.2023

more groups, the extent to which the statistical test used can reveal the difference that truly exists is referred to as statistical power. In other words, if there is actually a difference between two or more groups and this difference is confirmed by a statistical decision, then this situation indicates the power of the test for the respective research. Three factors determine the statistical power of a study. One is the significance level, the second is the effect size, and the third is the sample size (Field, 2005; Rossi, 2012; Stevens, 2009).

There are four different situations related to the formulated hypothesis in a research process. Two of these situations lead to a correct decision, while the other two result in an incorrect decision. One of these incorrect decisions is a Type I error, and the other is a Type II error. Type I error occurs when the null hypothesis (H_0) is actually true but is rejected based on a statistical decision. It is also known as alpha (α) error, indicating the significance level of the test. Type II error occurs when the null hypothesis (H_0) is actually false, but it is not rejected based on a statistical decision (Field, 2005). It is also referred to as b. However, 1-b indicates the power of the test. The power of a test takes values between 0 and 1. Values nearing 1 indicate an increase in statistical power. For many years, various studies in educational and social sciences have demonstrated that the power of tests has often been overlooked or that these tests have exhibited low power (Murphy et al., 2014). However, in recent years, studies conducted in these fields have emphasized the importance of having high power in tests. If the power of a study is less than 0.50, its results are often prone to misinterpretation (Murphy et al., 2014). Cozby and Bates (2018) state that the power of tests is generally preferred to range between 0.70 and 0.90 in studies. If researchers do not have a specific benchmark for statistical power regarding their studies, the minimum recommended value for this ratio is 0.80 (Cohen, 1988; Cohen et al., 2018; Süt, 2011). The higher the power of a study, the lower the risk of missing a true effect.

Sample size and effect size can be determined through various methods in power analysis studies. In clinical studies, relevant reference studies in the field are often taken into consideration (Howell, 2010). However, this may not be always feasible in social sciences. Therefore, the researcher can conduct a pilot study before the actual research to estimate the effect size (Ünalan, 2021). When we examine the literature, the power of tests has either been overlooked or not given due importance in many studies conducted in social and educational sciences. However, in recent years, the power of tests has become important even in studies conducted in education and psychology, and reputable journals expect reporting on the power of tests and effect sizes in studies to be published (Cozby & Bates, 2018; Meyners et al., 2020).

The main purpose of this study is to provide researchers in the fields of education, statistics, and psychology who conduct studies on power analysis with a framework related to the relevant literature in these fields. Additionally, it aims to guide researchers who conduct studies on power analysis about which journals and authors to refer to in this regard. Furthermore, it aims to provide insights into collaborations related to the topic, enabling international researchers to access the most frequently engaged institutions in such research. This study also aims to present new trends related to the topic to researchers, enabling them to access relevant information more quickly and easily. In this respect, answers were sought to the following research questions:

1. How are the studies related to power analysis distributed according to years?
2. How are the studies related to power analysis distributed according to countries?
3. How are the studies related to power analysis distributed according to journals?
4. How are the studies related to power analysis distributed according to authors?
5. How are the studies related to power analysis distributed according to collaborative (co-authored) studies?

6. How are the studies related to power analysis distributed according to the common keywords used?

Method

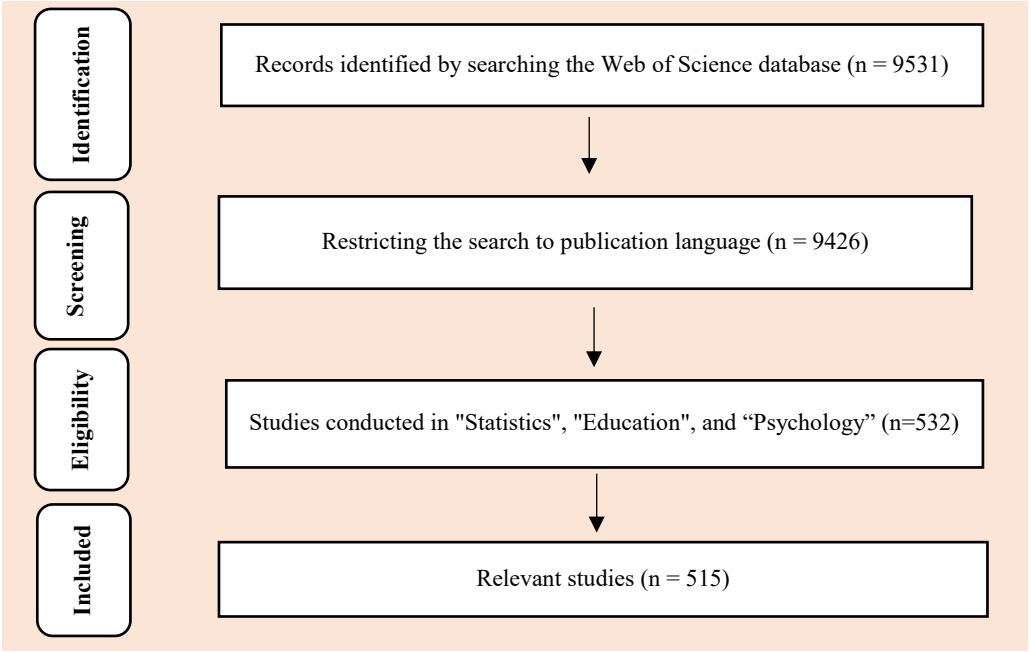
In this study, data related to power analysis were extracted from the Web of Science (WOS) database, following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher et al., 2009). The reason for choosing the WOS database was to access high-quality articles specifically on the mentioned topic. These data include articles focused on power analysis from January 1970 to July 2023.

Selection Strategies and Criteria

The sample of this study was determined using the criterion sampling technique of purposive sampling methods. A literature review was carried out focusing on topics linked to the keyword "Power analysis" within the WOS database. Specifically, studies concentrated on the fields of education, psychology, and statistics were prioritized for inclusion. After listing these studies, fields such as medicine, engineering, computer science, ecology, environmental sciences, law, communication, veterinary medicine, and women's studies were excluded. The inclusion criteria were employed in selecting the studies included in the research. Studies meeting the following inclusion criteria were included: Studies conducted between 1970 and May 2023, studies conducted in education, psychology, and statistics, the publication language is English, and studies including the concept of power analysis in the relevant fields. Figure 1 presents the PRISMA flow chart created based on these criteria.

Figure 1

The PRISMA Flow Chart



Two researchers independently retrieved a total of 515 studies on "Power analysis" from the WOS database using the same inclusion criteria. The retrieved studies were examined through bibliometric analysis. The study was conducted using document analysis, one of the descriptive analyses in qualitative research methods. Both descriptive and evaluative bibliometric analyses were used in this study. The reason for choosing descriptive bibliometrics was to reveal trends in studies in the literature related to power analysis according to countries, publication years, and subjects. Descriptive bibliometrics was employed because it targets measuring productivity, while evaluative bibliometrics was employed because it focuses on measuring the use of relevant literature. Descriptive bibliometrics enables revealing the distribution and trends of the literature according to authors, subjects, publication years, countries, languages, and so on. Evaluative bibliometrics, on the other hand, enables analyzing

the relationships between publications, authors, and countries through citations made by authors (Osareh, 1996).

Data Analysis Technique

This study employed a bibliometric analysis to analyze the data. Bibliometric analysis is a data analysis method used for statistical analyses and evaluation of scientific studies. The WOS database was used to search for relevant studies.

The Bibliometrix software was used to analyze the data in this study (Aria & Cuccurullo, 2017). The Biblioshiny interface was used through the R software for data inclusion criteria. Both descriptive and evaluative bibliometrics were used in the process of obtaining findings in the study. In descriptive bibliometrics, fundamental information about power analysis and descriptive information regarding the sources and authors were examined. In evaluative bibliometric analysis, common keyword analysis, co-authorship analysis, and other conceptual networks were determined to reveal trends, current topics, and research areas related to power analysis. Besides, graphics were generated for the networks of most-cited authors and most cited publications, respectively.

Results

The results are presented under two main headings (descriptive and evaluative bibliometrics).

Results of Descriptive Bibliometrics

This section presents findings related to the distribution of 515 studies on "Power Analysis" in the WOS database by years and researchers' collaboration and productivity. Data related to basic information regarding power analysis are reported in Table 1.

Table 1
Basic Information on Bibliometric Analysis

Timespan	1970:2023
Sources (Journals, Books, etc.)	183
Documents	515
Annual growth rate %	7.1
Average age of documents	10.6
Average number of citations per article	212.4
DOCUMENT CONTENT	
Keywords Plus (ID)	1213
Author's Keywords (DE)	1391
AUTHORS	
Authors	1246
Single-author articles	98
AUTHOR COLLABORATION	
Single-author documents	116
Co-authors per article	2.88
International co-authorship %	20.58
Articles	414
Articles, book chapter	15
Articles, early access	24
Articles, proceedings papers	6

According to Table 1, 515 studies on power analysis were published between 1970 and 2023. The number of citations was 212.4 on average. Of 1246 authors, 98 published single-author studies. The annual average citation graph for studies related to power analysis and the number of articles written over the years are illustrated in Figure 2a and Figure 2b.

Figure 2a
Annual publication rates

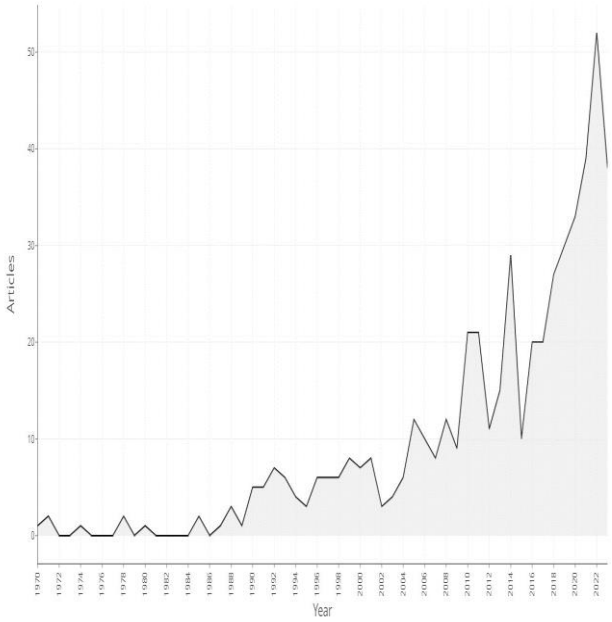
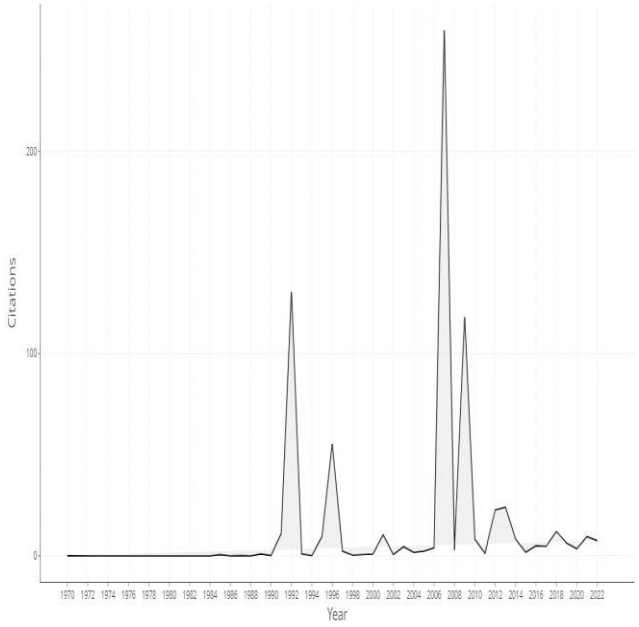


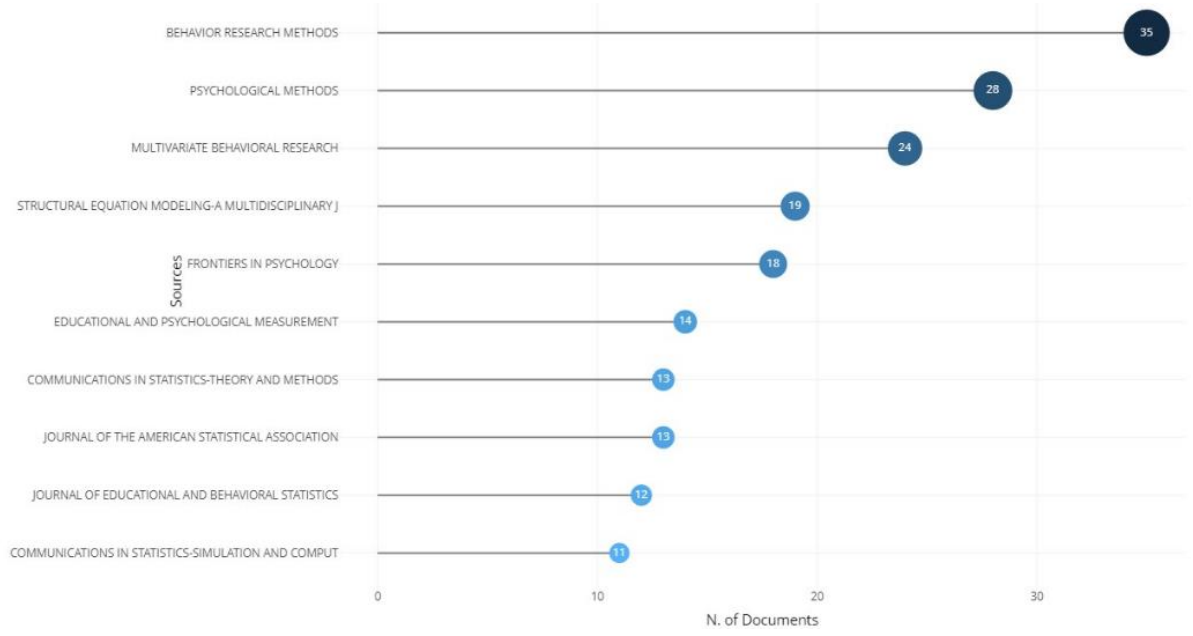
Figure 2b
Annual average citation graph



As illustrated in Figure 2a, the proportion of articles regarding power analysis started to increase from 2016 onwards. However, there were fluctuations in the rate of increase between 2006 and 2016, but there was a rapid increase after 2016. Before 2006, there were a very limited number of studies related to power analysis in education, psychology, and statistics.

The annual average citation graph illustrated in Figure 2b shows that the annual average citation count was below one before 1990. However, it increased from 118.01 in 2007 to 259.78 in 2009. In addition, the number of citations made on power analysis decreased from 2015 onward. The journals with the highest number of published articles are illustrated in Figure 3 to determine the most influential sources related to power analysis.

Figure 3
The most relevant journals



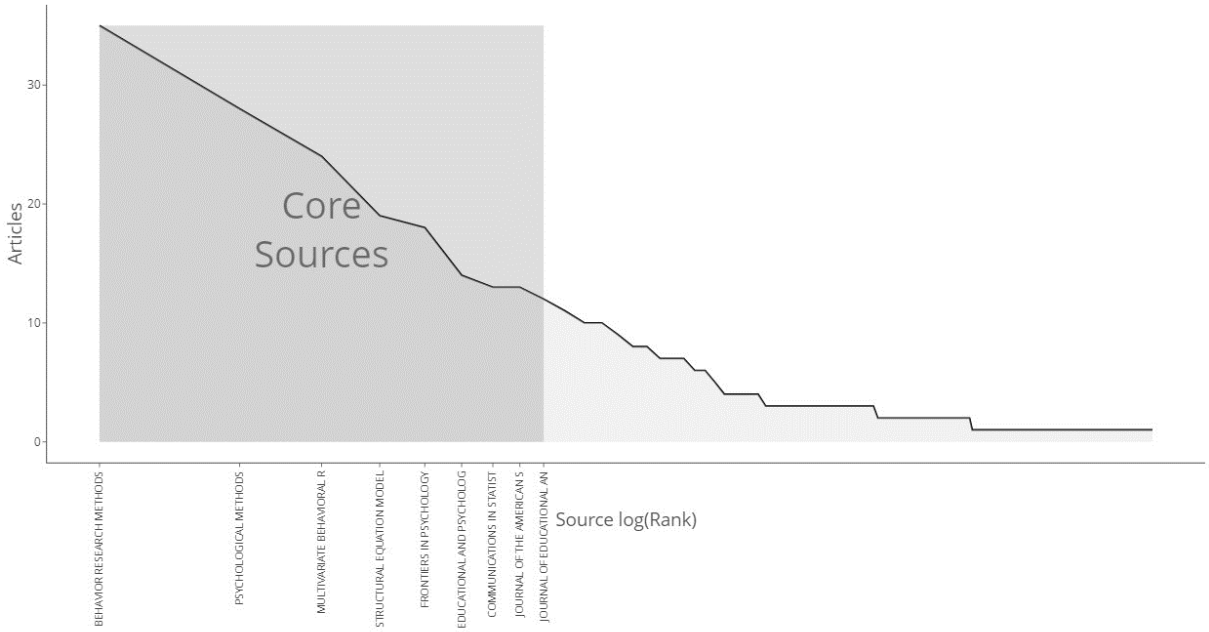
As seen in Figure 3, the journal with the highest number of articles published on power analysis was the Journal of Behavior Research Methods, with 35 published articles. Furthermore, 28 studies were published in the Journal of Psychological Methods and 24 in the Journal of Multivariate Behavioral Research. The h-indexes, total number of citations (TNC), and number of publications (NP) of the journals are shown in Table 2.

Table 2
The h-Index, TNC, and NP of Journals

Journal	h-Index	TNC	NP
Psychological Methods	20	7871	27
Behavior Research Methods	14	48812	30
Multivariate Behavioral Research	11	2766	24
Structural Equation Modeling-A Multidisciplinary Journal	10	564	19
Frontiers In Psychology	8	3457	15
Journal Of Educational And Behavioral Statistics	8	1028	11
Journal Of The American Statistical Association	8	440	12
Behavior Research Methods Instruments & Computers	7	2993	8
Educational And Psychological Measurement	6	417	10
Advances In Methods And Practices In Psychological Science	5	349	6

Table 2 presents ten journals with the highest h-indexes. Considering the results in Table 2, the journal with the highest h-index is the second in productivity ranking. Additionally, considering the total number of citations and publications, this journal holds the second position. Psychological Methods, which has the second highest number of publications and citations, ranks second according to the h-index value. The graph obtained based on Bradford's Law, showing the distribution within the journals in the literature regarding power analysis, is presented in Figure 4.

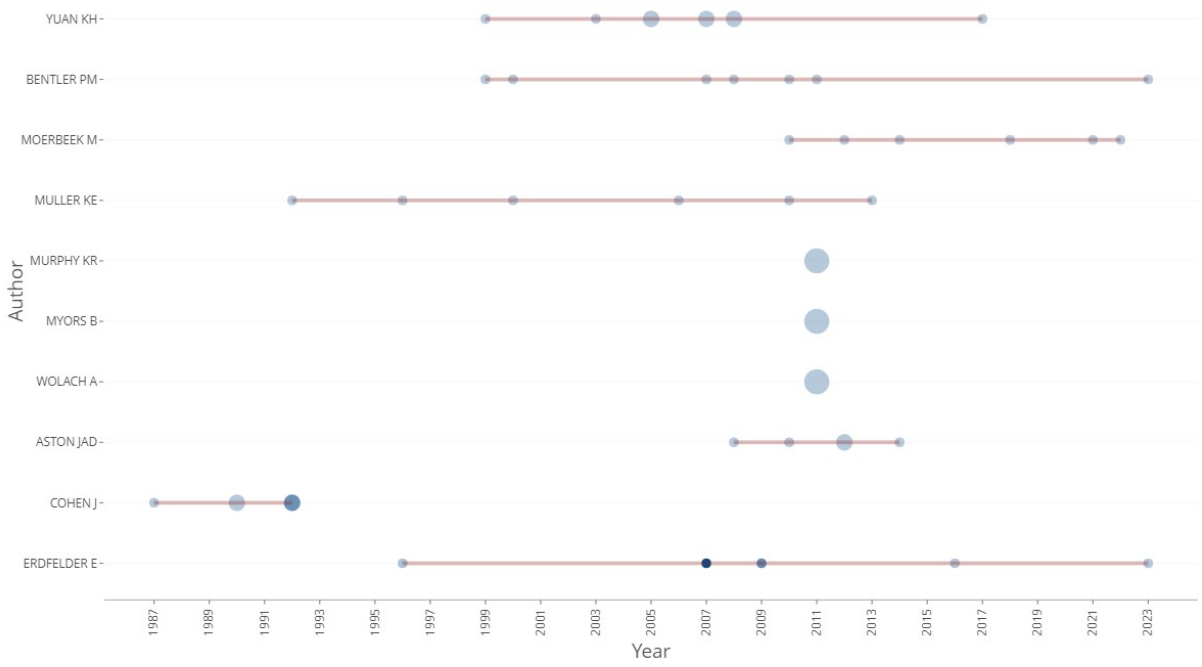
Figure 4
The Bradford law graph



The Bradford Law Graph presented in Figure 4 represents the productivity of the journals. According to this law, journals in a specific field are ranked based on the number of publications they contain. These journals are then grouped into three segments, each containing an equal number of publications. The group with the least number of journals is referred to as the core. The most productive journals are located in the core group, while in the other groups, the number of journals increases while the

publication count remains the same as the core group. In other words, productivity drops in other groups. According to Bradford's Law, the journals Behavior Research Methods, Psychological Methods, and Multivariate Behavioral Research occupy a larger proportion within the core sources compared to other journals. These journals are among the top three in terms of publication count, h-index, total citation count, and publication count rankings. These journals are followed by the journals Structural Equation Modeling-A Multidisciplinary Journal, Frontiers in Psychology, and Educational and Psychological Measurement, respectively. Findings regarding authors' publication productivity over the years are presented in Figure 5.

Figure 5
Authors' productivity over time



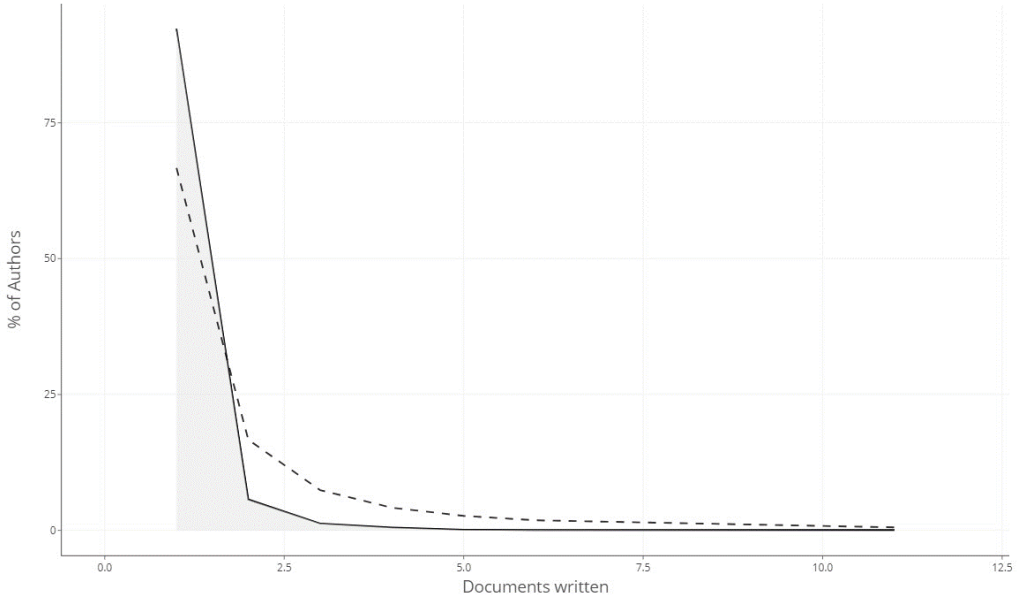
The size and darkness of the circles in Figure 5 indicate the citation strength of the publications. Figure 5 shows that the authors with the highest number of publications on power analysis were Yuan Kh, Bentler PM, and Moerbeek M. Yuan Kh, who continued conducting research on power analysis from 1999 to 2018. However, Yuan Kh., Cohen J., Murphy Kr., and Myors B. were cited more frequently than other authors.

The graph obtained according to Lotka's Law regarding authors' productivity is presented in Figure 6, and the table is provided in Table 3.

Table 3
Author Productivity

Number of written articles	Number of authors	Proportion of authors
1	1114	0.894
2	75	0.06
3	35	0.028
4	9	0.007
5	6	0.005
6	5	0.004
7	1	0.001

Figure 6
Scientific productivity according to Lotka's Law

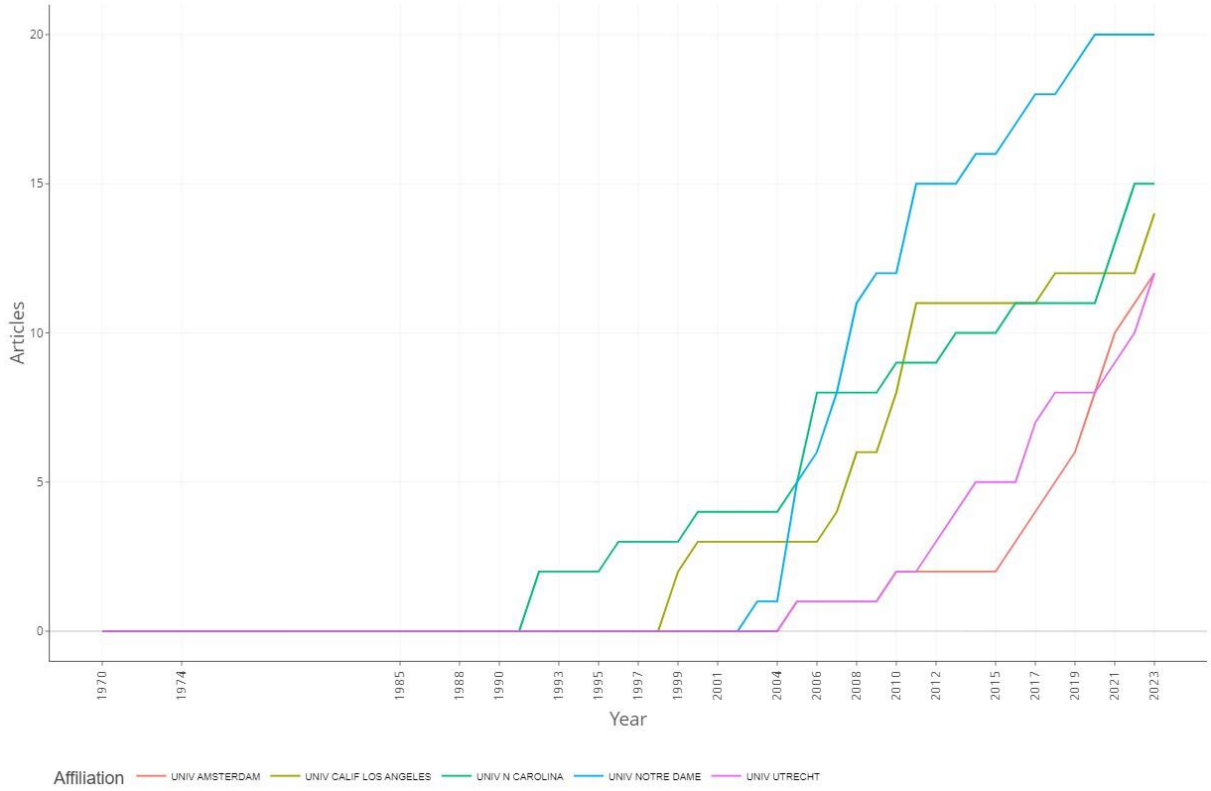


As seen in Figure 6 and Table 3, 89.4% (n = 1114) of the researchers working on power analysis have only published one study regarding this topic, while only 6% (n = 75) have published two studies. Accordingly, the majority of the authors have published only one study on power analysis. According to Lotka's Law, which quantitatively demonstrates the contribution of authors conducting studies in a specific field to the literature and is an indicator of scientific productivity, the number of authors who have made n number of contributions was approximately $1/n^2$ times the number of authors who have made a single contribution. In other words, the proportion of authors with a single contribution among all contributing authors should be a maximum of 60% (Lotka, 1926). In conclusion, it could be stated that the number of authors specializing in power analysis in education, psychology, and statistics is limited.

The changes over time in the productivity of institutions to which researchers producing studies on power analysis are affiliated are presented in Figure 7. According to Figure 7, 232 studies related to power analysis were conducted at the University of North Carolina between 1992 and 2023. After 2013, in particular, there has been an increase in the number of publications in the mentioned university. A total of 198 studies were conducted at the University of California-Los Angeles between 1999 and 2023, with an increase in the number of publications related to power analysis after 2011. Furthermore, 285 studies were conducted at the University of Notre Dame, 76 at the University of Amsterdam, and 93 at the University of Utrecht.

Figure 7

Productivity of researchers' affiliated institutions over time

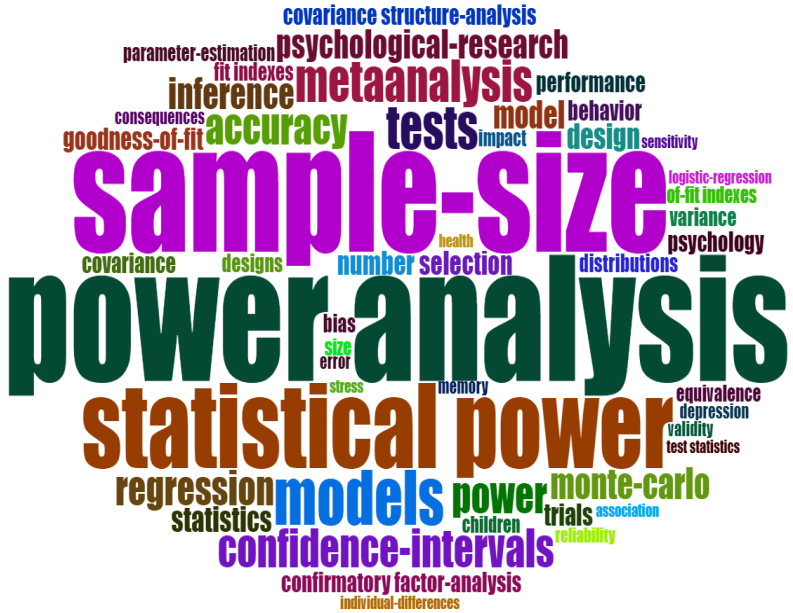


Results of Evaluative Bibliometrics

In the evaluative bibliometric analysis, common keyword analysis, co-authorship analysis, and other conceptual networks were identified to reveal trends in the field of power analysis. The findings related to them are presented below. The most frequently used keywords in publications related to power analysis are presented in Figure 8.

Figure 8

Frequency of keyword use



As illustrated in Figure 8, the keyword "power analysis", written in the largest font, was the most frequently used term. In addition, the keywords "sample size", "statistical power", "models", "tests", and "confidence intervals" were also among the frequently used keywords. Conceptual structure analyses provide valuable insights for researchers working on topics frequently studied in the field. Particularly, they are valuable for identifying trends related to the field.

Figure 9a

Network visualization of keyword formation

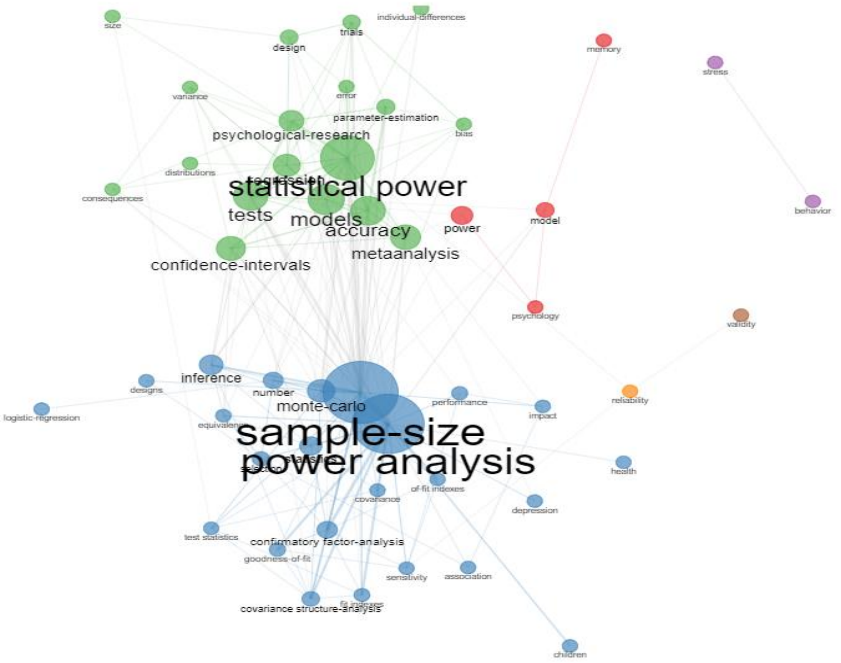
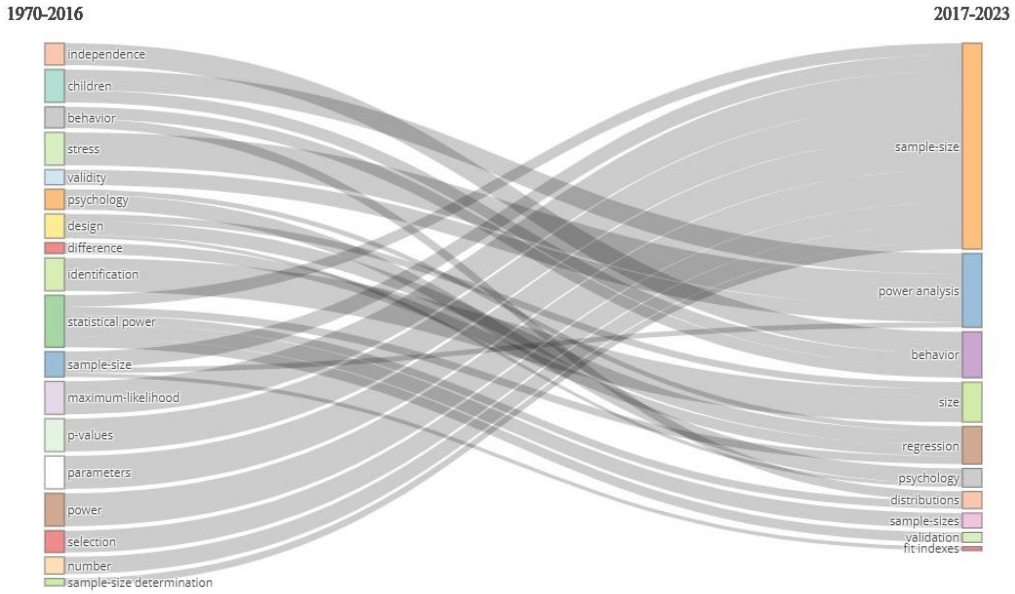


Figure 9b

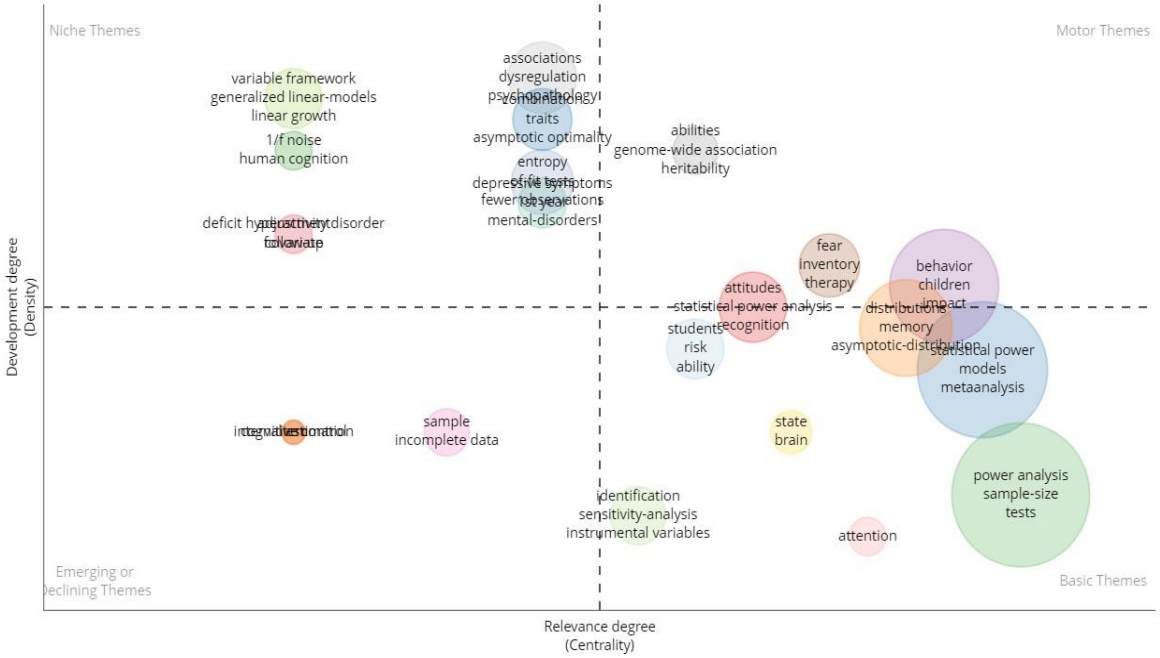
1970-2023 Thematic Evolution in Power Analysis



Conceptual images of the keywords are shown in Figure 9a and Figure 9b. As seen in Figure 9a, each circle represents a keyword, and considering the sizes of the circles, the visualization created using the Louvain Clustering Algorithm confirmed that the most commonly used keywords were “sample size”, “power analysis”, and “statistical power”. The thickening of the lines between circles indicates an increase in the intensity of the relationship between the corresponding words. Figure 9b illustrates how the most important keywords related to power analysis have transformed over time. The use of the thematic map in Figure 10 is common to examine the current state of the power analysis domain and provide insights for future research.

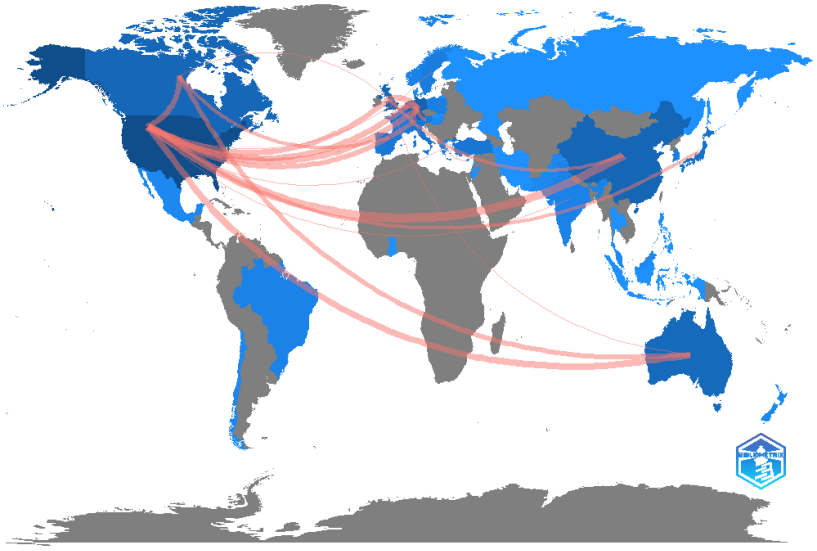
Figure 10

Thematic evolution map



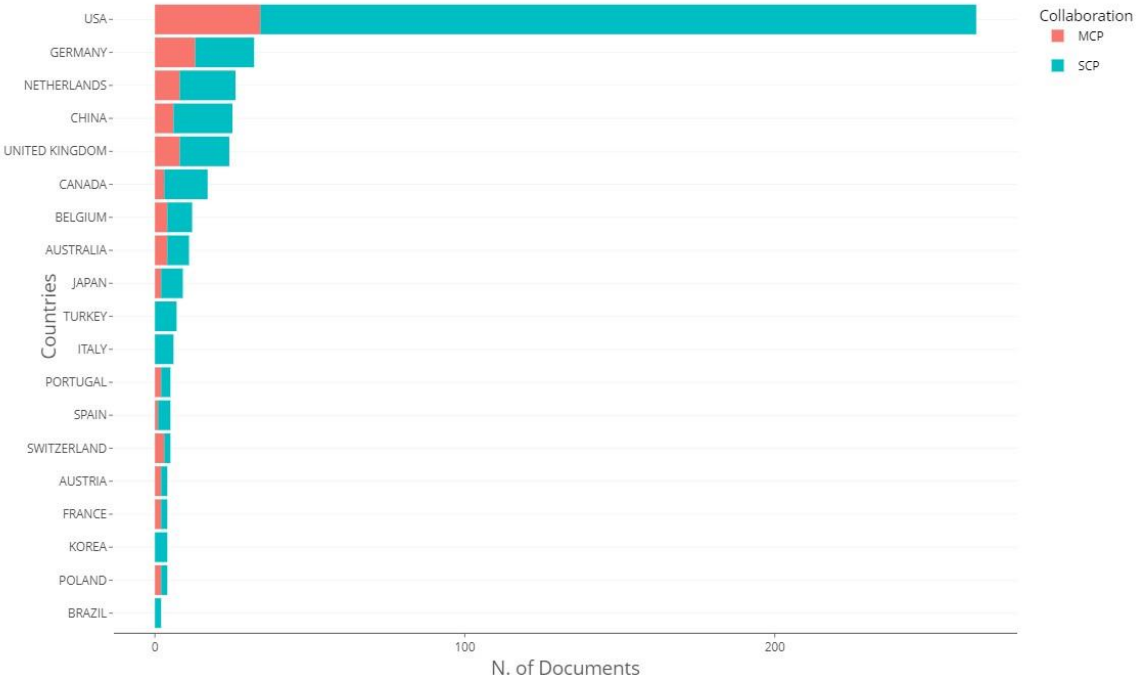
Based on the thematic evolution map depicted in Figure 10, the themes in the upper-right quadrant show significant advancement in power analysis and play a pivotal role in shaping the research area. In the upper-right quadrant, the themes represented by keywords have strong internal connections with each other. In the lower-right quadrant, there are fundamental themes for the field of power analysis. Power analysis is particularly clustered around the keywords statistical power, sample size, and distribution. The themes in this quadrant are of great importance for the research domain. In the upper-left quadrant, marginal themes can be observed. It seems that the themes in this quadrant are not significant enough to shape the research domain. Furthermore, the lower-left quadrant contains themes that are both weakly developed and marginally known. The themes in this quadrant are either in the early stages of development or in a declining trend. Figure 11 depicts a network of collaboration between countries. Figure 11 shows that there is a collaboration between the United States and many countries such as Australia, Canada, and the United Kingdom.

Figure 11
Country Collaboration Map



The number of articles by responsible authors' countries in Figure 12 distributions are included.

Figure 12
Distribution of studies related to power analysis according to countries



According to Figure 12, SCP (Single Country Publications) shows the number of publications by authors in the same country, and MCP (Multiple Country Publications) shows the number of publications made together by authors from different countries. According to both the number of publications by authors from the same country and the number of publications made by authors from different countries together, the USA ranks first, Germany ranks second, and the Netherlands ranks third.

Discussion

The data obtained from the bibliometric analysis were visualized and interpreted through graphs and tables. Both descriptive and evaluative bibliometric approaches were employed to thoroughly examine the study topic. The term "power analysis" was used as a keyword in the WOS database. This study was conducted based on 515 studies that were included considering specific criteria. The analyses were carried out using the R program through the Biblioshiny interface. It was concluded that studies published in the subject area between 1970 and 2023 were obtained from 183 sources. The total number of authors was 1246, the number of single-authored studies was 98, and the number of co-authors per study was 2.88.

An examination of the publication rates of studies on power analysis in education, psychology, and statistics over the years indicated that the proportion of articles related to power analysis began to increase from the year 2016 onward. There were fluctuations in the rate of increase between 2006 and 2016. However, there was a rapid increase after 2016. Before 2006, there were very limited studies related to power analysis in education, psychology, and statistics. One of the reasons for this could be that sample size in studies in the field of health has been considered important in terms of time, cost, and ethics for many years in research. Effect size and power analysis studies have been emphasized, and the required sample size for studies has been determined a priori before conducting the research. However, in recent years, this practice has also gained more attention in the social sciences and educational sciences. In light of all this information, it was concluded that there is a need to increase the number of studies on this subject in education and psychology.

According to Bradford's Law, Behavior Research Methods, Psychological Methods, and Multivariate Behavioral Research were the most productive journals on power analysis, occupying more space than other journals in core resources. These journals are among the top three in terms of the number of publications, h-index, total number of citations, and publication rankings. These journals were followed by Structural Equation Modeling-A Multidisciplinary Journal, Frontiers in Psychology, and Educational and Psychological Measurement journals, respectively. It is particularly important for new researchers who will work on power analysis in education, psychology, and statistics to follow these journals. Yuan Kh., Butler P. M., Moerbeek M., Muller K. E., Murphy Kr., Myors B., and Cohen J. are among the leading authors considering the number of articles they have published regarding power analysis. It is also believed that the works of relevant authors would be important for researchers who are interested in following the literature on the same subject.

The most frequently used keyword was "power analysis", indicating that this keyword has been commonly employed in the literature. Also, the terms "sample size", "statistical power", "models", "tests", and "confidence intervals" were the most frequently used keywords. Conceptual structure analyses provide valuable insights to researchers regarding frequently studied topics in the field. They are particularly valuable for observing trends in the field. The frequent use of keywords such as "sample size" and "confidence intervals" in many studies is likely because the primary purpose of power analysis is to determine the sample size.

This study was conducted using only the WOS database. Bibliometric studies conducted with studies from different databases could be compared with this study. Findings obtained through different programs such as VOSviewer, CiteSpace, and other bibliometric analysis tools that were not used in this study could be compared with the findings of this study.

A great majority of authors have published only once on power analysis, indicating that the number of authors specializing in the field remains limited. According to Lotka's Law, for a field to be considered developed, the number of authors who have published in that field should not exceed 60% of the total number of authors. According to Lotka's Law, those who have published two works should be ¼ of those who have published one work, and those who have published three works should be 1/9 of those who have published one work (Lotka, 1926). When one examines studies related to power analysis in

education, psychology, and statistics according to Lotka's Law, it could be concluded that the relevant literature is insufficient and needs further development.

Declarations

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

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

References

Aria M, & Cuccurullo C. (2017). Bibliometrix: an R-tool for comprehensive science mapping analysis. *J Informetrics*,11(4):959-975. https://doi.org/10.1016/j.joi.2017.08.007

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.),* Hillsdale, NJ: Lawrence Erlbaum.

Cohen J., Cohen P., West S. G. & Aiken L. S., (2003). *Applied multiple regression/correlation analysis for the behavioral sciences,* Third Edition, Lawrence Erlbaum Associates, Publishers, London.

Cohen, L., Manion L. & Morrison K. (2018). *Research methods in education*. (8th ed.) Abingdon: Routledge.

Cozby P. & Bates S. (2018). *Methods in behavioral research*. (13th ed.) New York: McGraw-Hill

Field, A.P. (2005) *Discovering statistics using SPSS*. (2nd ed.) Sage Publications, London.

Güler, G. (2022). Güç analizi ve örneklem büyüklüğü. S. Göçer Şahin & M. Buluş (Ed.), *Adım adım uygulamalı istatistik* (p. 535-560). Pegem Akademi

Howell, D. C. (2010). *Statistical methods for psychology*, 7th ed., Thomson Wadsworth, Cengage Learning. Canada.

Lotka, A.J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences,* 16(12), 317-323. https://www.jstor.org/stable/24529203.

Meyners, M. et al. (2020). To replicate or not to Replicate, or When Did We Start to Ignore the Concept of Statistical Power. *Food Quality and Preference 79.* https://doi.org/10.1016/j.foodqual.2019.01.005

Moher D, Liberati A, Tetzlaff J. & Altman DG. (2009). Preferred reporting items for systematic reviews. *the PRISMA statement. BMJ.* https://doi.org/10.1136/bmj.b2535

Murphy, K. R., Myors, B. & Wolach A. (2014). *Statistical power analysis: a simple and general model for traditional and modern hypothesis tests* (4nd ed.). Routledge, New York.

Osareh, F. (1996). Bibliometrics, citation analysis and co-citation analysis: A review of literature I. *Libri* 46(3):149-158. https://doi.org/10.1515/libr.1996.46.3.149.

Rossi, J. S. (2012). *Statistical power analysis.* In J. A. Schinka & W. F. Welicer(Eds.), Handbook of Psychology. Volume 2: Research Methods in Psychology (2nd. Ed.). John Wiley & Sons.

Sink, C. A., & Mvududu, N. H. (2010) Statistical Power, Sampling, and Effect Sizes: Three keys to Research Relevancy. Counseling Outcome Research and Evaluation 1, 1-18. https://doi.org/10.1177/2150137810373613.

Stevens, J. P. (2009). *Applied multivariate statistics for the social science* (5th Edition). Routledge. London.

Süt, N. (2011). Klinik araştırmalarda örneklem sayısının belirlenmesi ve güç (power) analizi, *RAED Dergisi* ;3(1-2):29-33. https://doi.org/10.2399/raed.11.005.

Tabachnick, B. G., Fidell & L.S. (2013). *Using multivariate statistics* (6th ed.), Boston: Allyn and Bacon.

Ünalan, A. (2021). Sample size in clinical researches: power of the test and effect size. *BSJ Health Sci*, 4(3): 221-227. https://doi.org/10.19127/bshealthscience.866556

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(3); 249-265

A New Weighting Method in Meta-Analysis: The Weighting with Reliability Coefficient*

Yıldız YILDIRIM** Şeref TAN***

Abstract

This study aimed to investigate the impact of various weighting methods for effect sizes on the outcomes of meta-analyses. For this purpose, a representative meta-analysis example examining the effect of the 5E teaching method on academic achievement in science education was discussed. Two effect size weighting methods were explored: one based on the inverse of the sampling error variance and the other utilizing the reliability of measures in primary studies. The study also assessed the influence of including gray literature on the meta-analysis results, considering factors such as high heterogeneity and publication bias. The research followed a basic research design and drew data from 112 studies, encompassing a total of 149 effect sizes. An exhaustive search of databases and archives, including Google Scholar, Dergipark, HEI Thesis Center, Proquest, Science Direct, ERIC, Taylor & Francis, EBSCOhost, Web of Science, and five journals was conducted to gather these studies. Analyses were performed by utilizing the CMA v2 software and employing the random effects model. The findings demonstrated divergent outcomes between the two weighting methods—weighting by reliability coefficient yielded higher overall effect sizes and standard errors compared to weighting by inverse variance. Ultimately, the inclusion of gray literature did not significantly impact any of the weighting methods employed.

Keywords: weighting methods, meta-analysis, reliability coefficient, gray literature

Introduction

Today, with the development of technology and the increase in globalization, science has become more rapidly developing and shared than in the past. As it is known, one of the essential features of scientific research is that it is reproducible and progresses cumulatively. The literature shows that many studies have been conducted in different fields within the framework of the same or similar research problems. For this reason, while there was no need to combine the findings in the past because the number of studies was less, over time, it has become necessary to combine these studies in many fields because of the increase in the number of studies conducted within the same framework and the repetition of studies. As a result, this necessity led to the birth of the meta-analysis method.

The method used to combine findings from repeated studies has a long history (Hedges & Olkin, 1985). Simpson and Pearson's (1904) study was one of the first examples of meta-analysis and evaluated the effectiveness of smallpox vaccine (National Research Council, 1992). Since studies are frequently repeated, it has led to the development of statistical techniques for combining results in different fields. The combining estimates from different studies were not used much in educational or psychological research until Glass proposed it in 1976 because, in studies conducted in these fields, certain psychological constructs or variables were not measured on the same scale in all studies. In 1976, Glass suggested using the effect size index to combine the results of studies conducted with different scales, making the studies comparable and combinable regardless of which scale was used (Hedges & Olkin, 1985). Glass (1976), the eponymist (Mutluer et al., 2020), called the combination of research findings in his study meta-analysis.

---

* This study is a part of doctoral thesis conducted under the supervision of Prof. Dr. Şeref TAN and prepared by Yıldız YILDIRIM GÖRGÜLÜ

** Assist. Prof. Dr., Aydin Adnan Menderes University, Faculty of Education, Aydin-Türkiye, e-mail: yildizyldrm@gmail.com, ORCID ID: 0000-0001-8434-5062
*** Prof. Dr., Retired from Gazi University, Faculty of Education, Ankara-Türkiye, sereftan4@yahoo.com, ORCID ID: 0000-0002-9892-3369

To cite this article:
Yıldırım, Y., & Tan, Ş. (2023). A new weighting method in meta-analysis: the weighting with reliability coefficient. *Journal of Measurement and Evaluation in Education and Psychology, 14*(3), 249-265. https://doi.org/10.21031/epod.1351485
Received: 28.08.2023
Accepted: 26.09.2023

In meta-analysis, an overall effect size is calculated by non-weighting or weighting the effect sizes of primary studies (Fuller & Hester, 1999). To calculate the overall effect size, summing the effect sizes of the primary studies and dividing by the total number of studies, i.e., averaging the effect sizes, is a method used mainly in the past and is called non-weighting in the literature. In addition to the average effect size (overall effect size) without weighting, there are different weighting methods in the literature. These methods generally assume that the error arises from the sample and are based on sample size and sampling error variance. Weighting the effect sizes in primary studies by sample size to obtain the overall effect size was proposed by Hunter and Schmidt (1990). Hunter and Schmidt (2004) stated that if the effect size in the population was assumed to be fixed across studies, to make the best estimation of this effect size, it is necessary to work not with the arithmetic mean of the studies but with a weighted average in which each effect size was weighted by the sample size in the study. Hedges and Vevea (1998) proposed a method called inverse-variance weighting, in which the effect sizes of primary studies are weighted by the inverse of the sampling error variance. In this method, the calculation of weights varies according to random effects and fixed effects models. In the random effects model, in addition to the sampling error variance, the between-studies variance is also taken into account. There are studies on the effects of weighting methods in the literature (Englund et al., 1999; Marín-Martínez & Sánchez-Meca, 2009; Schmidt et al., 2009; Shuster, 2010; Yıldırım & Şahin, 2023). In these studies, the effects of methods such as non-weighting, weighting by sample size, and weighting by the inverse of the sampling error variance were compared and examined.

In meta-analysis studies in the literature, primary studies are generally weighted by the inverse of the sampling error variance based on the sample size, and it is assumed that the error variance is caused only by the sample. However, there are sources of error variance other than the sample. The reliability coefficient is an index that also includes other sources of random error. The error can be caused by the measurement tool or the individual performing the measurement, as well as the environment in which the measurement is made and the construct of the trait. Rosenthal (1991) also stated that it is wise to weight studies in proportion to the quality of the studies using any weight between zero and one.

Based on the research on weighting in the literature, this study, unlike other studies, aimed to examine how the overall effect size and standard error obtained from the meta-analysis were affected by weighting with the reliability coefficient in addition to weighting with the inverse of the sampling error variance because assuming that the error is caused only by the sample is not exactly the right approach. No other study using weighting with a reliability coefficient was found in the literature. Using the reliability coefficient in synthesizing studies in meta-analysis and weighting effect sizes is this study's original and innovative aspect that will contribute to the literature. In this respect, the study differs from other methodological meta-analysis studies. The study discusses how these weighting methods change the results of meta-analysis. The research is essential since not many studies in the literature use a different weighting technique other than weighting by sampling error variance. In addition, the fact that weighting by reliability is used for the first time in this research by formulating weighting by reliability coefficient makes the research essential.

In the literature, it is frequently observed that meta-analysts in educational research do not include unpublished studies such as papers, reports, and theses (Altunoğlu et al., 2020; Bozdemir et al., 2017; Yeşilpınar Uyar & Doğanay, 2018). Such studies are called gray literature. In addition to this situation, it has been observed that there are also studies that include only theses in meta-analysis studies (Alacapınar & Ok, 2020; Basit, 2020; Başpınar, 2021; Saraç, 2018). However, there are meta-analyses that included both published and unpublished studies (e.g., Fabiano et al., 2021; Toraman et al., 2018; Özdemir, 2023). For this reason, it is another question of how the inclusion and exclusion of gray literature in meta-analysis studies affect the meta-analysis results. Based on this, how the inclusion of gray literature under different weighting methods affects the meta-analysis results is also examined within the scope of this study. Although there are studies in the literature that examine the effect of the inclusion of gray literature (Hartling et al., 2017; Moher et al., 1996), what makes this research different from other studies is that it examines this effect in the context of two weighting methods. This study is essential since reviewing the impact of gray literature under different weighting methods is a new issue.

Aim

This study aims to examine how the meta-analysis results are affected when the studies are weighted by sampling error variance and reliability in examining the effect of the 5E teaching method on academic achievement in science education by meta-analysis. In addition, within the scope of the research, it is also examined how the inclusion and exclusion of gray literature affect the meta-analysis results when weighting is done by sampling error variance and reliability in examining the effect of the 5E teaching method in science education on academic achievement by meta-analysis.

Method

Research Model

In this study, meta-analysis was conducted by using the weighting method with the reliability coefficient, which is different from the weighting method with the inverse of the sampling error variance since the error in measurement and evaluation processes is not only caused by the sample. Thus, a new weighting method was proposed to find a solution to the existing problem. According to Karasar (2013), basic research aims to add new knowledge to existing knowledge, and there are different levels of basic research. These are explication, elaboration, determination of cause-effect relationship, and theory development levels. A study at the explication level tries to determine exactly what an existing problem is, what variables are affected by it, and what the most appropriate approaches to explain the situation might be. In this context, the research is at the explication level of the basic research type. On the other hand, it was also examined how the inclusion of gray literature in meta-analysis studies affected the results of meta-analysis when the methods of the inverse of sampling error variance and weighting with reliability were considered. From this point of view, the research also has a descriptive purpose since an existing situation is tried to be revealed.

Data Collection Process

Primary studies constitute the study data in meta-analysis. In the meta-analysis study to be conducted, the study data consists of the studies to be selected according to the determined criteria. In order to strengthen this meta-analysis study methodologically, PICO (Participant/Population, Interventions, Comparisons, Outcomes) was followed. According to PICO, we need to determine which participants, interventions, control groups/comparisons, and outcomes will be taken into account and which we are interested in when constructing the problem. (Higgins & Green, 2008). Therefore, databases were searched with the keywords given in Table 1 to select primary studies to be included in the meta-analysis. In addition, the journals in Table 1 were also included in the search.

Table 1
Databases, keywords and number of studies

Databases	Keywords	Number of Studies
Google Scholar	"5E" + "fen" + "başarı"	1678
Dergipark	5E AND fen AND başarı	61
HEI Thesis Center	5E AND fen AND başarı	125
Proquest	5E AND fen AND başarı	37
Science Direct	5E AND fen AND başarı	0
Science Direct	5E AND science AND achievement	84
ERIC	5E AND fen AND başarı	0
ERIC	5E AND science AND achievement	47
Taylor & Francis	5E AND fen AND başarı	0

Taylor & Francis	5E AND science AND achievement	261
EBSCOhost	5E AND fen AND başarı	268
EBSCOhost	5E AND science AND achievement	130
Web of Science	5E AND fen AND başarı	0
Web of Science	5E AND science AND achievement AND Turkey	53
Journal Name		
Science Education		930
Journal of Research in Science Teaching		1036
Journal of Science Teacher Education		696
International Journal of Science and Mathematics Education		1149
Studies in Science Education		157
Total		6712

The databases presented in Table 1 were selected because these databases are frequently used in meta-analysis studies in the field of education (Arık & Yılmaz, 2020; Batdı & Batdı, 2015; Becker & Park, 2011; Lazonder & Harmsen, 2016; Sosa et al., 2011; Warfa, 2016 and Xie et al., 2018). The journals in Table 1 were selected because they have a high impact factor in the field. The databases were searched with relevant keywords, and all articles in the journals were searched without using keywords, and their full texts were analyzed. These full texts were analyzed according to the criteria determined. The criteria for selecting the study data for the meta-analysis study are listed as follows:

i. The period should be between January 2005 - December 2020,

ii. Papers, articles, dissertations, reports, etc., must have been conducted in a sample of Turkey,

iii. Designed as a weak experimental design, quasi-experimental design, true experimental design, or one of the mixed methods research that used one of the experimental designs in the quantitative research step,

iv. The language of publication must be Turkish or English,

v. Primary studies must have been conducted at the 4th, 5th, 6th7th, 8th, 9th, 10th,11th, or 12th grade or at a higher education level and must be in the field of science, physics, chemistry and biology,

vi. The teaching in the treatment group must have been done with the 5E teaching model or with the 5E teaching model supported by additional applications,

vii. In the control group, traditional methods such as lecture, question and answer, discussion, demonstration, exhibition etc., must have been used, and if not stated in the study, when the authors were contacted via e-mail/message, it was confirmed in their response that they used traditional methods.

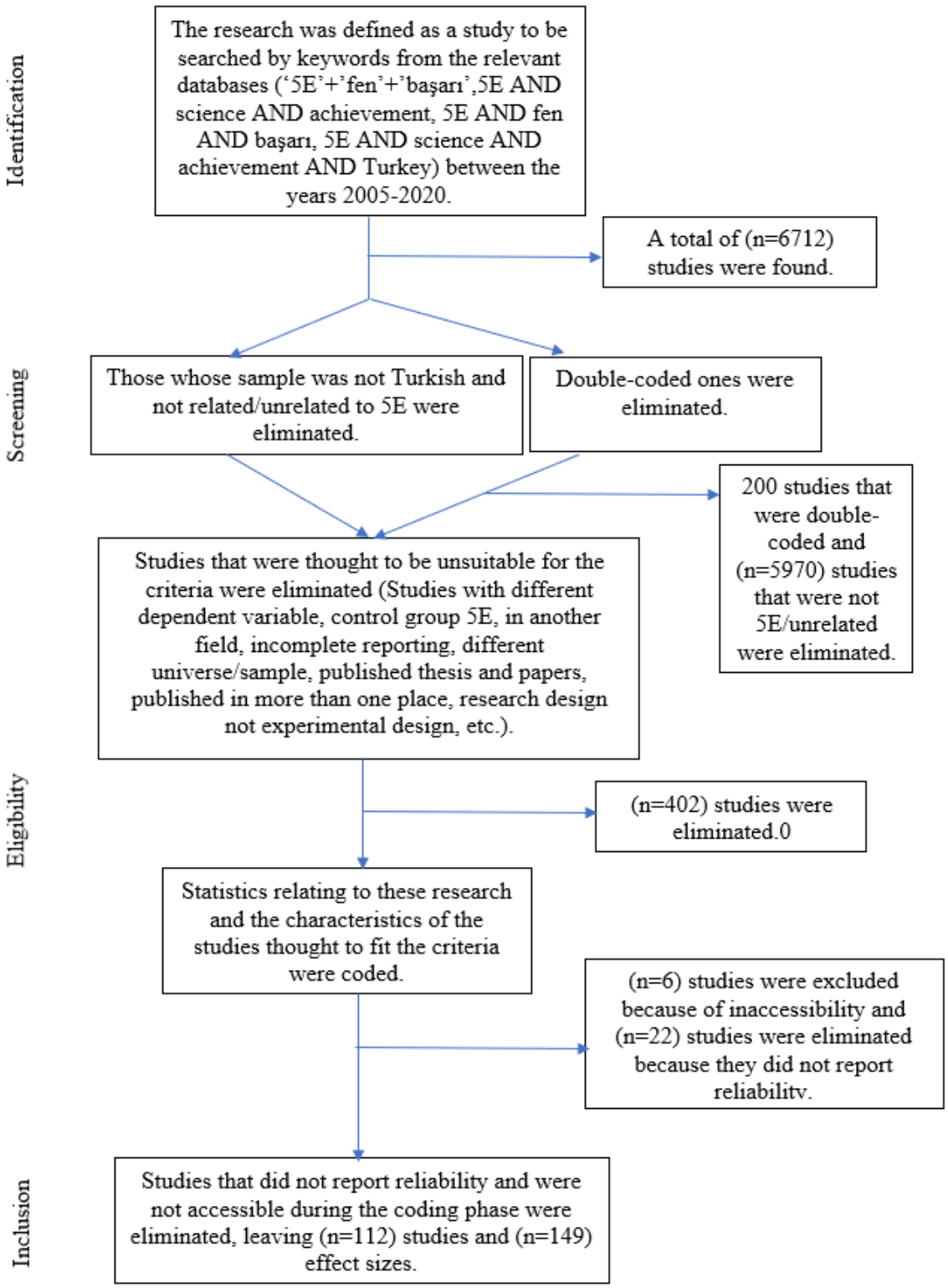
viii. As a data collection tool, tests such as multiple-choice achievement tests, concept tests, conceptual understanding tests, tests composed of open-ended items, and concept maps, which measure academic achievement and report reliability scores, must have been used.

ix. The dependent variable must be academic achievement or concept knowledge.

x. Report sufficient quantitative data and sample size to allow calculation of the effect size.

Primary studies to be included in the meta-analysis were identified according to the search criteria made with keywords in the databases. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart for the process of identifying these studies is given in Figure 1 (Liberati et al., 2009).

Figure 1
PRISMA flowchart



According to the PRISMA flowchart in Figure 1, 112 studies and 149 effect sizes were finally included in the meta-analysis. The studies included in the meta-analysis could not be presented in the article due to page limitations. Therefore, they are shown in the original thesis mentioned in the footnote on the article's first page. Those who want to access the primary studies can access the thesis presented in the references.

Coding of Data

The coding of the 112 studies included in the meta-analysis and the 149 effect sizes obtained from these studies were made in Microsoft Excel. The descriptive variables considered in the coding made in Microsoft Excel are: "publication code, name of the study, colophon (author surnames, year of publication), publication type, publication language, publication year, place of publication, volume-number, authors, database, index, models used, additional application in the treatment group, techniques used in the control group, research design, subject area, grade level, course area (science, physics, chemistry, biology), data collection tool, dependent variable, reliability coefficient, range of difficulty/mean difficulty, population-sample, number of activities, class hours, piloting status (yes/no), the piloting status of achievement test (yes/no), data analysis method, application time, school type". The categorical variables determined for coding and the number of studies and effect sizes in the categories of these variables are given in Table 2.

Table 2

Number of studies and effect sizes for coded categorical variables

	Number of studies (f)	Number of effect sizes (f)		Number of studies (f)	Number of effect sizes (f)
Study Type			**Study Language**		
Article	48	55	English	23	26
Proceeding	9	11	Turkish	89	123
Master's Thesis	37	45	**Databases**		
Doctoral Thesis	18	38	Google Scholar	67	80
Publishing Time			Dergipark	2	2
2005-2009	25	32	ERIC	6	7
2010-2014	46	69	Taylor & Francis	1	2
2015-2020	41	48	HEI Thesis Center	27	48
Study Design			Science Direct	5	5
True experimental	3*	3	Web of Science	3	4
Quasi experimental	96*	124	Proquest	1	1
Poor experimental	14	22			
Grade Level			**Subject**		
4. and 5.	10	13	Science	1	1
6., 7. and 8.	50	63	Physic	44	57
9.,10., 11. and 12.	36*	51	Chemistry	36	52
High education	18*	22	Biology	31	39
Academic Year			**School Type**		
Unspecified	14	20	Unspecified	4	4
(2001-2002)-(2007-2008)	28	36	Public	102	134
(2008-2009)-(2013-2014)	46	66	Private	5	10
(2014-2015)-(2019-2020)	24	27	Public and Private	1	1
Total	112	149		112	149

*One of the studies used both true experimental design and quasi-experimental design.

The statistics related to effect sizes were also coded in the same file for performing the meta-analysis study. Since some primary studies reported effect sizes directly, Cohen *d*, Hedges *g,* and *η2* effect sizes

were taken directly, and the sample size of the treatment groups and the sample size of the control group were also coded. In addition, in some primary studies, the statistics required to calculate effect sizes were coded, and thus effect sizes were calculated. For the true and quasi-experimental designs that calculated statistics such as mean and standard deviation, the mean and standard deviation for the post-test of the treatment group and the mean and standard deviations for the post-test of the control group were coded. If the research was conducted in a weak experimental design, the means and standard deviations for both the post-test and pre-test of the treatment group were included in the coding. In addition, if mean and standard deviation values were not reported in the studies that also used analyses such as *t*-test, ANOVA, Mann Whitney U Test, ANCOVA, MANOVA, MANCOVA, Wilcoxon Signed-Rank Test, and Kruskal-Wallis H Test, statistics related to these analyses were coded, and effect sizes were calculated according to these statistics. Finally, correlation was coded for primary studies that reported correlation coefficient as correlation directly means effect size.

Data Analysis

In the meta-analysis examining the effect of the 5E teaching method on academic achievement in science education, it was examined how the overall effect sizes were affected when weighting with the inverse of the sampling error variance and reliability were applied. In addition, it was also examined how the overall effect sizes were affected when gray literature was included and was not included. CMA program and random effects model were used to obtain the overall effect sizes. Two different types of weighting were used in the CMA program. The first one is weighting by the inverse of the sampling error variance (Hedges & Vevea, 1998), and how it is calculated is shown in Equation 1 (Borenstein et al., 2009);

$$w_i^* = \frac{1}{V_{yi}^*} \tag{1}$$

In Equation 1, w_i* represents the weight of the relevant study for the random effects model, while V_{yi}* is the sum of the sampling error variance (Vyi) of the relevant study to be weighted and the variance between studies (T^2). For weighting by reliability coefficient, the weighting is as in Equation 2 for fixed effects and random effects models. However, within the scope of the research, meta-analysis was conducted according to the random effects model.

$$w_i = r_{at} \qquad w_i^* = r_{at} + T^2 \tag{2}$$

In Equation 2, while w_i represents the weight of the related study, r_{at} represents the reliability coefficient for the measurements obtained with the achievement test used in the related study. T^2 represents the variance between studies and is used to calculate w_i^* in the random effects model.

The weighting types determined were used both for the cases where gray literature was included in the meta-analysis and for the cases where it was not included, and the overall effect sizes and standard errors obtained were interpreted. There were 149 effect sizes in the meta-analysis when gray literature was included, while there were 55 effect sizes when gray literature was excluded. In addition to interpreting the effect of the inclusion and exclusion of gray literature on the meta-analysis results, it was examined whether there was a significant difference between the effect sizes between the studies in the gray literature and the articles. Accordingly, a Q test based on analysis of variance was performed.

Before conducting the meta-analyses, the heterogeneity values for the data were examined with *Q, p(Q), T^2, I^2, H^2* and *R^2* statistics. For the I^2 statistic, 25% is interpreted as low, 50% as medium and 75% as high heterogeneity (Higgins et al., 2003). H^2 and R^2 statistics of 1 is an indication of homogeneity of effect sizes. Publication bias was examined with the funnel plot and trim-and-fill method by Duval and Tweedie (Duval & Tweedie, 2000a; 2000b), Rosenthal's fail-safe *N*, Begg and Mazumdar's rank correlation test and Egger's regression intercept methods. The number of missing studies calculated in

Rosenthal's fail-safe *N* method was compared with the criterion value of 5*k*+10 (k=number of studies) (Rosenthal, 1979). In Begg and Mazumdar's rank correlation and Egger's regression intercept methods, the significance of the correlation and intercept were interpreted, respectively (Begg & Mazumdar, 1994; Egger et al., 1997).

Results

Heterogeneity

Within the scope of the study, firstly, heterogeneity and publication bias regarding the primary studies included in the meta-analysis were examined. The heterogeneity statistics, Q, $p(Q)$, T^2, I^2, H^2 ve R^2, were analyzed and given in Table 3.

Table 3
Heterogeneity statistics

k	*Q*	*df*	*p*	T^2	I^2	H^2	R^2
149	1102.69	148	0.000*	0.455	%86.578	7.450	7.796

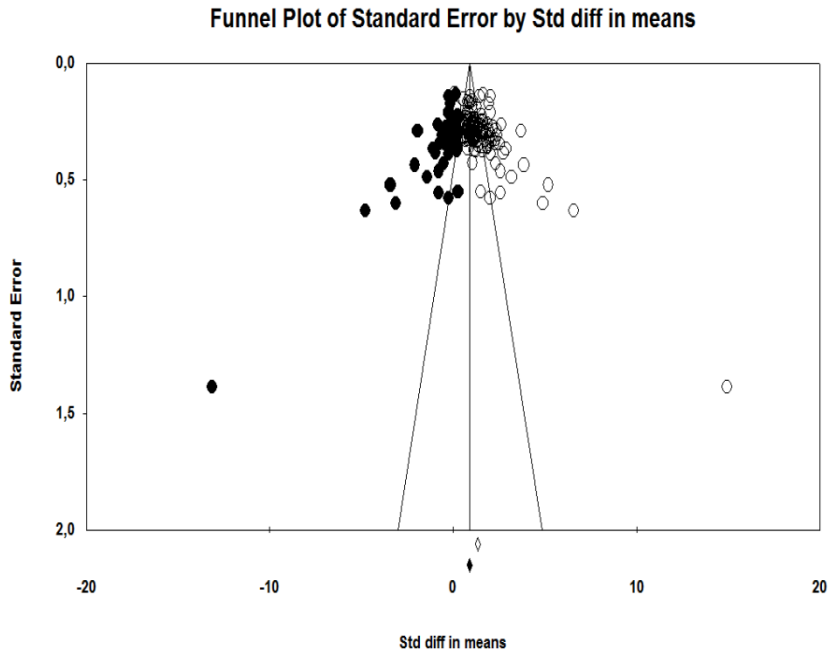
*p < .001

When Table 3 is analyzed, it is seen that the Q value is significant. While this is an indicator of heterogeneity, an I^2 value higher than 75% is an indicator of high heterogeneity (Higgins et al., 2003). Besides, the fact that the T^2 value is quite different from 0 indicates the presence of variance between studies. In addition, the fact that H^2 and R^2 statistics are quite different from 1 indicates that effect sizes are heterogeneously distributed (Higgins & Thompson, 2002). When all statistics are handled together, it is observed that heterogeneity exists. In addition to statistical evidence, there is also theoretical evidence for the existence of heterogeneity. The fact that the studies included in the meta-analysis belong to different populations is also a source of heterogeneity. For example, the research data has a wide range of education levels from secondary school to higher education. Furthermore, the regions where the primary studies were conducted differ from each other in many aspects, such as climate and culture. Moreover, the subject areas in the primary studies differ from each other in physics, chemistry, biology, and science. Based on this, when the statistical and theoretical evidence of heterogeneity is considered together, it can be said that the weighting methods in this study were compared under a condition where heterogeneity exists.

Publication Bias

The study analyzed publication bias using the funnel plot and Duval and Tweedie's trim-and-fill method, Rosenthal's fail-safe N method, Begg and Mazumdar's rank correlation, and Egger's regression intercept method. The funnel plot is given in Figure 2.

Figure 2
Funnel Plot



The funnel diagram in Figure 2 shows that studies (filled dots) had to be added to adjust the symmetry of the plot. This indicates publication bias and the diagram is evaluated together with Duval and Tweedie's trim-and-fill results in Table 4.

Table 4
The results of Duval & Tweedie's trim-and-fill

	Studies Trimmed	Overall Effect	Lower Limit	Upper Limit	Q Value
Observed Values		1.347	1.228	1.466	1102.690
Adjusted Values	48	0.912	0.777	1.046	2379.926

In Table 4, it was observed that 48 studies were added to make the funnel plot symmetrical and the added studies changed the overall effect. In addition, in Rosenthal's fail-safe *N* method, it was observed that the number of missing studies that should be added for the overall effect size to be non-significant was 177019, and this value was greater than the criterion value of 755 (5k+10) (Rosenthal, 1979). When Begg and Mazumdar's rank correlation results were analyzed, it was seen that Kendall's tau value was 0.326 and significant. Finally, in Egger's regression intercept method, the intercept was found to be 3.834 and significant. The fact that these statistics are significant is an indicator of publication bias. When all statistics are evaluated together, it is observed that there is publication bias. Based on this, it can be said that the weighting methods in this study were compared under a condition where publication bias exists.

Meta-Analysis Results

In this study, the effect of weighting with the inverse of the sampling error variance and reliability in the presence of high heterogeneity and publication bias on meta-analysis results was examined. We also examined the effect of the inclusion and exclusion of gray literature on the meta-analysis results and the results are presented in Table 5.

Table 5

The results of meta-analysis in different conditions (weighting methods and gray literature)

Weighting Methods	Number of Effect Sizes	Cohen *d*	SE	Variance	Lower Limit	Upper Limit	Z	p
Gray Literature Included								
Inverse variance	149	1.347	0.061	0.004	1.228	1.466	22.217	0.000
Reliability	149	1.474	0.119	0.014	1.242	1.707	12.426	0.000
Gray Literature Excluded								
Inverse variance	55	1.281	0.076	0.006	1.132	1.431	16.780	0.000
Reliability	55	1.324	0.152	0.023	1.026	1.622	8.705	0.000

When Table 5 was examined, it was seen that the largest overall effect size was obtained in the weighting method with a reliability of 1.474, and the smallest overall effect size was obtained in the weighting method with the inverse of sampling error variance with 1.347 when gray literature is included. When the standard error values were analyzed, it was seen that the lowest standard error value was obtained from weighting with a sampling error variance of 0.061. The highest standard error value was found in weighting by reliability coefficient, which was 0.119. Variance values also changed in parallel with the standard error values. When evaluated in terms of confidence interval, the narrowest confidence interval was found in the sampling error variance method, again in parallel with the standard error. In addition, the confidence interval was wider for the weighting method with the reliability coefficient. When the significance of the overall effect sizes was analyzed, it was observed that the overall effect sizes were significant in both methods. In addition, forest plots of both methods are presented in Appendix A and Appendix B, respectively. When the forest plots were analyzed, it was seen that the primary studies were more homogeneous in terms of confidence intervals due to the narrow range of weights in the reliability weighting method. On the other hand, when the weighting method with sampling error variance was used, it could be said that the forest plot was more heterogeneous due to the wide sample range.

In the case where gray literature was not included, the largest overall effect size was obtained from weighting methods with a reliability coefficient and was found to be 1.324. The lowest overall effect size was found to be 1.281 for the weighting by sampling error variance method. When the standard error values were analyzed, it was seen that the lowest standard error value was obtained from weighting with sampling error variance and was 0.076. The highest standard error value was found in weighting by reliability coefficient, which was 0.152. Variance values also changed in parallel with the standard error values. When the confidence intervals were evaluated, it could be said that the confidence interval was wider when weighting by reliability coefficient than when weighting by sampling error variance. It was observed that the meta-analysis study with the narrowest confidence interval was the meta-analysis using the weighting method with sampling error variance. When the significance of the overall effect sizes was analyzed, it was seen that the overall effect sizes were significant in both methods.

In addition to interpreting the effects of the inclusion and exclusion of gray literature on the meta-analysis results, it is also necessary to interpret the significance of these effects. In this context, Analog ANOVA was conducted to examine the significance of the effects. The results are given in Table 6.

Table 6
Analog ANOVA results of gray literature and articles for weighting methods

Weighting Method		Q values	df (Q)	p
Inverse Variance	Within Group	1101.062	147	0.000
$N_{GrayLiterature}$ **= 94**	Between Groups	1.629	1	0.202
$N_{Manuscript}$ **= 55**	Total	1102.690	148	0.000
Reliability	Within Group	240.668	147	0.000
$N_{GrayLiterature}$ **= 94**	Between Groups	1.597	1	0.206
$N_{Manuscript}$ **= 55**	Total	242.265	148	0.000

When Table 6 was examined, it was seen that the p-values for the intergroup Q values in the inverse of the sampling error variance and reliability weighting methods were 0.202 and 0.206, respectively. In this respect, it was clear that the difference between the average effect size obtained from the studies in the gray literature and the average effect size obtained from the articles was not significant in all weighting methods. Therefore, it can be said that the meta-analysis results obtained with and without the inclusion of gray literature did not differ significantly from each other.

Discussion

When the weighting methods were compared with each other, both when gray literature was included and not included in the meta-analyses, it was seen that the weighting method with the smallest overall effect size was the weighting method with the sampling error variance. The weighting method with the largest overall effect size was the weighting method with a reliability coefficient. The fact that the overall effect size obtained from weighting with sampling error variance is lower than the effect sizes obtained from weighting with reliability coefficient does not indicate that the weighting method with reliability coefficient synthesizes effect size more accurately than the weighting method with sampling error variance. The reason for the difference in the overall effect sizes between the two weighting methods may be that weighting by sampling error variance deals with the sampling error, whereas weighting by reliability coefficient deals not only with sampling error but also with sources of random error, including sampling error. In addition, the fact that the overall effect sizes are larger in the weighting method with reliability coefficient may be due to the fact that, as Rosenthal (1991) states, the contribution of studies that are weaker in terms of quality weight and have smaller effect sizes to the average effect size is less than other studies.

A similar situation is observed when standard error values are examined in the context of weighting methods. It was observed that the standard error values obtained from weighting by reliability coefficient were the highest, while the standard error values obtained from weighting by sampling error variance were the lowest, both in the conditions where gray literature was included and not included. The fact that the standard error values obtained from weighting with sampling error variance were lower than the standard error values obtained from weighting with reliability coefficient can be explained by the fact that it deals only with the dimension of the error arising from the sample. This is because weighting with the reliability coefficient addresses not only sampling error but also other sources of random error sources. Therefore, the standard error values obtained from the weighting methods with sampling error variance and reliability coefficient differ from each other. In parallel with the standard error, the narrowest confidence intervals were observed in the weighting method with sampling error variance in all studies, while the widest confidence intervals were observed in the weighting method with reliability coefficient. This is because the confidence interval is calculated directly using the standard error. The

fact that the lower and upper limit values obtained from weighting with sampling error variance are lower than the other lower and upper limit values and the confidence intervals are narrower can be explained by the fact that only the error arising from the sample is considered in parallel with the overall effect size and standard error.

When the meta-analysis results were compared according to the inclusion and exclusion of gray literature, it was observed that the overall effect size had different values and the overall effect sizes were higher when the gray literature was included. However, it was concluded that this difference was not significant in both weighting methods. Although the difference was not significant, the reason why the overall effect sizes were higher when the gray literature was included might be due to the fact that the effect sizes of the primary studies in the gray literature were larger than the scanned studies. In addition, higher average effect sizes may have been obtained due to the larger sample sizes of these studies where effect sizes might be larger.

Like the overall effect size, the standard error also took different values according to the inclusion of gray literature. In general, standard error values were higher when gray literature was not included. The standard error is expected to decrease as the sample size increases with the inclusion of gray literature. Conn et al. (2003) stated that when gray literature was included, the overall effect size was estimated with less error than when gray literature was not included, which is similar to the results of weighting with sampling error variance and reliability coefficient in this study. Moher et al. (1996), similar to the results of this study, found that there was a slight difference due to the reporting language of the studies but that this was not a significant bias and that the inclusion of non-English language publications may reduce the error and increase the accuracy of estimation. As stated by Conn et al. (2003) and Moher (1996) in their studies and as found in this study, the reason for the decrease in the standard error and more accurate estimations may be the increase in the number of included studies. Hartling et al. (2017) have also observed that the studies included in the gray literature generally constitute a very small part of the meta-analysis sample, and therefore, the results are not affected much by the inclusion of the gray literature. However, in this study, the studies in the gray literature constitute a larger portion of the studies rather than a small portion of the studies. Despite this, the effect of the inclusion of gray literature is not significant and is similar to Hartling et al. (2017). Contrary to the results of this study, Corlett (2011) also stated that ignoring the gray literature might lead to biased results. Although Corlett (2011) did not statistically examine the effect of gray literature, the reason why he made such a suggestion is that he worked in the tropics and gray literature is the only source in the tropics. Based on the findings of this study and the literature, it is obvious that it is important to investigate the impact of gray studies in order to make a correct decision about whether there is bias in a meta-analysis study.

Conclusions, Suggestions and Limitations

The study results showed that the overall effect size changed with the inverse of the sampling error variance and when weighted by reliability. It was also concluded that the standard error was highest when weighted by the reliability coefficient because it included all random errors. In this regard, meta-analysts may also be recommended to try weighting with a reliability coefficient because it is thought that weighting by reliability may provide a more accurate confidence interval.

When the results regarding the inclusion of gray literature were examined, it was observed that the results were not significantly different in the inclusion and exclusion cases. In this study, although there was no significant difference between the overall effect sizes according to the inclusion of gray literature, it is recommended that researchers should also scan the gray literature in all weighting methods since the estimation accuracy will increase due to the lower standard error when gray literature is included.

When the weighting methods were compared with each other, it was seen that weighting with sampling error variance gave the closest results when gray literature was included and not included. Therefore, when weighting with sampling error variance, the exclusion of gray literature may be less important for educational research. However, since clinical research requires more precise results, it may be recommended to include the gray literature since these studies have little differentiation. The weighting

method with the highest differentiation was found to be weighting with a reliability coefficient. Although this differentiation is not significant, there may be significance between the inclusion and exclusion of gray literature in other studies. For this reason, researchers are strongly recommended to review the gray literature and examine the significance of the difference when using weighting with a reliability coefficient.

Within the scope of this study, the results were compared with each other by weighting with reliability coefficient in addition to weighting with sampling error variance used in classical meta-analysis. Other researchers can compare meta-analysis results by formulating different weighting methods or choosing not to weight. They can also contribute to the mathematical formulation of the weighting method with reliability. In addition, other researchers can choose another study topic instead of the effect of the 5E teaching model on science achievement, which was selected as the subject of the meta-analysis study in this study, or they can compare the methods in this study in fields such as sports sciences, health sciences, etc. instead of using data in the field of education.

In the present study, there is a situation of publication bias and high heterogeneity, which are the limitations of the study. Other researchers can examine the method of weighting effect sizes with the reliability coefficient developed in this study under different conditions. For this purpose, they can design a simulation study and test this new method under conditions of different sample sizes, number of studies, estimation methods, heterogeneity, publication bias, fields, etc. As a result, this study is expected to encourage new studies on weighting the measures from which effect sizes are obtained with reliability coefficients in synthesizing studies in meta-analysis and to add the options of the reliability of measures for weighting effect sizes to meta-analysis softwares.

Declarations

Conflict of Interest: No potential conflict of interest was reported by the authors.
Ethical Approval: Permission was obtained from Gazi University Ethics Commission dated 22.02.2021 and numbered E-77082166-302.08.01-33880.
Author Contribution: Yıldız YILDIRIM: conceptualization, investigation, methodology, writing - original draft, formal analysis, visualization, editing. Şeref TAN: conceptualization, investigation, methodology, data curation, supervision, writing - review & editing

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

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

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

References

Alacapınar, F. G., & Ok, M. (2020). Meta-analysis covering studies on problem-based learning. *Research on Education and Psychology*, *4*(Special Issue), 53-73. https://dergipark.org.tr/tr/pub/rep/issue/54042/703272

Altunoğlu, B.D., Bozdemir Yüzbaşıoğlu, H., Candan Helvacı, S., & Kurnaz, M.A. (2020). Genetik kavramlara ilişkin eğitim çalışmalarının meta analiz yöntemi ile incelenmesi. *Batı Anadolu Eğitim Bilimleri Dergisi, 11*(2), 643-661. https://dergipark.org.tr/en/pub/baebd/issue/58594/702868

Arık, S., & Yılmaz, M. (2020). The effect of constructivist learning approach and active learning on environmental education: A meta-analysis study. *International Electronic Journal of Environmental Education, 10*(1), 44-84. https://dergipark.org.tr/tr/pub/iejeegreen/issue/49969/605746

Basit, O. (2020). *Türkiye'de yapılan okul öncesi dönem çocuklarının gelişim alanlarını destekleyici çalışmaların incelenmesi: Bir meta analiz çalışması.* Doktora Tezi, Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Başpınar, M. M. (2021). Türkiye'de yapılan tez çalışmalarında sigara içiminin uyku kalitesi üzerine etkisinin değerlendirilmesi: meta-analiz. *Journal of Turkish Sleep Medicine*, *8*(1), 7-15. https://doi.org/10.4274/jtsm.galenos.2021.98698

Batdı, V., & Batdı, H. (2015). Effect of creative drama on academic achievement: A meta-analytic and thematic analysis. *Educational Sciences: Theory & Practice, 15*(6), 1459-1470. https://doi.org/10.12738/estp.2015.6.0156

Becker, K. H., & Park, K. (2011). Integrative approaches among science, technology, engineering, and mathematics (STEM) subjects on students' learning: A meta-analysis. *Journal of STEM Education Innovation and Research, 12*(5), 23-37. https://doi.org/10.12691/education-2-10-4

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 1088-1101. PMID: 7786990. https://doi.org/10.2307/2533446

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons.

Bozdemir, H., Çevik, E. E., Altunoğlu, B. D., & Kurnaz, M. A. (2017). Astronomi konularının öğretiminde kullanılan farklı yöntemlerin akademik başarıya etkisi: Bir meta analiz çalışması. *Alan Eğitimi Araştırmaları Dergisi*, *3*(1), 12-24. https://dergipark.org.tr/tr/download/article-file/266427

Conn, V. S., Valentine, J. C., Cooper, H. M., & Rantz, M. J. (2003). Grey literature in meta analyses. *Nursing Research*, 52, 256–261. https://doi.org/10.1097/00006199-200307000-00008

Corlett, R. T. (2011). Trouble with the gray literature. *Biotropica*, *43*(1), 3-5. https://doi.org/10.1111/j.1744-7429.2010.00714.x

Duval, S., & Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*(449), 89-98. https://doi.org/10.1080/01621459.2000.10473905

Duval, S., & Tweedie, R. (2000b). Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455-463. https://doi.org/10.1111/j.0006-341x.2000.00455.x.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, *315*(7109), 629-634. https://doi.org/10.1136/bmj.315.7109.629

Englund, G., Sarnelle, O., & Cooper, S. D. (1999). The importance of data-selection criteria: meta-analyses of stream predation experiments. *Ecology*, *80*(4), 1132-1141. https://www.jstor.org/stable/177060?seq=1#metadata_info_tab_contents

Fabiano, G., Schatz, N., Aloe, A., Pelham, W., Smyth, A., Zhao, X., Merrill, B. M., Macphee, F., Ramos, M., Hong, N., Altszuler, A., Ward, L., Rodgers, D. B., Liu, Z., Karatoprak Ersen, R., & Coxe, S. (2021). Comprehensive meta-analysis of attention-deficit hyperactivity disorder psychosocial treatments investigated within between group studies. Review of Educational Research, 91(5), 718–760. https://doi.org/10.3102/00346543211025092

Fuller, J. B., & Hester, K. (1999). Comparing the sample-weighted and unweighted meta-analysis: An applied perspective. Journal of Management, 25(6), 803-828. https://doi.org/10.1177/014920639902500602

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3-8. https://doi.org/10.2307/1174772

Hartling, L., Featherstone, R., Nuspl, M., Shave, K., Dryden, D. M., & Vandermeer, B. (2017). Grey literature in systematic reviews: a cross-sectional study of the contribution of non-English reports, unpublished studies and dissertations to the results of meta-analyses in child-relevant reviews. *BMC Medical Research Methodology*, *17*(1), 1-11. https://doi.org/10.1186/s12874-017-0347-z

Hedges, L. V. & Olkin, I.(1985). *Statistical methods for meta-analysis*. Academic.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3, 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Higgins, J. P., & Green, S. (2008). *Cochrane handbook for systematic reviews of interventions.* The Cochrane Collaboration and John Wiley & Sons.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta analysis. *Statistics in Medicine, 21*(11), 1539-1558. https://doi.org/10.1002/sim.1186

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal*, *327*(7414), 557. https://doi.org/10.1136/bmj.327.7414.557

Hunter, J. E., & Schmidt, F. L. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Beverly Hills, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (2004). Methods of meta-analysis: Correcting error and bias in research findings (2nd ed.). Thousand Oaks, CA: Sage.

Karasar, N. (2013). *Bilimsel araştırma yöntemleri* (25th ed.). Nobel.

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research*, *86*(3), 681-718. https://doi.org/10.3102/0034654315627366

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care

interventions: explanation and elaboration. *Journal of Clinical Epidemiology*, *62*(10), 1-34. https://doi.org/10.1136/bmj.b2700

Marín-Martínez, F., & Sánchez-Meca, J. (2009). Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educational and Psychological Measurement*, *70*(1), 56-73. https://doi.org/10.1177/0013164409344534

Moher, D., Fortin, P., Jadad, A. R., Jüni, P., Klassen, T., Le Lorier, J., ... & Linde, K. (1996). Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. *The Lancet*, *347*(8998), 363-366. https://doi.org/10.1016/s0140-6736(96)90538-3

Mutluer, C., Gündüz, T., Çelikten Demirel, S., & Çakan, M. (2020). *Meta analiz çalışmasında sabit etki modeline karşı rastgele etki modeli.* Presented at the VIIth International Eurasian Educational Research Congress, Eskişehir.

National Research Council. (1992). *Combining information: statistical issues and opportunities for research*. National Academies. https://doi.org/10.17226/20865

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638-641. https://doi.org/10.1037/0033-2909.86.3.638

Rosenthal, R. (1991). Quality-weighting of studies in meta-analytic research. *Psychotherapy Research, 1*(1), 25-28. https://doi.org/10.1080/10503309112331334031

Saraç, H. (2018). Yapılandırmacı yaklaşım öğrenme halkası modellerinin öğrenilen bilgilerin kalıcılığına etkisi: Meta analiz çalışması. *Kastamonu Eğitim Dergisi*, *26*(3), 753-764. https://doi.org/10.24106/kefdergi.413322

Schmidt, F. L., Oh, I. S., & Hayes, T. L. (2009). Fixed-versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, *62*(1), 97-128. https://doi.org/10.1348/000711007X255327

Shuster, J. J. (2010). Empirical vs natural weighting in random effects meta-analysis. *Statistics in Medicine*, *29*(12), 1259-1265. https://doi.org/10.1002/sim.3607

Simpson, R. J. S., & Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *The British Medical Journal*, *2*(2288), 1243-1246. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2355479/

Sosa, G. W., Berger, D. E., Saw, A. T., & Mary, J. C. (2011). Effectiveness of computer-assisted instruction in statistics: A meta-analysis. *Review of Educational Research*, *81*(1), 97-128. https://doi.org/10.3102/0034654310378174

Toraman, Ç. , Çelik, Ö. C. & Çakmak, M. (2018). Oyun-Tabanlı Öğrenme Ortamlarının Akademik Başarıya Etkisi: Bir Meta–Analiz Çalışması . *Kastamonu Eğitim Dergisi , 26* (6) , 1803-1811 . https://doi.org/10.24106/kefdergi.2074

Özdemir, V. (2023). *Okuma becerisi ile sosyoekonomik ve kültürel değişkenler arasındaki ilişkilerin incelenmesi meta-analiz çalışması*. Presented at the X International Eurasian Educational Research Congress, Ankara.

Warfa, A. R. M. (2016). Using cooperative learning to teach chemistry: A meta-analytic review. *Journal of Chemical Education, 93*(2), 248-255. https://doi.org/10.1021/acs.jchemed.5b00608

Xie C., Wang, M., & Hu, H. (2018). Effects of constructivist and transmission instructional models on mathematics achievement in mainland China: A meta-analysis. *Front. Psychol., 9*(1923), 1-18. https://doi.org/10.3389/fpsyg.2018.01923

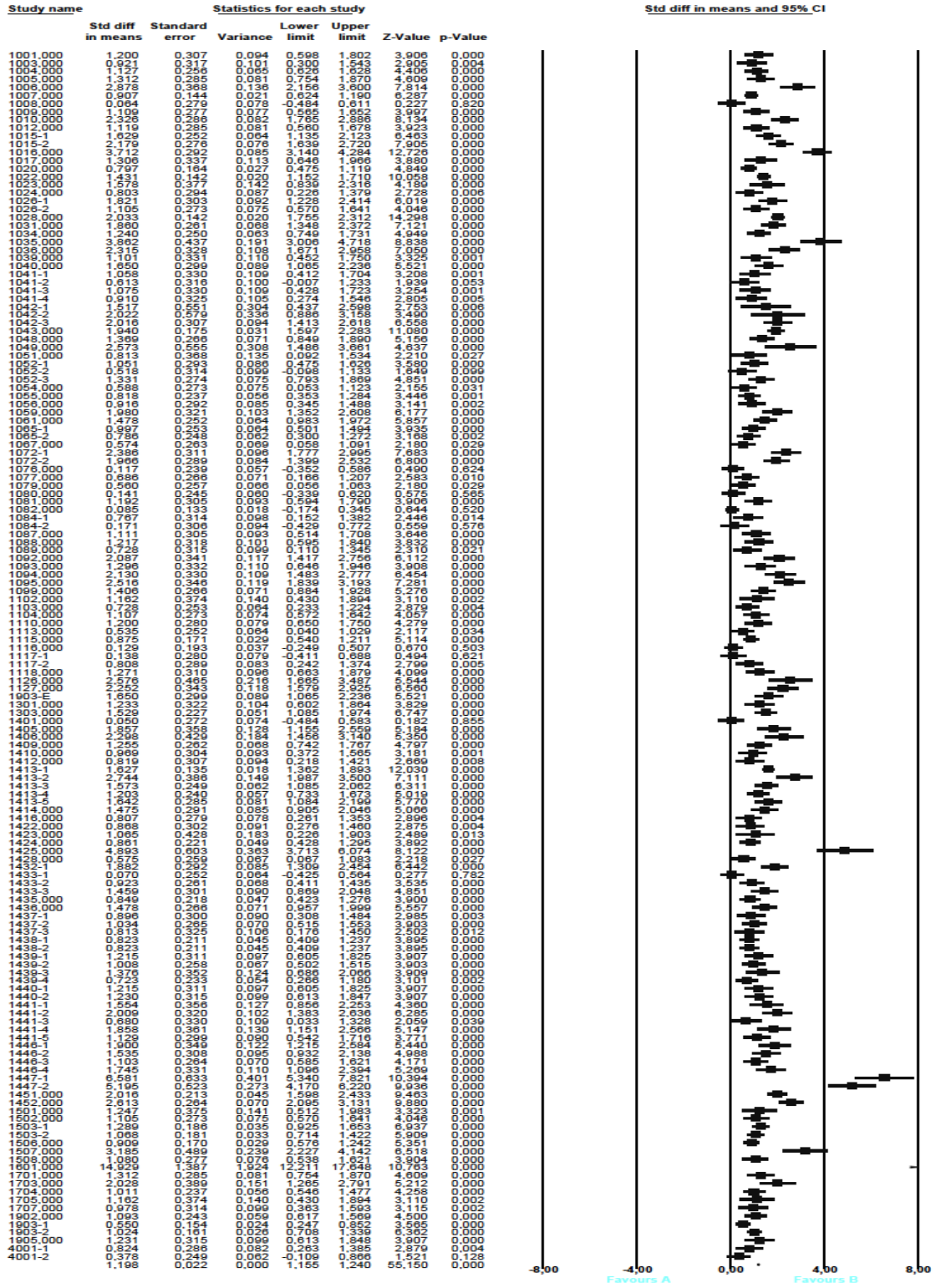
Yeşilpınar Uyar, M. & Doğanay, A. (2018). Öğrenci merkezli strateji, yöntem ve tekniklerin akademik başarıya etkisi: Bir meta-analiz çalışması. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 14*(1), 186-209. https://doi.org/10.17860/mersinefd.334542

Yıldırım, Y., & Şahin, M. G. (2023). How Do Different Weighting Methods Affect the Overall Effect Size in Meta-Analysis?: An Example of Science Attitude in Türkiye Sample. *International Journal of Psychology and Educational Studies, 10*(3), 744-757. https://doi.org/10.52380/ijpes.2023.10.3.1049

APPENDIX

APPENDIX A

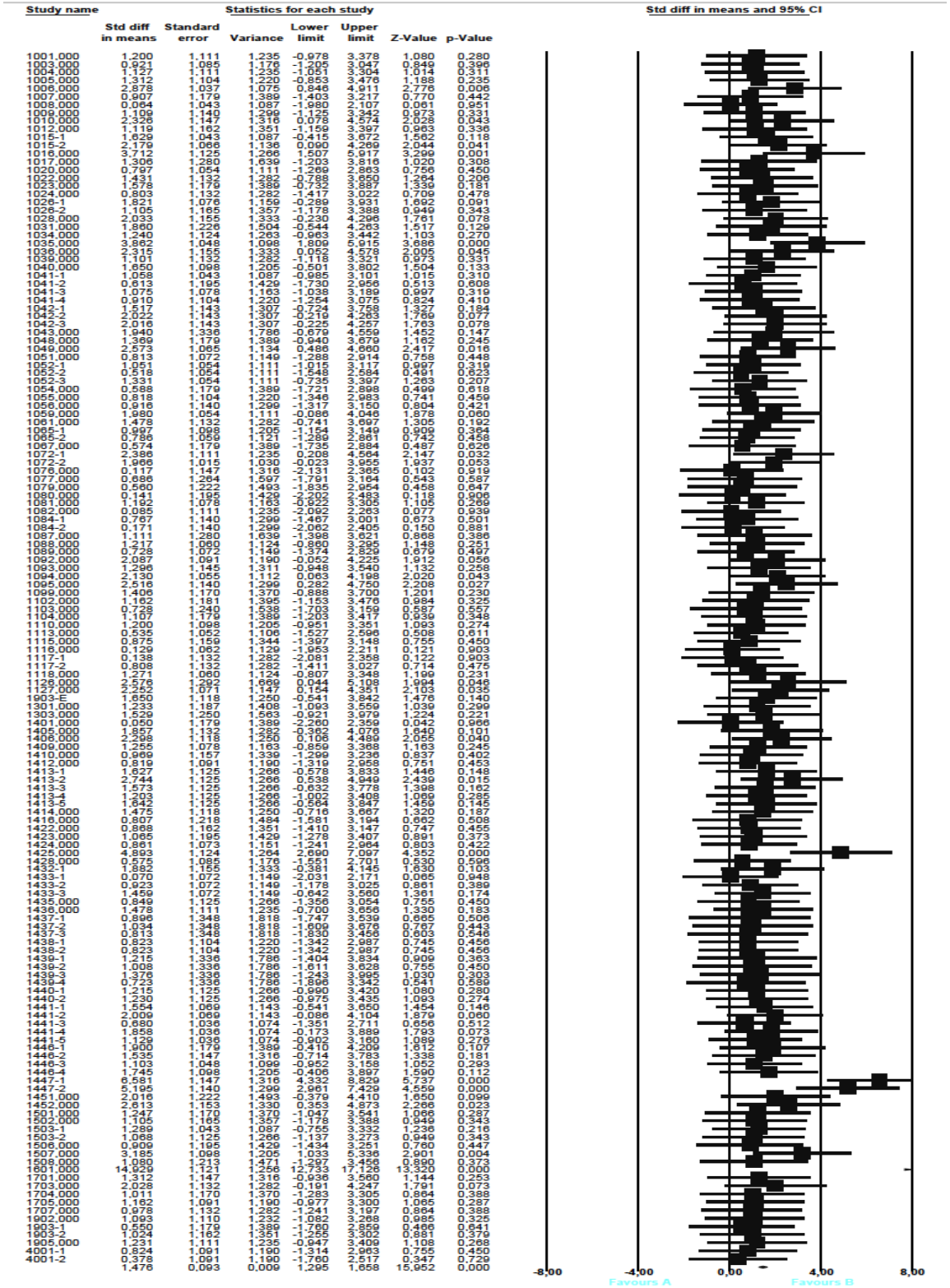
The forest plot for inverse variance method



Study name	Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
1001.000	1.200	0.307	0.094	0.598	1.802	3.906	0.000
1003.000	0.921	0.317	0.101	0.300	1.543	2.905	0.004
1004.000	1.127	0.256	0.065	0.626	1.628	4.406	0.000
1005.000	1.312	0.285	0.081	0.754	1.870	4.609	0.000
1006.000	2.878	0.368	0.136	2.156	3.600	7.814	0.000
1007.000	0.907	0.144	0.021	0.624	1.190	6.287	0.000
1008.000	0.064	0.279	0.078	-0.484	0.611	0.227	0.820
1009.000	1.109	0.277	0.077	0.565	1.652	3.997	0.000
1010.000	2.326	0.286	0.082	1.765	2.886	8.134	0.000
1012.000	1.119	0.285	0.081	0.560	1.678	3.923	0.000
1015-1	1.629	0.252	0.064	1.135	2.123	6.463	0.000
1015-2	2.179	0.276	0.076	1.639	2.720	7.905	0.000
1016.000	3.712	0.292	0.085	3.140	4.284	12.726	0.000
1017.000	1.306	0.337	0.113	0.646	1.966	3.880	0.000
1020.000	0.797	0.164	0.027	0.475	1.119	4.849	0.000
1022.000	1.431	0.142	0.020	1.152	1.710	10.058	0.000
1023.000	1.578	0.377	0.142	0.838	2.318	4.189	0.000
1024.000	0.803	0.294	0.087	0.226	1.379	2.728	0.006
1026-1	1.821	0.303	0.092	1.228	2.414	6.019	0.000
1026-2	1.105	0.273	0.075	0.570	1.641	4.046	0.000
1028.000	2.033	0.142	0.020	1.755	2.312	14.298	0.000
1031.000	1.860	0.261	0.068	1.348	2.372	7.121	0.000
1034.000	1.240	0.250	0.063	0.749	1.731	4.949	0.000
1035.000	3.862	0.437	0.191	3.006	4.718	8.838	0.000
1036.000	2.315	0.328	0.108	1.671	2.958	7.050	0.000
1039.000	1.101	0.331	0.110	0.452	1.750	3.325	0.001
1040.000	1.850	0.399	0.089	1.065	2.636	5.521	0.000
1041-1	1.058	0.330	0.109	0.412	1.704	3.208	0.001
1041-2	0.613	0.316	0.100	-0.007	1.233	1.939	0.053
1041-3	1.076	0.330	0.109	0.428	1.723	3.254	0.001
1041-4	0.910	0.325	0.105	0.274	1.546	2.805	0.005
1042-1	1.517	0.551	0.304	0.437	2.598	2.753	0.006
1042-2	2.022	0.579	0.336	0.886	3.158	3.490	0.000
1042-3	2.016	0.307	0.094	1.413	2.618	6.558	0.000
1043.000	1.940	0.175	0.031	1.597	2.283	11.080	0.000
1048.000	1.369	0.266	0.071	0.849	1.890	5.156	0.000
1049.000	2.573	0.308	0.095	1.968	3.178	8.361	0.000
1051.000	0.813	0.368	0.135	0.092	1.534	2.210	0.027
1052-1	1.051	0.293	0.086	0.475	1.626	3.580	0.000
1052-2	0.516	0.314	0.099	-0.098	1.133	1.649	0.099
1052-3	1.331	0.274	0.075	0.793	1.869	4.851	0.000
1054.000	0.588	0.273	0.075	0.053	1.123	2.155	0.031
1055.000	0.818	0.237	0.056	0.353	1.284	3.446	0.001
1056.000	0.916	0.292	0.085	0.345	1.488	3.141	0.002
1059.000	1.980	0.321	0.103	1.352	2.608	6.177	0.000
1061.000	1.478	0.252	0.064	0.983	1.972	5.857	0.000
1065-1	0.947	0.253	0.064	0.501	1.392	3.735	0.000
1065-2	0.786	0.248	0.062	0.300	1.272	3.168	0.002
1067.000	0.574	0.263	0.069	0.058	1.091	2.180	0.029
1072-1	2.366	0.311	0.096	1.777	2.955	7.603	0.000
1072-2	1.966	0.289	0.084	1.399	2.532	6.800	0.000
1076.000	0.117	0.239	0.057	-0.352	0.586	0.490	0.624
1077.000	0.886	0.266	0.071	0.166	1.207	2.583	0.010
1079.000	0.990	0.257	0.066	0.056	1.063	2.180	0.029
1080.000	0.141	0.245	0.060	-0.339	0.620	0.575	0.565
1081.000	1.192	0.306	0.093	0.594	1.790	3.906	0.000
1082.000	0.085	0.133	0.018	-0.174	0.345	0.644	0.520
1084-1	0.767	0.314	0.098	0.152	1.382	2.446	0.014
1084-2	0.171	0.306	0.094	-0.429	0.772	0.559	0.576
1087.000	1.111	0.305	0.093	0.514	1.708	3.646	0.000
1088.000	1.217	0.318	0.101	0.595	1.840	3.832	0.000
1089.000	0.728	0.315	0.099	0.110	1.345	2.310	0.021
1092.000	2.087	0.341	0.117	1.417	2.758	6.112	0.000
1093.000	1.296	0.332	0.110	0.646	1.946	3.908	0.000
1094.000	2.130	0.330	0.109	1.483	2.777	6.454	0.000
1095.000	2.516	0.346	0.119	1.839	3.193	7.281	0.000
1099.000	1.406	0.266	0.071	0.884	1.928	5.276	0.000
1102.000	1.162	0.374	0.140	0.430	1.894	3.110	0.002
1103.000	0.728	0.253	0.064	0.233	1.224	2.879	0.004
1104.000	1.107	0.273	0.074	0.572	1.642	4.057	0.000
1110.000	1.200	0.280	0.079	0.650	1.750	4.279	0.000
1113.000	0.535	0.252	0.064	0.040	1.029	2.117	0.034
1115.000	0.875	0.171	0.029	0.540	1.211	5.114	0.000
1116.000	0.129	0.193	0.037	-0.249	0.507	0.670	0.503
1117-1	0.138	0.280	0.079	-0.411	0.688	0.494	0.621
1117-2	0.808	0.289	0.083	0.242	1.374	2.799	0.005
1118.000	1.271	0.310	0.096	0.663	1.879	4.099	0.000
1126.000	2.576	0.465	0.216	1.665	3.487	5.544	0.000
1127.000	2.252	0.343	0.118	1.579	2.925	6.560	0.000
1903-E	1.650	0.299	0.089	1.065	2.236	5.521	0.000
1301.000	1.233	0.322	0.104	0.602	1.864	3.829	0.000
1303.000	1.529	0.227	0.051	1.085	1.974	6.747	0.000
1401.000	0.050	0.272	0.074	-0.484	0.583	0.182	0.855
1405.000	1.857	0.358	0.128	1.155	2.559	5.184	0.000
1406.000	2.298	0.429	0.184	1.456	3.140	5.350	0.000
1409.000	1.255	0.262	0.068	0.742	1.767	4.797	0.000
1410.000	0.969	0.304	0.093	0.372	1.565	3.181	0.001
1412.000	0.819	0.307	0.094	0.218	1.421	2.669	0.008
1413-1	1.627	0.135	0.018	1.362	1.893	12.030	0.000
1413-2	2.744	0.386	0.149	1.987	3.500	7.111	0.000
1413-3	1.573	0.249	0.062	1.085	2.062	6.311	0.000
1413-4	1.203	0.240	0.057	0.733	1.673	5.019	0.000
1413-5	1.642	0.285	0.081	1.084	2.199	5.770	0.000
1414.000	1.475	0.291	0.085	0.905	2.046	5.066	0.000
1416.000	0.807	0.279	0.078	0.261	1.353	2.896	0.004
1422.000	0.988	0.302	0.091	0.276	1.468	2.873	0.004
1423.000	1.065	0.428	0.183	0.226	1.903	2.489	0.013
1424.000	0.861	0.221	0.049	0.428	1.295	3.892	0.000
1425.000	4.893	0.603	0.363	3.713	6.074	8.122	0.000
1428.000	0.575	0.259	0.067	0.067	1.083	2.218	0.027
1432-1	1.882	0.292	0.085	1.309	2.454	6.442	0.000
1433-1	0.070	0.252	0.064	-0.425	0.564	0.277	0.782
1433-2	0.923	0.261	0.068	0.411	1.435	3.535	0.000
1433-3	1.459	0.301	0.090	0.869	2.048	4.851	0.000
1435.000	0.849	0.218	0.047	0.423	1.276	3.900	0.000
1436.000	1.478	0.266	0.071	0.957	1.998	5.557	0.000
1437-1	0.896	0.300	0.090	0.308	1.484	2.985	0.003
1437-3	1.034	0.265	0.070	0.515	1.553	3.903	0.000
1437-5	0.813	0.326	0.106	0.176	1.450	2.502	0.012
1438-1	0.823	0.211	0.045	0.409	1.237	3.895	0.000
1438-2	0.823	0.211	0.045	0.409	1.237	3.895	0.000
1439-1	1.215	0.311	0.097	0.605	1.825	3.907	0.000
1439-2	1.008	0.258	0.067	0.502	1.515	3.903	0.000
1439-3	1.376	0.352	0.124	0.686	2.066	3.909	0.000
1439-4	0.723	0.233	0.054	0.266	1.180	3.101	0.002
1440-1	1.215	0.311	0.097	0.605	1.825	3.907	0.000
1440-2	1.230	0.315	0.099	0.613	1.847	3.907	0.000
1441-1	1.554	0.356	0.127	0.856	2.253	4.360	0.000
1441-2	2.009	0.320	0.102	1.383	2.636	6.285	0.000
1441-3	0.680	0.330	0.109	0.033	1.328	2.059	0.039
1441-4	1.858	0.361	0.130	1.151	2.566	5.147	0.000
1441-5	1.109	0.249	0.062	0.621	1.598	4.461	0.000
1446-1	1.900	0.349	0.122	1.215	2.584	5.440	0.000
1446-2	1.535	0.308	0.095	0.932	2.138	4.988	0.000
1446-3	1.103	0.264	0.070	0.585	1.621	4.171	0.000
1446-4	1.745	0.331	0.110	1.096	2.394	5.269	0.000
1447-1	6.581	0.633	0.401	5.340	7.821	10.394	0.000
1447-2	5.195	0.523	0.273	4.170	6.220	9.936	0.000
1451.000	2.018	0.213	0.045	1.598	2.433	9.463	0.000
1452.000	2.613	0.264	0.070	2.095	3.131	9.880	0.000
1501.000	1.247	0.375	0.141	0.512	1.983	3.323	0.001
1502.000	1.105	0.273	0.075	0.570	1.641	4.046	0.000
1503-1	1.289	0.186	0.035	0.925	1.653	6.937	0.000
1503-2	1.068	0.181	0.033	0.714	1.422	5.909	0.000
1506.000	0.909	0.170	0.029	0.576	1.242	5.351	0.000
1507.000	3.185	0.489	0.239	2.227	4.143	6.518	0.000
1508.000	1.080	0.277	0.076	0.538	1.621	3.904	0.000
1601.000	14.929	1.387	1.924	12.211	17.648	10.763	0.000
1701.000	1.312	0.285	0.081	0.754	1.870	4.609	0.000
1703.000	2.028	0.389	0.151	1.265	2.791	5.212	0.000
1704.000	1.011	0.237	0.056	0.546	1.477	4.258	0.000
1705.000	1.162	0.374	0.140	0.430	1.894	3.110	0.002
1707.000	0.978	0.314	0.099	0.363	1.593	3.115	0.002
1902.000	1.093	0.243	0.059	0.617	1.569	4.500	0.000
1903-1	0.750	0.154	0.024	0.447	1.052	4.865	0.000
1903-2	1.024	0.161	0.026	0.708	1.339	6.362	0.000
1905.000	1.231	0.315	0.099	0.613	1.848	3.907	0.000
4001-1	0.824	0.286	0.082	0.263	1.385	2.879	0.004
4001-2	0.378	0.249	0.062	-0.108	0.866	1.521	0.128
	1.198	0.022	0.000	1.155	1.240	55.150	0.000

APPENDIX B

The forest plot for reliability coefficient method



ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology Research Article; 2023; 14(3); 266-286

Analysis of Peer and Self-Assessments Using the Many-facet Rasch Measurement Model and Student Opinions

Seda DEMİR*

Abstract

The aim of this study is to analyze the peer and self-assessments of higher education students' oral presentation skills with the many-facet Rasch measurement model and to determine students' opinions on peer and self-assessment. In the study, the convergent parallel method, one of the mixed-method research approaches, was used. The study group consisted of 11 university students studying at a state university in the 2022-2023 academic year. The FACETS program was used to analyze the data. The three facets identified in the study were the assessee (11 students), the assessor (11 students), and the items (16 items). Therefore, 11 participants scored (peer and self-assessment) on a 16-item assessment form. In addition, students' opinions on peer and self-assessment were obtained through three open-ended interview questions prepared by the researcher. According to the results of the study, it was determined that there was a statistically significant difference between the students in terms of their oral presentation skills, between the assessors in terms of their strictness/generosity in scoring, and between the criteria (items) in terms of the level of difficulty in realization. In addition, the participant opinions obtained from each interview question were analyzed through themes and sub-themes formed according to the general thoughts on peer and self-assessment, experiences, and whether the participants considered themselves as a reliable rater or not. In terms of practice, it can be suggested to provide detailed and enlightening information to students before peer and/or self-assessment in the classroom environment, and to give quick feedback to those who have not done the assessment appropriately. In addition, the reasons for the biases identified in peer and self-assessments in the current study can be investigated in future studies.

Keywords: Peer assessment, Self-assessment, Many-facet Rasch measurement model, Oral presentation skills

Introduction

Effective assessment of the educational process can be considered as one of the basic requirements that contribute to the discovery and development of students' true potential. In this context, it can be said that alternative (performance-based) assessment methods that support a student-centered education approach offer the opportunity to assess students' different aspects, learning styles or abilities and thus provide a more comprehensive learning process. In this process, students are generally expected to be able to apply knowledge to real-world situations. Peer and self-assessment, alternative methods, are among the most widely researched assessment methods in the literature that encourage students' active participation in assessment processes and develop their self-confidence (Falchikov & Goldfinch, 2000). According to Cheong et al. (2023) peer and self-assessment are processes in which students judge the quality of their peers or their own work. Peer assessment involves students assessing each other's work and providing feedback (Evans et al., 1993), while self-assessment allows students to observe and assess their own learning processes (Boud & Falchikov, 1989). Therefore, self-assessment is known to be closely related to reflection (Yan & Brown, 2017) and during peer assessment, students can benefit from both giving and receiving feedback (Hoo et al., 2021; Liu & Carless, 2006; Lundstrom & Baker, 2009). Peer and self-assessment practices help students identify their own strengths and weaknesses and create motivation for lifelong learning (Hanrahan & Isaacs, 2001; Panadero et al., 2023; Sande & Godino-Llorente, 2014).

Various studies in the literature indicate that peer assessment provides cognitive, affective, pedagogical, and metacognitive benefits to students (Butler & Winne, 1995; Nicol & Macfarlane-Dick, 2006; Orluwene & Ekim, 2020; Tseng & Tsai, 2007; Zhan et al., 2023). According to Brown (2004), if the

* Assist. Prof. Dr., Tokat Gaziosmanpasa University, Faculty of Education, Tokat-Turkey, seddadmr@gmail.com, ORCID ID: 0000-0003-4230-5593

To cite this article:
Demir, S. (2023). Analysis of peer and self-assessments using the many-facet rasch measurement model and student opinions. *Journal of Measurement and Evaluation in Education and Psychology, 14*(3), 266-286. https://doi.org/10.21031/epod.1344196

Received: 16.08.2023
Accepted: 27.09.2023

students selected for peer assessment are adequately informed and the assessment process is planned correctly, this approach can save considerable time for the teachers and the process, give students a chance to learn in depth, and help support students' higher-order cognitive activities. In their study, Crisp and Ward (2008) stated that peer assessment is a method of assessment that offers constant feedback while also improving students' academic achievement, class participation, and motivation. Panadero et al. (2023) examined the relationships between peer assessment and intrapersonal and interpersonal factors through a systematic review on peer assessment. According to the study, there are six intrapersonal factors including motivation, self-efficacy, emotions, trust in the self as an assessor, fairness, and comfort; and five interpersonal factors including social connections, trust in the other as an assessor, psychological safety, value diversity/congruence, and interdependence in peer assessment. In Boud and Falchikov's (1989) study, it was determined that students with high achievement gave themselves lower grades in self-assessment, while students with low achievement gave themselves higher grades. In this direction, Hanrahan and Isaacs (2001) emphasized in their study that students should be informed and trained in detail about how to assess self-assessment. Yan et al. (2021) examined the effect of self-assessment on academic performance in a meta-analysis study. The results of the study show that the overall effect of self-assessment was significant. Topping et al. (2000) compared the reliability of peer assessment and self-assessment results and found that peer assessment was a more reliable method than self-assessment. Gürlen et al. (2019) examined the reliability coefficients of teacher, self- and peer evaluations of primary school students with the help of generalizability theory. The results of the study revealed that the variance component estimated for the student main effect was the largest component of the total variance. According to the literature, it is possible to say that peer and self-assessment practices help students develop high-level skills such as taking ownership of their own learning and abilities to think critically, creatively, and analytically, solve problems, present information clearly, and conduct research. In addition, they aim to involve students in the assessment process and support student learning rather than grading (Pantiwati & Husamah, 2017; Stefani, 1994). In addition, in the study conducted by Cheong et al. (2023) with undergraduate students, peer and self-assessment were used together for the academic writing task and it was investigated how self-assessment complemented peer assessment. As a result of the study, self-assessment has been found to complement peer assessment in five ways: it guides students to make corrections when peer assessment is incomplete; when students have access to peer assessment, self-assessment effectively supports peer assessment; even when a student has access to quality peer assessment, self-assessment complements peer assessment because of the different reflections in the two processes; self-assessment can support peer assessment on issues related to social emotional burdens; and self-assessment also complements peer assessment in that it benefits high and low-achieving students. Therefore, it can be said that complementing peer assessment with self-assessment is an effective solution to overcome possible problems that may be encountered in the peer assessment process, such as students' limited ability to provide feedback and non-objective assessment.

In addition to peer and self-assessment, presentations are frequently preferred, especially in higher education, in order to ensure students' active participation in the course. In measuring presentation skills, the use of rubrics provides an objective assessment process and offers the opportunity to give more qualified and meaningful feedback. According to Fete et al. (2017), meaningful feedback enables students to be more responsible for their behaviors while ensuring their personal growth and development. In addition, students who know the scoring criteria produce better-quality work (Liu et al., 2001; Lu & Law, 2011). However, scoring may not always be based solely on performance. Various sources of variance (factors/facets) may be involved in scoring , which may negatively affect the validity of scoring (Prieto & Nieto, 2014). As stated in the study conducted by Gu (2020), there are some problems in peer assessment, such as students' hesitation to criticize their peers and students' doubts about each other's ability to make correct decisions. In addition, there are also studies showing that students score themselves lower or higher than they should be in self-assessments (Mumpuni et al., 2022; Semerci, 2011a). Therefore, it is necessary to examine in depth whether students make objective evaluations or not. For this reason, in the current study, peer and self-assessments of students' oral presentation skills were analyzed using a many-facet Rasch measurement model in which assessor characteristics were added as facets to the measurement model.

The Rasch measurement model (Linacre, 1993) is utilized to objectively calculate the precise intervals between options in tests, scales, and rubrics. This method aids in determining the interval unit with greater precision and accuracy (Elhan & Atakurt, 2005). In the many-facet Rasch measurement model, there is no facet limitation and it is suitable for multiple scoring (Eckes, 2005). With this model, the facets (such as assessor, assessee, and items) that may affect the predictions for the latent trait measured are considered. Semerci (2011a) analyzed faculty member, peer, and self-assessments within the framework of doctoral qualifications with the Rasch measurement model and determined the differences observed in student performances, jury strictness/generosity, and the difficulty/ease of the tasks expected to be performed. Similarly, Köse et al. (2016) analyzed rater, criterion, and presentation skills using peer assessments of student presentations with the many-facet Rasch measurement model. As a result, it was exemplified that the many-facet Rasch measurement model is an alternative measurement model that can be used to determine student performance. Mumpuni et al. (2022), in their study aiming to analyze how peer assessment takes place, concluded that students have the ability to make peer assessments and make their assessments objectively.

When the related literature is examined, it is seen that there are qualitative studies on peer assessment and/or self-assessment, or quantitative studies designed according to the many-facet Rasch measurement model. However, depending on the rubric used, the interview questions or the study group, the results of the studies differ from each other, and the need for new research arises.

As a result, the aim of this study was to analyze the peer and self-assessments of the students taking the Teaching Probability and Statistics course on oral presentation skills with the many-facet Rasch measurement model and to determine the students' views on peer and self-assessment.

In line with this overall objective;

i. General analysis of opinions on oral presentation skills in the Teaching Probability and Statistics course,

ii. Analysis of assessors' rigor/generosity,

iii. Task difficulty analysis of oral presentation skills,

iv. Assessor bias analysis,

v. Qualitative data obtained from interviews with all students participating in the study will be analyzed.

It is thought that the current study will provide students and educators with ideas about the objective use of peer and self-assessments, which are alternative assessment methods. In general, the accuracy of an assessment is directly related to the validity of the previous assessment. Accordingly, the fact that the scoring of oral presentation skills contains bias errors will directly affect the validity negatively. In addition, students' views on peer and self-assessment also enriched the study in terms of qualitative data. Therefore, this study, designed as a mixed research, is thought to be important in terms of its contribution to related literature.

Methods

This study employs the convergent parallel method, one of the mixed method research approaches, to collect and analyze both quantitative and qualitative data simultaneously (Creswell, 2014). In this regard, the peer and self-assessments of higher education students' oral presentation skills with the many-facet Rasch measurement model and students' opinions on peer and self-assessment were merged for a more complete understanding.

Study Group

In the Rasch measurement model, there is no assumption that sample statistics generalize to the population (Linacre, 1993). Therefore, the study group was determined for the research. The same study

group took part in both quantitative and qualitative parts of the research. The study group consisted of 11 university students (four males and seven females) who took the Teaching Probability and Statistics course at a state university in the 2022-2023 academic year. The convenience sampling method was used for the selection of participants. The participants were selected on a purely voluntary basis among the students who made oral presentations within the course.

Data Collection

In the study, all students enrolled in the Teaching Probability and Statistics course were first told about the goals of peer and self-assessment and what they should pay attention to in their assessments with the use of the peer and self-assessment guide prepared by the researcher. Then, 11 students who volunteered to participate in the study were identified. The participants both made oral presentations and then self-assessed and made peer assessments by listening to other oral presentations. In order to score the students' oral presentation skills, the "Oral Presentation Skills Peer Assessment Form" developed by the researcher and the "Oral Presentation Skills Self-Assessment Form" consisting of the same items were used. The items used to evaluate students' oral presentation skills are presented in Table 1.

Table 1

Items Used to Evaluate Oral Presentation Skills

Heading	Skills
Form of presentation	1. The subject is emphasized with main lines.
	2. Fluent language is used.
	3. The tone of voice is used correctly.
Content	4. The ideas put forward on the subject are supported by solid evidence.
	5. The examples given on the subject are interesting and original.
	6. There are no contradictory explanations about the subject.
Understanding the Subject and Participation in Discussions	7. The subject is fully understood.
	8. Sufficient technical information is given.
	9. The subject is presented in a convincing way.
	10. An overall evaluation including important points has been made.
	11. Thoughts have been expressed clearly.
	12. Questions and comments have been successfully answered.
	13. The more complex parts of the subject have been sufficiently emphasized.
Communication Skills and Time Management	14. Good communication has been established with the audience.
	15. The listeners who asked questions or made comments were not interrupted.
	16. Time was used efficiently and there were no problems in time management.

As seen in Table 1, the peer and self-assessment forms consisted of 16 items. These items-range from 1 (very inadequate) to 5 (very adequate) and scored on a five-point Likert scale. In addition, students' views on peer and self-assessment were obtained through three open-ended interview questions prepared by the researcher.

Data Analysis

In the quantitative part of the study, students' peer and self-assessments were analyzed using the many-facet Rasch measurement model. In this model, multiple sources of variability (ability, item, rater, situation, task, etc.) can be analyzed simultaneously and independently (Mulqueen et al., 2000; Sudweeks et al., 2005). In addition, the analysis results obtained from the sample are not intended to be generalized to the population (Linacre, 1993). In the analysis of the data, the FACETS program developed by Linacre (1993, 2023), which deals with three facets as ability, item/measure/task, and rater

as a general use, was used. The three facets identified in the current study were the assessee (11 students), the assessor (11 students), and the items (16 items). Therefore, 11 participants scored (peer and self-assessment) on the 16-item assessment form and a total of 1936 (11x11x16) data were obtained. 11 students both scored as assessors and were scored as assessee. In this context, Assessee 1 and Assessor A, Assessee 2 and Assessor B, Assessee 3 and Assessor C, Assessee 4 and Assessor D, Assessee 5 and Assessor E, Assessee 6 and Assessor F, Assessee 7 and Assessor G, Assessee 8 and Assessor H, Assessee 9 and Assessor I, Assessee 10 and Assessor J, and Assessee 11 and Assessor K are codes representing the same student. The interpretation of peer or self-assessments was made by considering these codes. In addition, each item was coded according to the order in which it appeared in the form, for example, Item 1, Item 2. With the many-facet Rasch measurement model, the study explored factors such as assessors' fairness, bias, the ease or difficulty of criteria, and identified which students had stronger oral presentation skills based on the established criteria.

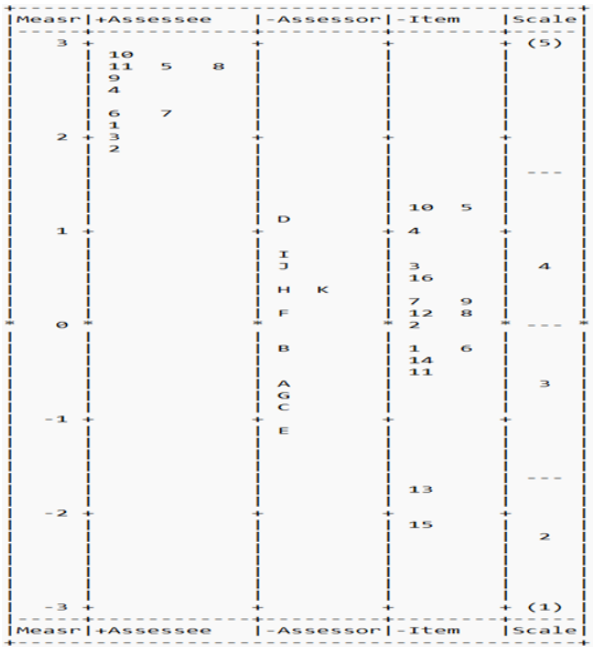
In the qualitative part of the study, themes and sub-themes were formed by content analysis of 11 students' responses to three open-ended questions to determine their views on peer and self-assessment. The opinions of the participant students were given in the form of quotations. For the quotations, the codes representing the assessors (Assessor A, Assessor B, Assessor C, Assessor D, Assessor E, Assessor F, Assessor G, Assessor H, Assessor I, Assessor J, and Assessor K) were used to represent the same students. For the reliability of the study, firstly, the participants' responses to each open-ended question were combined in a single document. Then, another expert was consulted for the codes and themes determined based on these responses. In addition, the confirmation of the findings obtained from the current study with the participants can be considered as evidence for the internal validity of the study, and the fact that the findings are compatible with the literature can be regarded as evidence for the external validity of the study.

Results

In the analysis of oral presentation skills with the many-facet Rasch measurement model, three facets (assessee, assessor, and items) were used. The Wright Map containing general information about these facets is presented in Figure 1.

Figure 1

Calibration Map of the Distribution of Assessee, Assessor and Items



When Figure 1 is examined, it is seen that the oral presentation students are ranked on the same logit scale according to their ability level (assessee), the strictness/generosity of the raters (assessor) and the difficulty level of the tasks (items). In this distribution, the assessee facet is ranked from the best oral presentation performance to the lowest, the assessor facet is ranked from the strictest rater to the most generous, and the item facet is ranked from the most difficult task to the easiest, from top to bottom. Accordingly, in terms of the oral presentation, Assessee 10 had the best performance, while Assessee 2 had the lowest performance. However, Assessor D gave the strictest assessment and Assessor E gave the most generous assessment. Based on the data obtained, it can be said that Assessor D, who gave the strictest score, realized a moderately good oral presentation, while Assessor E, who gave the most generous score, realized the second best oral presentation. In addition, Item 5 and Item 10 were determined as the most difficult items (the most difficult criterion/task to perform), while Item 15 was determined as the easiest item (the easiest criterion/task to perform).

The detailed measurement report on the oral presentation skills of 11 undergraduate students (assessees) who took the Teaching Probability and Statistics course is presented in Table 2.

Table 2

Oral Presentation Skills Measurement Report of the Assessees

Assessee	Logit	Standart Error	Infit MnSq	Outfit MnSq
10	2.87	0.17	1.07	1.18
8	2.79	0.16	0.97	0.82
5	2.69	0.16	1.03	1.22
11	2.69	0.16	0.93	0.76
9	2.58	0.15	1.09	1.09
4	2..45	0.14	1.08	0.86
7	2.31	0.14	0.78	0.69
6	2.27	0.14	0.89	0.93
1	2.15	0.13	1.11	1.54
3	1.98	0.13	1.08	1.10
2	1.88	0.12	1.03	1.15
RMSE= 0.14	sd= 0.28	Separation= 1.96	Strata= 2.94	Reliability= 0.79
chi-square= 56.2	df= 10	*p*= 0.00		

When Table 2 is examined, it is seen that the Root Mean Square Standard Error- RMSE of the measurement values is calculated as 0.14 and the standard deviation is calculated as 0.28. In addition, the separation index was calculated as 1.96 and the strata value as 2.94. The discrimination index shows the ability of the measurement tool in Rasch analysis to distinguish participants with different ability levels (Linacre, 1994). The higher the discrimination index, the better the measurement tool is understood to be (Mumpuni et al., 2022). The strata value calculated as approximately three indicates that there are three groups of students in terms of oral presentation skills. The reliability coefficient obtained from the analysis shows that the students assesses in terms of oral presentation skills are ranked with 0.79 confidence. In addition, according to the chi-square test results ($\chi 2$ =56.2, df=10, p=0.00), the null hypothesis was rejected. Therefore, it was determined that there was a statistically significant difference between the students in terms of oral presentation skills.

The general order of the students assessed in terms of oral presentation skills from the best performer to the lowest performer is as follows: 10, 8, 5, 11, 9, 4, 7, 6, 1, 3, 2. Here, Assessees 5 and 11 have the same performance. In self-assessment, this order is: Assessee 10, 1, 5, 8, 9, 3, 7, 2, 6, 4, 11, and Assessee 1, 5, 8, 9 perform similarly. On the other hand, in peer assessment, Assessee 10, 8, 5, 11, 6, 9, 4, 7, 1, 3, 2, and Assessee 4, 7 have the same performance. When these rankings are analyzed collectively, it can be said that it is a remarkable finding that the position of Assessee 10 did not change.

In Rasch analysis, the fit and misfit values of the facets, which indicate the degree of fit between the data and the model, are also calculated. The out-of-fit statistic, which is more sensitive to unexpected extreme values compared to the in-fit statistic, is equal to the mean squares of the residuals between the observed data and the expected values (Randall and Engelhard, 2009). A fit statistic of 1 indicates that the variance between the data is greater than expected; a fit statistic of less than 1 indicates that the variance between the data is less than expected. The range of 0.5 to 1.5 for fit statistics is the range of values considered appropriate as an indicator of accurate and effective measurements (Turner, 2003; Wright & Linacre, 1994). Accordingly, it can be concluded that the model fits the data obtained from all the assessed data.

The detailed measurement report for the 11 assessors who scored oral presentation skills is presented in Table 3.

Table 3

Assessors' Strictness/Generosity Measurement Report

Assessor	Logit	Standart Error	Infit MnSq	Outfit MnSq
D	1.18	0.11	0.93	0.85
I	0.80	0.11	1.01	0.89
J	0.66	0.12	0.96	0.87
H	0.40	0.13	1.06	0.81
K	0.39	0.13	1.22	1.58
F	0.16	0.14	1.12	1.11
B	-0.26	0.16	0.73	1.41
A	-0.62	0.19	0.79	0.80
G	-0.69	0.19	1.53	1.68
C	-0.94	0.21	0.82	0.77
E	-1.08	0.23	0.72	0.56
RMSE= 0.16	sd= 0.71	Separation= 4.41	Strata= 6.21	Reliability= 0.95
chi-square= 235.5	df= 10	*p*= 0.00	Inter-rater exact agreements= 60.1%	

According to Table 3, with a measurement value of 1.18, it is seen that Assessor D is the strictest, and with a measurement value of -1.08, Assessor E is the most generous in scoring. Therefore, the general order of the assessors is from the most strict to the most generous in terms of scoring oral presentation skills is Assessor D, I, J, H, K, F, B, A, G, C, and E. In addition, as seen in Table 3, the standard error of the measurement values was calculated as 0.16, the standard deviation as 0.71, the separation index as 4.41 and the strata value as 6.21. The strata value calculated as approximately six indicates that there are six groups of assessors in terms of strictness/generosity in scoring oral presentation skills. The reliability coefficient of 0.95 obtained from the Rasch analysis shows that the students who were assessed in terms of their strictness/generosity were ranked with very high reliability. Moreover, according to the results of the chi-square test ($\chi 2$ =235.5, df=10, p=0.00), the null hypothesis was rejected. Therefore, there was a statistically significant difference between the assessors regarding their strictness/generosity in scoring.

When the congruent and incongruent values of the facets in the Rasch analysis are analyzed, it is seen that only the incongruent value of Assessor K and Assessor G is outside the recommended value range (0.5 to 1.5 range). Accordingly, it can be said that there are some inconsistencies in the scoring of Assessor K and Assessor G. Finally, according to Table 3, the absolute inter-rater agreement value was calculated as 60.1%.

A detailed measurement report on the criteria/tasks (items) in the form used to assess students' oral presentation skills is presented in Table 4.

Table 4

Measurement Report of the Items Used to Assess Oral Presentation Skills

Item	Logit	Standart Error	Infit MnSq	Outfit MnSq
5	1.25	0.13	0.81	0.83
10	1.21	0.13	1.19	1.09
4	1.05	0.13	0.81	0.79
3	0.61	0.15	0.97	1.12
16	0.56	0.15	1.57	1.26
7	0.24	0.17	0.79	1.02
9	0.21	0.17	0.85	0.78
8	0.07	0.18	0.74	0.68
12	0.07	0.18	0.89	0.74
2	0.00	0.18	0.99	1.14
6	-0.24	0.20	1.81	2.34
1	-0.28	0.20	0.74	0.57
14	-0.37	0.21	1.05	0.92
11	-0.50	0.22	1.01	0.70
13	-1.76	0.38	1.03	1.52
15	-2.11	0.45	0.98	0.97
RMSE= 0.22	sd= 0.88	Separation= 4.00	Strata= 5.67	Reliability= 0.94
chi-square= 243.5	df= 15	*p*= 0.00		

When the item measurement report in Table 4 is examined, according to the measurement values obtained, it is seen that the most difficult criterion (the criterion with the lowest rate of high score) is Item 5: "The examples given on the subject are interesting and original." with a measurement value of 1.25, followed by Item 10: "A general assessment including the important points of the subject was made." The easiest criterion (with the highest rate of high scores) was Item 15: "Listeners who asked questions or made comments were not interrupted." with a measurement value of -2.11. This was followed by Item 13: "The more complex parts of the topic were sufficiently emphasized." A visual of these results is given in Figure 1. In addition, as seen in Table 4, the standard error of the measurement values was calculated as 0.22, the standard deviation as 0.88, the separation index as 4.00 and the strata value as 5.67. In addition, the calculated reliability value is quite high at 0.94. The significant results of the chi-square test ($\chi 2$= 243.5, df=15, p=0.00) indicate a statistically significant difference between the difficulty levels of the criteria. When the fit statistics for the criteria are examined, it is observed that all criteria except the sixth criterion are between acceptable values within and outside the acceptable fit. Accordingly, it can be said that only the sixth criterion is an obstacle to data-model fit.

With the help of the many-facet Rasch analysis, unexpected responses obtained with the measurement tool can also be identified. Unexpected responses show which rater scored the response of which individual in an unexpected way. In addition, it provides information (such as training of raters and revision of items) for determining the sources of decreased reliability and planning the measurement process more reliably (Güler, 2014; Nakamura, 2002).

In the current study, a sample of unexpected responses between the assessee, the assessor, and the item is presented in Table 5.

Table 5

Unexpected Responses between Assessee, Assessor, and Item

Sequence	Score	Expected	StRes	Assessor	Asseessee	Item
102	2	4.9	-8.2	1	G	6
454	2	4.9	-7.7	3	G	6
278	2	4.8	-7.2	2	G	6
111	4	5.0	-6.9	1	G	15
1613	4	5.0	-6.7	10	B	13
733	4	5.0	-6.1	5	B	13
1437	4	5.0	-5.7	9	B	13
879	4	5.0	-5.2	5	K	15
909	4	5.0	-4.9	6	B	13
1622	4	5.0	-4.3	10	C	6
1581	4	4.9	-4.1	9	K	13
175	4	4.9	-3.9	1	K	15
974	3	4.8	-3.9	6	F	14
1298	4	4.9	-3.9	8	E	2
1490	3	4.8	-3.7	9	F	2
527	4	4.9	-3.6	3	K	15
1590	4	4.9	-3.6	10	A	6
112	3	4.7	-3.5	1	G	16
694	3	4.7	-3.5	4	K	6
351	4	4.9	-3.4	2	K	15
1053	4	4.9	-3.4	6	K	13
1474	4	4.9	-3.4	9	E	2
1884	3	4.7	-3.4	11	H	12
1625	4	4.9	-3.3	10	C	9
1754	2	4.4	-3.3	10	K	10
173	4	4.9	-3.2	1	K	13
174	3	4.7	-3.2	1	K	14
775	4	4.9	-3.2	5	E	7
962	3	4.7	-3.2	6	F	2
1667	3	4.7	-3.2	10	F	3
542	4	4.9	-3.1	4	A	14
1271	4	4.9	-3.1	8	C	7
1262	4	4.9	-3.0	8	B	14

When the standardized StRes values given in Table 5 are examined, it is seen that all of them have a minus (-) sign. Accordingly, it can be said that all of the unexpected data resulted from the fact that some students gave lower than expected scores to other students. It is seen that the most unexpected data stemmed from the score given by Assessor G to Item 6 for Assessee 1. Here, while the expected value for Item 6: "There were no contradictory explanations about the topic." was 4.9, Assessor G gave 2 points to Assessee 1 for this item and the standardized StRes value was calculated as -8.2. In addition, the first four most unexpected data belong to Assessor G; all were scored below the expected value. As

a remarkable finding from the study, this means that Assessor G performed worse than expected. It is also seen that the most recurrent rater in terms of giving unexpected scores was Assessor K, and the top three items with the highest recurrence of bias were Item 7, Item 6, and Item 15, respectively.

When the data presented in Table 5 related to self-assessment are analyzed, it is seen that Assessor F gave himself lower scores than expected for Item 14: "Good communication with the audience was established." (three points were given while the expected score was 4.8) and Item 2: "Fluent language was used" (three points were given while the expected score was 4.7). Similarly, Assessor E gave herself a lower than expected score for Item 7: "The topic was fully understood" (four points were given when the expected score was 4.9).

The bias analysis of self- and peer-assessors is presented in Table 6.

Table 6

Assessor and Assessee Interaction Bias Report

Observed Score	Expected Score	Obs-Exp Average	Bias	Standart Error	*z* Score	Infit MnSq	Outfit MnSq	Assessee	Assessor
67	73.69	-0.42	-0.97	0.34	-2.84	1.3	1.2	6	F
63	70.61	-0.48	-0.89	0.32	-2.80	0.8	0.7	8	D
60	67.32	-0.46	-0.77	0.31	-2.46	0.9	0.9	3	I
73	77.09	-0.26	-1.01	0.42	-2.41	0.4	0.4	7	G
63	69.02	-0.38	-0.68	0.32	-2.12	0.8	0.7	9	D
60	66.40	-0.40	-0.66	0.31	-2.12	0.9	0.9	2	I
73	76.63	-0.23	-0.85	0.42	-2.02	2.3	3.7	1	G
76	69.72	0.39	1.16	0.53	2.19	1.2	1.0	1	J
78	71.74	0.39	1.63	0.73	2.24	1.0	0.9	9	I
80	67.99	0.75	3.58<	1.43	2.51	0.0	0.0	4	D
chi-square= 156.7		df= 121		*p*= 0.02					

The fact that the z scores given in Table 6 are outside the commonly accepted range of -2 to +2 points to interaction bias between assessors and assessees. Assessor F gave a significantly (p<0.05) rigid scoring by giving himself 67 points when he should have given himself approximately 74 points in his self-assessment. Similarly, Assessor G made a significantly rigid peer assessment for Assessee 7 and Assessee 1. In addition, it is seen that Assessor D and Assessor I gave lower scores to some students than expected in their peer assessments and made a significantly strict scoring, while they gave higher scores to some students and made a significantly generous scoring. Assessor J gave a significantly generous peer assessment for Assessee 1.

In addition to the analyses conducted with the many-facet Rasch measurement model, the participants' responses to three questions regarding their views on peer and self-assessment were also analyzed and themes and sub-themes were formed.

1. The themes and sub-themes determined in line with the answers to the question "What are your general thoughts about the peer/self-assessment practice you participated in?" are presented in Table 7 and Table 8, respectively.

Table 7

General Thoughts about the Peer Assessment Practice

Themes	Sub Themes
Benefits of Peer Assessment	Gaining a critical perspective
	Increasing awareness of responsibility
	Being respectful for different ideas
	Increasing motivation
	Gaining different perspectives
	Focusing on the learning process without worrying about grades
	Developing reasoning skills
	Developing empathy skills
	Recognizing professional values
	Gaining an objective perspective
	Providing students with the drive to be better
	Improving academic performance
	Developing reflective thinking skills
	Gaining awareness of assessment
	Supporting future development
Characteristics of the Assessment Process	Performance based
	Objectivity
	Process and product oriented
	Based on criteria
Learning Process	Increasing teacher-student coordination
	Ensuring effective participation in the lesson
	Providing a student-centered learning environment
	Creating work discipline
	Providing feedback
	Sharing responsibility for learning
Problems in the Peer Assessment Process	Time consuming
	Complexity
	Performing assessments in line with prejudices
	Lack of experience
	Increasing the level of anxiety

Some sample responses reflecting the participants' general thoughts about the peer assessment practice are given below.

Assessor B: "While doing peer assessment, I had the opportunity to assess the process as well as the product. I think that with this practice, the course was carried out in teacher-student coordination and student-centered. Although at first I found peer assessment complex and difficult due to my lack of experience and my prejudices against some of my friends, I realized that the assessments I made improved my ability to empathize and reason over time.

Assessor C: "While doing peer assessment, it is useful to know that it is important to make an assessment. In other words, it has many benefits both for ourselves and for our friends we assess. From our own point of view, we see that it develops critical thinking. For our friends, we see that it is important for them to see their shortcomings and good sides."

Assessor H: "Peer assessment makes the lesson environment more productive by making the lesson more active and attentive. I think peer assessment should be done for every lesson. The only negative aspect I can say is that the peer assessment process is a bit laborious and time-consuming. Other than that, I think it is a good assessment that should be done."

Assessor K: "I think this practice is useful for us because we make presentations by taking into consideration which criteria our friends who listen to the presentation may pay attention to while

assessing. I also think that we cannot be fully objective in scoring the individuals with whom we are in closer contact and this is the disadvantage of the application."

Table 8

General Thoughts on Self-Assessment Practices

Themes	Sub Themes
Benefits of Self-Assessment	Developing self-awareness
	Recognizing mistakes/deficiencies
	Recognizing strengths and weaknesses
	Developing self-criticism
	Improving oral communication skills
	Improving presentation performance
	Improving metacognitive thinking strategy use
	Developing creativity skills
	Improving decision-making skills
	Creating a perception of success
	Feeling valued
	Contribution to lifelong learning
	Tracking the development process
	Increasing self-confidence
	Developing multiple perspectives
	Providing personal development
	Reinforcing learning
	Creating cognitive awareness
	Developing awareness of democracy
Learning Process	Encouraging active participation in the lesson
	Taking responsibility for own learning
	Increasing the efficiency of the course
	Providing professional development
	Developing metacognition about their own learning
Problems in the Self-Assessment Process	Not assessing their own performance objectively
	Being overly critical and scoring rigidly
	Being too generous in scoring
	Not being conscious enough
	Loss of self-confidence
	Reluctance to learn

Some sample responses reflecting the participants' general thoughts about the self-assessment practice are given below.

Assessor A: "I think that self-assessment is a study developed for us to notice our mistakes or shortcomings. I believe that self-assessment will shed light on our future studies and enable us to take care not to make the same mistakes again and to continue our studies in this direction. Thanks to self-assessment, we have developed a metacognitive perspective on our learning and performance by taking responsibility for our own learning."

Assessor B: "I can honestly say that this practice leaves the person alone with himself/herself. And in this way, the person wants to be more honest with himself/herself and makes his/her assessment accordingly. Therefore, I can say that I found this practice useful. The biggest difficulty I had while trying to make an objective self-assessment was trying not to be more optimistic or pessimistic towards myself than I should be."

Assessor D: "The self-assessment practice made a great contribution to my ability to look at myself objectively and criticize myself. It enabled me to discover myself and see my strengths and weaknesses.

It contributed to gaining a realistic perspective and being impartial. I also think that self-assessment is very important not only in lessons but also in every aspect of life."

Assessor J: "Thanks to the self-assessment, I had the opportunity to realize where I was lacking and what I could do to improve myself. Self-assessment will help me perform better in other presentations by improving myself."

2. The themes and sub-themes determined in line with the responses to the question "What are your experiences with peer/self-assessment?" are presented in Table 9 and Table 10, respectively.

Table 9

Experiences with Peer Assessment

Themes	Sub Themes
Positive Experiences	Developing an empathic approach
	Gaining a critical perspective
	Listening to lesson effectively
	Analyzing the lesson process
	Objective thinking
	Gaining high-level cognitive skills
	Improving social relations
	Progression of competencies
	Improving communication skills
	Fair assessment
	Development of presentation skills
	Development of teaching skills
	Identifying misconceptions
	Developing a sense of responsibility
	Providing permanent learning
	Handling the process holistically and analytically
	Developing research skills
	Interacting with the environment
	Providing feedback
Awareness-Building Experiences	Importance of criteria-based assessment
	Recognizing the importance of making assessments independent from personal feelings and thoughts
	Identifying knowledge gaps
	Impact of peer assessment on social relationships
	Importance of fair/objective assessment

Some sample responses reflecting the participants' experiences in peer assessment are given below.

Assessor A: I realized that peer assessment is a difficult task, especially because we are at similar ages and when it comes to the negative aspects of your friends whom you like very much, whom you are sincere with, it is more difficult to point out these aspects. I gained a more critical perspective. I closed the deficiencies in myself by seeing the deficiencies of my friends. I made an effort to be fair and since I tried to assess from an objective point of view, my learning developed in parallel with this. I based my peer assessment on certain criteria. I learned that such assessments are very necessary. Finally, I realized that peer assessment is not as easy as it seems.

Assessor B: I have developed critical thinking skills and gained experience by assessing the work of my peers. I had never listened to someone's oral presentation before and reached a conclusion or seen the shortcomings of this person and thought about how to overcome these shortcomings while I was explaining. Peer assessment provided me with the opportunity to be objective and to analyze a person or myself from an objective point of view, thus forming the basis of my future experiences. The notes I

took during the lesson for assessment purposes and focusing on my friend who made the presentation made me listen to the lesson more carefully and made the lesson more productive, making my learning more permanent.

Assessor F: The peer assessment practice reminded me that my responsibility for the lesson continues. I think it contributed to my development in terms of objective assessment. In addition, since we need to have knowledge on the subject presented while performing these assessments, it directed me to listen to the presentation more effectively. While doing peer assessment, I tried to look at both positive and negative aspects at the same time. Although I avoided making comparisons between individuals, I realized that at first I filled out the form a little bit influenced by the presentation of the previous presenter.

Assessor I: I realized that in order to analyze the process correctly in peer assessment, the presentation should be listened to carefully. I saw that the assessments of almost all presenters were close to each other when they were listened to carelessly. I think that the presentation should be listened to with focus and calm mind to catch the details.

Table 10

Experiences with Self-Assessment

Themes	Sub Themes
Experiences Supporting Development	Questioning the level of self-efficacy
	Being open to development
	Developing planning skills
	Taking responsibility for own learning
	Developing the ability to make observations
	Striving for perfection
	Contribution to organizing the learning environment
	Providing academic development
	Developing affective skills
	Improving time management
	Discovering different learning methods
	Increasing attention level
	Creating active learning environment
	Providing in-depth learning
	Developing self-regulation strategies
	Creating a desire to learn
	Recognizing aspects open for improvement
	Gaining experience in the learning process
	Making original inferences
	Preparing instructional content
	Developing creative thinking skills
Various Educational Experiences	Complexity of self-assessment
	Revealing one's potential
	The difficulty of conducting objective scoring
	Necessity of process management
	Need for assessment away from comparisons
	Cognitive adaptation to the process

Some sample responses reflecting participants' experiences of self-assessment are given below.

Assessor C: The self-assessment practice contributed positively to my learning process by helping me discover myself and recognize my strengths and weaknesses. It paved the way for me to objectively and realistically assess my own performance and development throughout my life. In addition, self-assessment reminded me that my responsibility for the course continues even after I finish my presentation. If I make progress in my next presentation in terms of the issues I observe in myself and need to work on, the self-assessment practice will have made a concrete contribution to my learning.

Assessor G: "To be honest, assessing myself was more difficult than assessing someone else, but it was also useful for me to see my mistakes. I realized that when I was doing self-assessment, I was doing it by comparing myself with my other friends. Instead of assessing myself, I saw that I was ranking myself from the most successful to the least successful. When

I realized this, I did my self-assessment from the beginning. In the meantime, I approached myself with the same tolerance as I did when assessing my other friends."

Assessor H: "I think I made a good presentation, but there may be shortcomings. I think that self-assessment improved my research skills and contributed to my permanent learning. Examining the process holistically and analytically and working in a planned way before the presentation helped me to cope with my excitement. While doing self-assessment, I realized that one can give feedback even to oneself, and that while we see ourselves positively at certain points, we have mistakes at certain points. I believe that this application is suitable for eliminating these mistakes."

Assessor I: "I realized that self-assessment is actually a difficult task and that one can improve oneself according to some criteria while considering oneself adequate. I think that self-assessment enables us to manage time more easily before or during the presentation and improves self-regulation skills after the presentation. I have experienced different learning methods. I think that my creative thinking skills have improved thanks to the research and studies I have done in order to make a more effective presentation. Self-assessment has enabled me to improve my self-control, knowledge, understanding and skills and to gain the experience of looking at myself objectively even in different areas. It also gave me the experience of understanding each other in the relationship with my fellow listeners, respecting different opinions, etc."

3. Would you describe yourself as a reliable assessor when doing peer/self-assessment? Why? The themes and sub-themes determined in line with the answers given to the question are presented in Table 11 and Table 12, respectively.

Table 11

Whether the Participant Considers Him/herself Reliable in Peer Assessment

Themes	Sub Themes
Characteristics of a Self-Reliable Assessor	Compliance with the principle of impartiality
	Making assessments in line with objective criteria
	Having professional experience
	Considering only the performance
	Performing rational assessment
	The ability to utilize prior knowledge
	Mastering alternative assessment techniques
	The ability to think critically
	Having a collaborative perspective
	Being respectful for the person being assessed
	Having a constructive attitude
	Having ability to make comparisons

Table 11

Whether the Participant Considers Him/herself Reliable in Peer Assessment (Continued)

Themes	Sub Themes
Characteristics of a Partially ReliableAssessor	Not being sure about their assessments
	Lack of self-confidence
	Thinking of missing something due to inattention
	Thinking that assessments may need correction
	Having a competitive perspective
	Influenced by group dynamics
	Feeling incompetent for assessment
	Seeing oneself as inadequate for assessment
	Inability to act impartially

Some sample responses reflecting the participants' views on whether they consider themselves reliable in peer assessment are given below.

Assessor D: I define myself as a reliable assessor because I have always looked at people and situations objectively. I have not hesitated to emphasize my friends' shortcomings or strengths.

Assessor E: Yes. I consider myself to be a reliable assessor because I think I was objective in assessing even the people I was closest to. I tried to be very careful and attentive during the assessments.

Assessor F: Yes, I define myself as a reliable assessor because I listened to everyone's presentations in the group that week in line with the criteria in the scale and reflected my own views transparently in the practice by critically and analytically filtering my mind.

Assessor G: I don't think I'm completely reliable, but I would say I'm mostly reliable because I haven't done a lot of negative assessments, I'm not sure about the assessments I've done because I'm not fully qualified to assess.

Assessor J: I define it partially because as I listened to my friends, I looked at their performances in the presentation and revised the assessment scale of those whom I thought I was unfair in my previous assessments and corrected the places where I needed to make corrections. However, I may not have answered the assessment scale completely correctly for the places I missed or could not listen to, so I think I am a partially reliable assessor.

Assessor K: Of course. I listened carefully to my friends who made presentations and scored them after assessing whether the given criteria were met or not.

Table 12

Whether the Participant Considers Him/herself Reliable in Self-Assessment

Themes	Sub Themes
Requirements for Reliable Self-Assessment	Avoiding overly generous scoring
	Objectivity
	Being open to criticism
	Transparency
	Acting independently from prejudices
	Integrity
	Empathic thinking skills
	Having belief in benefits of fair assessment
	Being constructive
	Being realistic
	Having introspective skills

Table 12

Whether the Participant Considers Him/herself Reliable in Self-Assessment (Continued)

Themes	Sub Themes
Factors Affecting Self-AssessmentNegatively	Ignoration of deficiencies
	Being more tolerant/generous with oneself
	Experiencing cognitive contradiction
	Perfectionism
	Past experiences
	Lack of goal-oriented assessment
	Defensive attitude

Some sample responses reflecting the participants' views on whether they consider themselves reliable in their self-assessment are given below.

Assessor D: I define myself as reliable. Because I looked at the events objectively in my assessment. I judged myself impartially. I did not include contradictory statements.

Assessor E: Yes, I do. Because when I assessed myself, I assessed myself by taking into account my deficiencies.

Assessor F: Yes, I define myself as a reliable assessor. While sharing my personal views, I transparently conveyed what I experienced during the practices. I tried to concretize my views with additional explanations and examples I gave for clarity.

Assessor G: I don't think I am very reliable, people tend to consider themselves as perfect, I believe that people who look at me from the outside can be more objective.

Assessor J: Yes, because I think I assess myself as transparently as possible and I think I am a reliable assessor because what is important for me is to recognize my deficiencies and mistakes.

Assessor K: Yes, I see myself as a good assessor because I commented on my own performance objectively.

Discussion

In this study, the results of peer and self-assessment of 11 students' oral presentation skills in an undergraduate course using a 16-item rubric were analyzed using the many-facet Rasch measurement model. In addition, the opinions of the students participating in the study regarding peer and self-assessment were also determined simultaneously. In the current study, first of all, the data calibration map was examined to obtain general information about the relationship between the facets (assessee, assessor, and items) used in the many-facet Rasch measurement model (Nakamura, 2000) and it was seen that all facets were sorted on the same logit ruler.

The results of the study showed that there were statistically significant differences between the students' oral presentation skills, the assessors' strictness/generosity in scoring, and the criteria's (items') level of difficulty in realization. In support of this finding, in many studies in the literature (Baştürk, 2008; Baştürk, 2010; Köse et al., 2016; Mumpuni et al., 2022; Semerci, 2011a; Semerci, 2011b; Semerci et al., 2013; Uyanık et al., 2019; Yüzüak et al., 2015), it was determined that different rater characteristics created statistically significant differences between raters.

According to the oral presentation skills measurement report, the compliance statistics were among the desired values. According to the overall, peer and self-assessments, it was observed that the rankings from the best-performing student to the lowest-performing student changed in general. However, it is noteworthy that the ranking of the top-performing Assessee 10 remained the same in both peer and self-assessment. It can be interpreted that this situation indicates that the reliability of the ranking of the Assessee 10 is higher.

According to the strictness/generosity measurement report of the assessors, it was found that Assessor D was the most strict and Assessor E was the most generous in scoring. In addition, the non-compliance value of Assessor K and Assessor G was outside the desired value range. This can be interpreted as some inconsistencies in scoring of Assessor K and Assessor G. This problem can be solved by giving extra training to Assessor K and Assessor G on peer and self-assessment. In support of this finding, in most of the studies in the literature using the many-facet Rasch model (Atılgan, 2005; Baştürk, 2008, Baştürk, 2010; Semerci, 2011a, Semerci, 2011b, Akın & Baştürk, 2012; Semerci et al. 2013; Uyanık et al. 2019; Yüzüak et al. 2015), it was stated that the raters can sometimes be objective and sometimes biased.

According to the measurement report of the items used to assess oral presentation skills, the most difficult criterion is item 5: "The examples given on the topic are interesting and original." The easiest criterion is item 15: "Listeners who asked questions or made comments were not interrupted." which can be handled under the heading of communication skills and time management. In addition, the agreement statistics for Item 6: "There were no contradictory explanations about the topic." were outside the desired value range. Therefore, it can be interpreted that this item with double negativity is not a suitable item for measuring oral presentation skills. The reason for this situation may be that the item contains double negativity, both conceptual (contradictory explanation statement) and structural (not done statement).

When the unexpected responses between the assessee, assessor and item were analyzed, it was determined that Assessor G and Assessor K gave lower scores than expected in peer assessment and showed a poor performance. When the unexpected responses were analyzed in terms of self-assessment, it was seen that Assessor F and Assessor E gave themselves lower scores than expected in some items. In support of these results, when the assessee and assessor interaction bias report was examined, it was seen that Assessor F made a significantly strict self-assessment and Assessor G made a significantly strict peer assessment for some assessees. In this case, how Assessor G, Assessor K, Assessor F and Assessor E made sense of the items and how they scored them can be investigated and feedback can be given on how to make appropriate peer and self-assessment. Thus, these unexpected situations can be eliminated. However, it is seen that the first three items with the highest recurrence of bias are Item 7: "The topic was fully understood.", Item 6 and Item 15, respectively. Therefore, it can be said that the assessment forms can be further improved by reviewing and revising these items. Based on the results obtained, it can be said that examining unexpected responses is very useful in improving peer and self-assessment practices.

In the light of the results obtained from the quantitative part of the current study, which was designed as a mixed research, it can be interpreted that the many-facet Rasch measurement model provides very useful information in measurement studies where there is more than one rater and the facets determined will be examined in detail. In the qualitative part of the study, the participants' responses to three open-ended questions were analyzed to determine their views on peer and self-assessment. Regarding the first question, the participants' general thoughts about the peer assessment practice were grouped under four themes: Benefits of Peer Assessment, Characteristics of the Assessment Process, Learning Process, Problems Experienced in the Peer Assessment Process. For their general thoughts on self-assessment, three themes were identified as Benefits of Self-Assessment, Learning Process, and Problems Experienced in the Self-Assessment Process. In the second question, in which the opinions of the participants about their experiences were taken, the experiences for peer assessment were grouped under two themes as Positive Experiences and Awareness-Building Experiences, and the experiences for self-assessment were grouped under two themes as Experiences Supporting Development and Various Educational Experiences. In the third question, which asked whether the participants defined themselves as a reliable assessor, two themes were identified for peer assessment: Characteristics of a Self-Reliable Assessor and Characteristics of a Partially Reliable Assessor, and for self-assessment: Requirements for Reliable Self-Assessment and Negative Factors Affecting Self-Assessment. When the participant opinions obtained from the third question were compared with the results of the many-facet Rasch measurement model, it was seen that qualitative and quantitative partially supported each other. As a result of the analysis conducted with the many-facet Rasch measurement model, Assessor D, who was

determined as the strictest rater, Assessor E, who was determined as the most generous rater, Assessor K, who was seen to score more strictly than expected in peer assessment, and Assessor F, who was seen to score more strictly than expected in self-assessment, stated in the interview that they considered themselves as a reliable assessor, which contradicts these findings. In addition, Assessor G, who was found to have some inconsistencies in his scoring according to quantitative data, stated that he did not consider himself as a fully reliable rater in both peer and self-assessment. Therefore, it can be said that quantitative and qualitative data support each other for Assessor G.

As a result, the many-facet Rasch measurement model highlights through the designated facets, which assessors perform the bias, its source, and direction. In addition, with this study, it was tried to develop suggestions that can be effective in minimizing the errors that may be encountered in the scoring process and minimizing these errors. Participants' views are related to general thoughts and experiences about peer and self-assessment and awareness of bias in scoring. There may be many different reasons for the biases observed in peer and self- assessments. The reasons for the identified biases can be investigated in future studies. In terms of practice, it can be suggested to give detailed and enlightening information to the students before the peer and/or self-assessment in the classroom environment and to give quick feedback to those who have not done the assessment appropriately. Thus, possible biases can be minimized and students' assessment skills and indirectly the teaching process can be improved. It is recommended to employ peer assessment as an impartial instrument for assessing student performances in teaching and learning practices. It is suggested that more applications and experimental investigations related to peer assessment should be conducted in the future.

Declarations

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

Ethical Approval: This study was approved by the Ethical Committee of Tokat Gaziosmanpasa University Social and Human Sciences Researches dated 13.06.2023 and numbered E-46052777-100-307436.

References

Akın, Ö. ve Baştürk, R. (2012). The evaluation of the basic skills in violin training by many facet Rasch model. *Pamukkale University Journal of Education, 31*(1), 175-187. https://dergipark.org.tr/en/pub/pauefd/issue/11112/132860

Atılgan, H. (2005). Analysis of Special Ability sellection examination for music education department using many-facets Racsh measurement (İnönü University case). *Eurasian Journal of Educational Measurement, 0*(20), 62-73. https://web.s.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=0&sid=1c78c0f3-93e0-43b5-9499-026878d2fcb1%40redis

Baştürk, R. (2008). Applying the many facet Rasch model to evaluate PowerPoint presentation performance in higher education. *Assessment and Evaluation in Higher Education, 33*(4), 431-444. https://doi.org/10.1080/02602930701562775

Baştürk, R. (2010). Bilimsel araştırma ödevlerinin çok yüzeyli Rasch ölçme modeli ile değerlendirilmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 1*(1), 51-57. https://dergipark.org.tr/en/pub/epod/issue/5808/77254

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. https://doi.org/10.1007/BF00138746

Brown, S. (2004). Assessment for learning. *Learning and Teaching in Higher Education, 5*(1), 81-89. https://eprints.glos.ac.uk/3607/

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245-281. https://doi.org/10.3102/00346543065003245

Cheong, C. M., Luo, N., Zhu, X., Lu, Q., & Wei, W. (2023). Self-assessment complements peer assessment for undergraduate students in an academic writing task. *Assessment & Evaluation in Higher Education, 48*(1), 135-148. https://doi.org/10.1080/02602938.2022.2069225

Creswell, J. W. (2014). *A concise introduction to mixed methods research*. SAGE.

Crisp, V., & Ward, C. (2008). The development of a formative scenario-based computer assisted assessment tool in psychology for teachers: The PePCAA project. *Computers & Education, 50*(4), 1509-1526. https://doi.org/10.1016/j.compedu.2007.02.004

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal, 2*(3), 197-221. https://doi.org/10.1207/s15434311laq0203_2

Elhan, A. H., & Atakurt, Y. (2005). Why is it necessary to use Rasch analysis when evaluating measures? *Ankara Üniversitesi Tıp Fakültesi Mecmuası, 58*(1), 47-50. https://doi.org/10.1501/Tipfak_0000000134

Evans, A. T., Mcnutt, R. A., Fletcher, S. W., & Fletcher, R. H. (1993). The characteristics of peer reviewers who produce good-quality reviews. *Journal of General Internal Medicine, 8*(8), 422-428. https://doi.org/10.1007/bf02599618

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287-322. https://doi.org/10.3102/00346543070003287

Fete, M. G., Haight, R. C., Clapp, P., & McCollum, M. (2017). Peer evaluation instrument development, administration, and assessment in a team-based learning curriculum. *American Journal of Pharmaceutical Education, 81*(4), 1-10. https://doi.org/10.5688/ajpe81468

Gu, C. (2020). Student Peer Assessment. *Review of Educational Theory, 3*(2), 74-78. https://doi.org/10.30564/ret.v3i2.1762

Güler, N. (2014). Analysis of open-ended statistics questions with many facet Rasch model. *Eurasian Journal of Educational Research, 55*, 73-90. http://dx.doi.org/10.14689/ejer.2014.55.5

Gürlen, E. , Boztunç Öztürk, N. & Eminoğlu, E. (2019). Investigation of the reliability of teachers, self and peer assessments at primary school level with Generalizability Theory. *Journal of Measurement and Evaluation in Education and Psychology, 10*(4), 406-421. http://dx.doi.org/10.21031/epod.583891

Hanrahan, S. J., & Isaacs, G. (2001). Assessing self- and peer-assessment: The students' views. *Higher Education Research & Development, 20*(1), 53-70. https://doi.org/10.1080/07294360123776

Hoo, H. T., Deneen, C., & Boud, D. (2022). Developing student feedback literacy through self and peer assessment interventions. *Assessment & Evaluation in Higher Education, 47*(3), 444-457. https://doi.org/10.1080/02602938.2021.1925871

Köse, İ. A., Usta, H. G., & Yandı A. (2016). Evaluation of presentation skills by using many facets Rasch model. *Bolu Abant İzzet Baysal University Journal of Faculty of Education, 16*(4), 1853-1864. https://dergipark.org.tr/en/pub/aibuefd/issue/28550/304600

Linacre, J. M. (1993). Rasch-based generalizability theory. *Rasch Measurement Transactions, 7*(1), 283-284. https://www.rasch.org/rmt/rmt71h.htm

Linacre, J. M. (1994). *Many-facet rasch model* (2nd ed.). Mesa Press.

Linacre, J. M. (2023). A user's guide to FACETS: Rasch-model computer programs (Program manual 3.85. 1). Chicago, IL. https://www.winsteps.com/a/Facets-Manual.pdf

Liu, N. F., and D. Carless. 2006. "Peer feedback: The learning element of peer assessment." *Teaching in Higher Education 11*(3), 279-290. https://doi.org/10.1080/13562510600680582

Liu, E. Z.-F., Lin, S. S. J., Chiu, C.-H., & Yuan, S.-M. (2001). Web-based peer review: the learner as both adapter and reviewer. *IEEE Transactions on Education, 44*(3), 246-251. https://doi.org/10.1109/13.940995

Lu, J., & Law, N. (2011). Online peer assessment: effects of cognitive and affective feedback. *Instructional Science, 40*(2), 257-275. https://doi.org/10.1007/s11251-011-9177-2

Lundstrom, K., and W. Baker. 2009. "To give is better than to receive: The benefits of peer review to the reviewer's own writing." *Journal of Second Language Writing 18*(1), 30-43. https://doi.org/10.1016/j.jslw.2008.06.002

Mulqueen C., Baker D. & Dismukes R.K. (2000) *Using multifacet Rasch analysis to examine the effectiveness of rater training*. SIOP.

Mumpuni, K. E., Priyayi, D. F., & Widoretno, S. (2022). How do students perform a peer assessment? *International Journal of Instruction, 15*(3), 751-766. https://doi.org/10.29333/iji.2022.15341a

Nakamura, Y. (2000). Many-facet Rasch based analysis of communicative language testing results. *Journal of Communication Students, 12*, 3-13. https://eric.ed.gov/?id=ED449678

Nakamura, N. (2002). Teacher assessment and peer assessment in practice. *Educational Studies, 44*, 204-215. https://files.eric.ed.gov/fulltext/ED464483.pdf

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199-218. https://doi.org/10.1080/03075070600572090

Orluwene, G. W., & Ekim, D. K. (2020). Promoting self-regulated learning through self-and peer-assessment techniques among secondary school students. *International Journal of Arts and Commerce, 9*(4), 1-16. https://ijac.org.uk/articles/9.4.1.1-16.pdf

Panadero, E., Alqassab, M., Fernández Ruiz, J., & Ocampo, J. C. (2023). A systematic review on peer assessment: intrapersonal and interpersonal factors. *Assessment & Evaluation in Higher Education*, 1-23. https://doi.org/10.1080/02602938.2023.2164884

Pantiwati, Y., & Husamah. (2017). Self and peer assessments in active learning model to increase metacognitive awareness and cognitive abilities. *International Journal of Instruction, 10*(4), 185-202. https://doi.org/10.12973/iji.2017.10411a

Prieto, G., & Nieto, E. (2014). Analysis of rater severity on written expression exam using many faceted Rasch measurement. *Psicologica, 35*(2), 385-397. https://www.redalyc.org/pdf/169/16931314011.pdf

Randall, J. & Engelhard, G. Jr. (2009). Examining teacher grades using Rasch measurement theory. *Journal of Educational Measurement, 46*(1), 1-18. https://doi.org/10.1111/j.1745-3984.2009.01066.x

Sande, J. C. G., & Godino-Llorente, J. I. (2014). Peer assessment and self-assessment: Effective learning tools in higher education. *International Journal of Engineering Education, 30*(3), 711-721. https://oa.upm.es/35804/

Semerci, Ç. (2011a). Doktora yeterlikler çerçevesinde öğretim üyesi, akran ve öz değerlendirmelerin Rasch ölçme modeliyle analizi. *Journal of Measurement and Evaluation in Education and Psychology, 2*(2), 164-171. https://dergipark.org.tr/en/pub/epod/issue/5804/77226

Semerci, Ç. (2011b). Analyzing microteaching applications with many-facet Rasch measurement model. *Education and Science, 36* (161), 14-25. http://egitimvebilim.ted.org.tr/index.php/EB/article/view/145

Semerci, Ç., Semerci, N., & Duman, B. (2013). Analysis of seminar presentation performances of postgraduate students with many-facet Rasch model. *The Journal of Sakarya Unıversity Education Faculty, 0*(25), 7-22. https://dergipark.org.tr/tr/pub/sakaefd/issue/11221/133978

Stefani, L. A. J. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education, 19*(1), 69-75. https://doi.org/10.1080/03075079412331382153

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*(3), 239-261. https://doi.org/10.1016/j.asw.2004.11.001

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative Peer Assessment of Academic Writing Between Postgraduate Students. *Assessment & Evaluation in Higher Education, 25*(2), 149-169. https://doi.org/10.1080/713611428

Tseng, S.-C., & Tsai, C.-C. (2007). On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education, 49*(4), 1161-1174. https://doi.org/10.1016/j.compedu.2006.01.007

Turner, J. (2003). *Examining on art portfolio assessment using a many facet Rasch measurement model* [Unpublished dissertation]. Boston College.

Uyanık, G. K., Güler, N., Teker, G. T., & Demir, S. (2019). The analysis of elementary science education course activities through many-facet Rasch model. *Kastamonu Education Journal, 27*(1), 139-150. https://doi.org/10.24106/kefdergi.2417

Wright, B.D., & Linacre, J.M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370-382. https://www.rasch.org/rmt/rmt83b.htm

Yan, Z., & Brown, G. T. (2017). A cyclical self-assessment process: Towards a model of how students engage in self-assessment. *Assessment & Evaluation in Higher Education, 42*(8), 1247-1262. https://doi.org/10.1080/02602938.2016.1260091

Yan, Z., Wang, X., Boud, D., & Lao, H. (2023). The effect of self-assessment on academic performance and the role of explicitness: a meta-analysis. *Assessment & Evaluation in Higher Education, 48*(1), 1-15. https://doi.org/10.1080/02602938.2021.2012644

Yüzüak, A. V., Yüzüak, B., & Kaptan, F. (2015). A many-facet Rasch measurement approach to analyze peer and teacher assessment for authentic assessment task. *Journal of Measurement and Evaluation in Education and Psychology, 6*(1), 1-11. https://doi.org/10.21031/epod.57425

Zhan, Y., Yan, Z., Wan, Z. H., Wang, X., Zeng, Y., Yang, M., & Yang, L. (2023). Effects of online peer assessment on higher-order thinking: A meta-analysis. *British Journal of Educational Technology, 54*, 817-835. https://doi.org/10.1111/bjet.13310