Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309-6575 *Kış* 2023
Winter 2023
Cilt: 14-Sayı: 4
Volume: 14-Issue: 4

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 sorumluluğu 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

Hakem Kurulu / Referee Board

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Aralık 2023, Sayı: 14-4)

Journal of Measurement and Evaluation in Education and Psychology (December 2023, Issue: 14-4)

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(4); 396-412

Comparison of Different Computerized Adaptive Testing Approaches with Shadow Test Under Different Test Lengths and Ability Estimation Method Conditions

Mahmut Sami YİĞİTER*   Nuri DOĞAN**

Abstract

Adaptive testing approaches have been used and adopted in many international large-scale assessments (PISA, TIMSS, PIRLS, etc.). The shadow test approach, on the other hand, is an innovative testing approach that both meets all test specifications and constraints and aims to provide maximum information at the test taker's true ability level. The aim of this study is to investigate the effectiveness of four different adaptive testing approaches created with shadow test (CAT, 2-Stage O-MST, 3-Stage O-MST, and LOFT) according to the test length and ability estimation method. With the Monte Carlo (MC) study in R software, 200 item parameters and 2000 test takers were generated under the 3PL model and the results were calculated over 50 replications. The results show that CAT, 2-Stage O-MST, and 3-Stage O-MST are quite similar in effectiveness, while LOFT is less effective than these techniques. As the test length increases, the measurement precision increases in all different types of adaptive tests. Although the EAP method generally presents better measurement precision than the MLE method, at the extremes of the ability scale, MLE has been found to present good measurement precision. In the research, it is discussed that large-scale assessments can benefit from adaptive testing created with a shadow test approach.

Keywords: computerized adaptive testing, shadow test, on-the-fly multistage testing, linear on-the-fly test

Introduction

Linear tests have been the most popular way of measuring knowledge, skill, and ability in the field of education for centuries. With the advancements in computer hardware and software, Computer Adaptive Testing (CAT) has been adopted and used in many applications worldwide, including the Graduate Record Examination (GRE), Graduate Management Admission Test (GMAT), and Medical College Admission Test (MCAT), as it provides efficient ability estimation and shortens test time (Kirsch & Lennon, 2017; Gökçe & Glas, 2018; Khorramdel et al., 2020; Akhtar et al., 2023; Ebenbeck, 2023).

In linear tests (LT), test takers take all items. A large number of items are needed to obtain effective ability estimation from linear tests (Huang et al., 2009). In CAT, on the other hand, individual tests are obtained by analyzing the properties of the items by algorithms (Raborn & Sari, 2021). CAT is a computer-based test; it can have a fixed or varying length. The test management algorithm presents items to the test taker consecutively and adjusts the difficulty of the items to estimate the test taker's ability level as the test progresses (Wainer, 1990; Hendrickson, 2007; Choi & van der Linden, 2018; Gündeğer & Doğan, 2018). In Computerized Multistage Testing (MST), a group of items called "module" is administered to test takers. In MST, the difficulty of the test is adjusted between modules according to the answers given by the test taker (Yigiter & Dogan, 2023). Although there are studies showing that CAT is more effective than MST and LT in terms of measurement precision (Patsula, 1999; Schnipke & Reese, 1999), MST has beneficial aspects for test administration and test takers (Kim & Plake, 1993; van der Linden, 2010).

* Dr., Social Sciences University of Ankara, Distance Education Application and Research Center, Ankara-Türkiye, e-mail: mahmutsamiyigiter@gmail.com, ORCID ID: 0000-0002-2896-0201
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Türkiye, e-mail: nuridogan2004@gmail.com, ORCID ID: 0000-0001-6274-2016

To cite this article:
Yiğiter, M. S., & Doğan, N. (2023). Comparison of different computerized adaptive testing approaches with shadow test under different test lengths and ability estimation method conditions. *Journal of Measurement and Evaluation in Education and Psychology, 14*(4), 396-412. https://doi.org/10.21031/epod.1202599

Received: 11.11.2022
Accepted: 20.10.2023

It is important to choose a good model with parameters that reflect the characteristics of the items and the abilities of the test takers in order to maintain the adaptive test successfully. Item Response Theory (IRT) has been successfully operating in its nearly century-old historical development. Many models have been developed under IRT models. In addition to the statistical model, the creation of a well-designed item pool for the attribute to be measured plays an important role in the success of the adaptive test. The selection of optimal items at each ability level, updated with a well-designed item pool and a successful statistical model, can be defined as a mathematical optimization problem. In current adaptive testing algorithms, many methods have been developed as item selection methods, including the maximum Fisher information (MFI), the maximum likelihood weighted information (MLWI) (Veerkamp & Berger, 1997), the maximum posterior weighted information (MPWI) (van der Linden, 1998), the maximum expected information (MEI) (van der Linden, 1998), the minimum expected posterior variance (MEPV), the Kullback-Leibler (KL) divergency criterion (Chang & Ying, 1999), the posterior Kullback-Leibler (KLP) criterion (Chang and Ying, 1996), the global-discrimination index (GDI) (Kaplan et al., 2015). MFI criterion can work successfully in a simple adaptive test where only the item selection from the item pool is based on the amount of information criterion. However, as the number of test specifications and constraints increases in item selection from the item pool, the number of combinations increases rapidly in item selection. Therefore, the complexity of the solution gap in the MFI criterion can make things unsolvable. When the test specifications and constraints such as different contents, item types, word count, expected response time, common stem items, enemy items (items where one item indicates the solution of another item due to the similarity of their content), word count, answer key distribution are considered together, the combinatorial complexity of the problem increases rapidly with each additional constraint. Under such a set of constraints, it is necessary to check each of the possible solutions until the information criterion has the largest value. Also, considering that the test termination rule will end with an incomplete test, there is no guarantee that the test taker will be presented with a test that meets all the constraints (van der Linden, 2022).

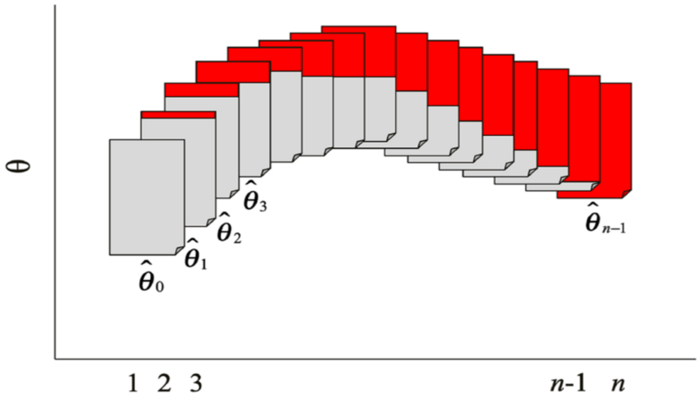
The basis of this mathematical optimization problem in adaptive tests is discrete optimization, which requires items to be found in order. Instead of discrete optimization that selects items one by one in each ability estimation, test forms that meet all test specifications and constraints can be created with a mixed integer programming (MIP) methodology that combines a fixed test form. In adaptive testing, Shadow Test Approach has come to the fore with the idea of offering test-takers a test that meets all test specifications and constraints (van der Linden & Chang, 2003).

CAT with Shadow Test Approach

The idea behind the shadow test approach is to create a test that both meets all test constraints and offers maximum information at the test taker's true ability level. The shadow test is obtained by assembling a full-length test that satisfies all the constraints set by the algorithm. When a shadow test is created, the item to be administered to the test taker is the item with the most information. In addition, each shadow test is assembled in such a way that the information function at the interim ability level has the maximum value, and the item to be selected from the shadow test and applied to the test taker has the maximum contribution to this function (van der Linden, 2009). Each subsequent shadow test also includes all items that have already been implemented by the test taker. Therefore, the final shadow test is true adaptive testing and always satisfies all constraints. The basic structure of the shadow test approach is shown in Figure 1.

Figure 1.

Basic Structure of the Shadow Test (van der Linden, 2022)



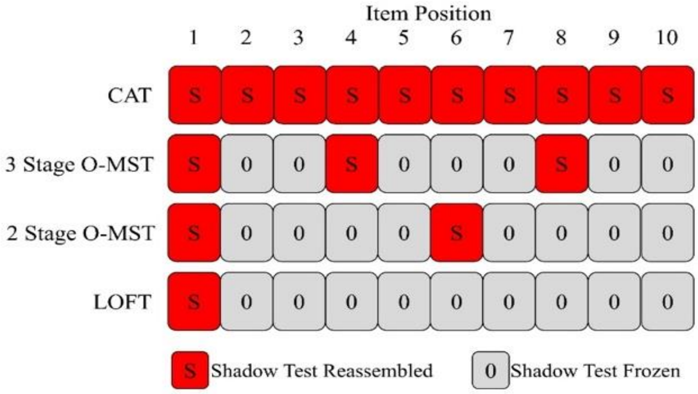
In Figure 1, the horizontal axis of the graph shows the position of the items in the test; the vertical axis represents the ability level that is updated after the implementation of each item. The higher the vertical position of the shadow tests, the higher the current ability estimation. Towards the end, the convergence of the positions of the shadow tests represents that the final ability estimation has become stable. The red part of the shadow test represents the items answered by the test taker. The gray part represents the part of the ability estimation that is reassembled after a new update (taking into account the items in the red part). The final shadow test includes all of the items actually taken by the test taker (van der Linden, 2009). Different adaptive testing approaches emerged by assembling the freeze-refresh mechanism introduced by van der Linden and Diao (2014) and shadow test at different item locations.

Freeze-Refresh Mechanism and Different Adaptive Testing Approaches

The original shadow test approach is based on reassembling the shadow test at every θ ability update. However, it is stated that instead of a new test assembly after each item, test assembly can be performed at predetermined item locations of the test. With this freeze-refresh mechanism, which was first introduced by van der Linden and Diao (2014), different adaptive testing approaches can be obtained by adjusting the test adaptation points to different item positions. Some adaptive testing approaches that can be obtained by changing the adaptation points according to the locations of the items are shown in Figure 2.

Figure 2.

Different Adaptive Testing Approaches with Freeze-Refresh Mechanism



	Item Position									
	1	2	3	4	5	6	7	8	9	10
CAT	S	S	S	S	S	S	S	S	S	S
3 Stage O-MST	S	0	0	S	0	0	0	S	0	0
2 Stage O-MST	S	0	0	0	0	S	0	0	0	0
LOFT	S	0	0	0	0	0	0	0	0	0

S Shadow Test Reassembled   0 Shadow Test Frozen

Figure 2 shows that different adaptive testing approaches can be obtained by reassembling the shadow test at different adaptation points on a test consisting of 10 items. Here, the letter "S" in the red box means that the shadow test was reassembled at those item positions and can be referred to as the "adaptation point". In the “0” item positions shown in the gray box, it indicates that the test is frozen (the shadow test is not reassembled) (Choi & van der Linden, 2018). These test approaches, hybrid adaptive tests can be created with the freeze-refresh mechanism. The four different approaches created by the freeze-refresh mechanism and discussed in this study are as follows;

- CAT is a fully adaptive test in which the shadow test is reassembled at each item position;
- 3-Stage O-MST (3 Stage On-The-Fly Multistage Testing) is a three-stage adaptive test in which the shadow test is reassembled at three specified item positions (item positions 1, 4, and 8);
- 2-Stage O-MST (2 Stage On-The-Fly Multistage Testing) is a two-stage adaptive test in which the shadow test is reassembled at two specified item positions (item positions 1 and 6);
- LOFT (Linear On-the-Fly Testing) is a uniquely created fixed test that is brought together at the test taker's initial ability level (Choi & van der Linden, 2018).

Ability Estimation

Many different methods have been developed in the estimation of a test taker's ability in IRT. Frequently used ability estimation methods can be listed as Maximum Likelihood Estimation (MLE) (Birnbaum, 1968), Weighted Likelihood Estimation (WLE) (Warm, 1989), Marginal Maximum Likelihood Estimation (MMLE) (Bock & Aitkin, 1981), Expected a Posteriori (EAP) (Bock & Aitkin, 1981) and Maximum a Posteriori (MAP) (Samejima, 1977; Embretson & Reise, 2000).

The MLE method efforts to get the θ value that maximizes the likelihood function. In cases where the test taker's responses to the items are independent of each other, the product of the response probabilities is defined as the likelihood function, and the point at which this function reaches its maximum is estimated as the ability level. The likelihood function is shown in Equation 1.

$$L(u|\theta) = \prod_{i=1}^{n} P_i(u_i|\theta)^{u_i} * Q_i(u_i|\theta)^{(1-u_i)} \tag{1}$$

In Equation 1, *u* represents the response vector. $P_i(u_i|\theta)$ indicates the probability that the test taker will correctly answer item i at ability level θ. $Q_i(u_i|\theta)$ is equal to 1-$P_i(u_i|\theta)$. n represents the number of items. The value at which this likelihood function is maximum is estimated as the test taker's ability level (θ). Ability level (θ) is solved by iterative methods by taking the derivative of the likelihood function given in the above equation. The most common method used for this purpose is the Newton-Raphson method (Wang & Vispoel, 1998).

The EAP method, on the other hand, is one of the Bayesian ability estimation methods that utilize a priori distributions in ability estimation. Its general formula is shown in Equation 2 (Borgatto et al., 2015):

$$EAP\ (\vartheta) = E(\vartheta|u) = \frac{\int_R \vartheta * L(\vartheta|u) * f(\vartheta) d\vartheta}{\int_R L(\vartheta|u) * f(\vartheta) d\vartheta} \tag{2}$$

In the equation, $f(\theta)$ is the prior distribution function and $L(\theta|u)$ is the likelihood function. In the EAP method, the prior distribution of ability levels must be known. If the a priori distribution is incorrect, the EAP may estimate ability parameters incorrectly (Embretson & Reise, 2000). While the MLE method is an iterative method, the EAP method is not. Therefore, ability estimations can be obtained faster with EAP.

Choi, Moellering, Li, and van der Linden (2016) compared CAT, O-MST, and a hybrid method combining these two methods using the freeze-refresh mechanism. In this study, the authors generated the parameters of the item pool consisting of 1000 items, assuming that they reflect a real case. The results are quite similar for all three approaches. The researchers concluded that the freeze-fresh mechanism works quite successfully, especially when there are common stem items and test constraints that need to be met, and there is no significant decrease in the measurement accuracy of the test.

Zheng and Chang (2015) compared CAT, O-MST, and F-MST in terms of measurement accuracy. In the study, a real item pool of 352 items from a large-scale assessment was used. The results of this study indicate that CAT and O-MST offer very similar measurement precision and that these two methods offer better measurement precision than F-MST.

van der Linden and Diao (2014) compared five different testing approaches, namely CAT, hybrid CAT, O-MST, F-MST, and LT, with real data sets by simulation. The results of this study show that LT is the least efficient, followed by F-MST. The other approaches, namely CAT, hybrid CAT, and O-MST, are reported to be approximately equally efficient.

Han and Guo (2014) propose an O-MST design that does not include pre-combined test modules and combines a new module on the fly at each stage. The researchers compared the CAT, F-MST, and O-MST designs. The 1-3-3-MST design produced similar measurement accuracy results with the newly developed O-MST design at low iteration shaping, while the new O-MST design produced better measurement accuracy results than F-MST when the number of iterations increased to 100. CAT produces better measurement accuracy than both methods.

Choi and van der Linden (2018) compared the O-MST, which they created using the shadow test approach, with the MST and fixed linear test designs. They concluded that although the measurement accuracy obtained from the O-MST is slightly lower than the CAT, it is better than the fixed linear test.

Purpose of the Study and Research Questions

The purpose of this study is to examine the effectiveness of different adaptive testing approaches created with shadow test according to test length and ability estimation methods. Since the test lengths and ability estimation methods that will be discussed in this study have not been included in any previous study; it is thought that this study will contribute to the development of new approaches. For this purpose, the main research question is:

How does the measurement precision of different adaptive testing approaches change according to different test lengths and different ability estimation methods?

According to the main purpose of the study, the three sub-research questions examined in order to examine this research question in detail are as follows:

1. How does the measurement precision change if different adaptive testing approaches (CAT, 3-Stage O-MST, 2-Stage O-MST, LOFT) are used in the adaptive testing approach?

2. How does the measurement precision change if the different fixed-test lengths (20, 30, 40) are used in the adaptive testing approach?

3. How does the precision of measurement change if different ability estimation methods (MLE, EAP) are used in different adaptive testing approaches?

Method

In this study, Monte Carlo (MC) simulations were performed to compare different adaptive testing approaches (Harwell et al., 1996). MC is a simulation method used to analyze the behavior of statistical models. In this method, the computer generates data according to probabilistic distributions and allows a comparison of the outputs from the model(s) (Sigal & Chalmers, 2016). In the research, a simulation study was carried out by changing the conditions of four different adaptive testing approaches (CAT, 2-Stage O-MST, 3-Stage O-MST, and LOFT), three different test lengths (20, 30, and 40) and two different ability estimation methods (EAP and MLE). All conditions are crossed with each other. Therefore, in this study, 4x3x2=24 conditions were examined. Analyzes were performed by making 50 replications for each condition.

Data Generation

The R program was used to generate the data and the "TestDesign" package in R was used for the analysis (Choi et al., 2022). Within the scope of the research, the parameters of 200 items were generated based on the 3PL model considering the distributions suggested in the literature (Feinberg & Rubright, 2016; Mooney, 1997; Bulut & Sünbül, 2017). Item discrimination parameters were obtained from $a \sim lnN$ (0.2, 0.3) log-normal distribution, item difficulty parameters were obtained from $b \sim N$ (0, 1) normal distribution, and item prediction parameters were obtained from $c \sim Beta$ (5, 16) beta distribution. Ability parameters were produced from the normal distribution $b \sim N$ (0, 1), with 2000 test takers. In addition, assuming that the item pool of the test will consist of three different contents, the item pool is randomly divided into three different content: Content 1 40 items (20%), Content 2 100 items (50%), and Content 3 60 items (30%). Descriptive statistics on item parameters and test taker parameters are shown in Table 1.

Table 1

Descriptive Statistics of Item and Ability Parameters

Parameter	N	Mean	Sd	Min	Max
a	200	1.36	0.24	0.87	1.93
b	200	-0.06	1.07	-2.86	2.81
c	200	0.24	0.08	0.07	0.49
Theta	2000	0.00	1.00	-3.11	2.96

Simulation Conditions

There are conditions that are varied in different adaptive tests created with the shadow test approach. The details of these conditions are explained in the sub-headings below.

Starting Rule

In adaptive testing, the test taker's starting level must be determined before the test can be started. If some information about the test takers is available, it can be used as a starting rule. This information can be students' information (previous course scores, graduation scores, student point average, etc.) or the average of the population (Wang & Vispoel, 1998; Stafford et al., 2019). Since this research was conducted with the simulation and since the population was produced from a normal distribution, the initial ability level was determined as θ=0 for all participants.

Item Selection

Many item selection methods have been developed in adaptive testing approaches with shadow test, especially Maximum Fisher Information (MFI), Maximum Posterior Weighted Information (MPWI), Goal Fisher Information (GFI), Full Bayesian (FB), Empirical Bayes (EB). In this study, the MFI method, which maximizes the amount of information at the interim ability level, was used (Choi et al., 2022).

Content Balancing

In this study, it was assumed that the item pool consisted of three different contents. Content 1, Content 2, and Content 3 are comprised of 40, 100, and 60 items, respectively (20%, 50%, and 30%, respectively). The contents of the items were determined randomly. The number of items obtained from the contents according to the test lengths is presented in Table 1. In all different adaptive testing approaches, the content distributions given in Table 2 are limited. The number of items that will come from the contents is determined according to their percentages. In Table 2, the content distributions of the adaptive tests according to the test lengths are given.

Table 2

Distribution of the Number of Items to be Obtained from the Contents

Content	Test Length		
	20	30	40
Content 1 (%20)	4	6	8
Content 2 (%50)	10	15	20
Content 3 (%30)	6	9	12

Automated Test Assembly Method

There are many methods used for automated test assembly (glpk, lpSolve, lpsymphony, gurobi, etc.). In this study, the “glpk” (Theussl et al., 2019) method was used for automated test assembly.

Freeze-Refresh Mechanism Item Positions

In computerized adaptive tests created with shadow tests, the item locations where the shadow tests are reassembled and re-presented to the participant student in the freeze-refresh mechanism should be determined. The item locations where the shadow tests are reassembled are given in Table 3.

Table 3

Item Positions in which Shadow Tests Reassembled

Adaptive Test	Test Length		
Approach	20	30	40
CAT	All item position	All item position	All item position
3-Stage O-MST	1., 8. and 14.	1., 11. and 21.	1, 14 and 28.
2-Stage O-MST	1. and 11.	1. and 16.	1. and 21.
LOFT	Only 1.	Only 1.	Only 1.

Ability Estimation Method

In computerized adaptive testing, the method of ability estimation should also be determined. In this study, EAP and MLE methods were used for ability estimation.

Termination Rule

Fixed test length, minimum or maximum test length limit, and standard error threshold methods can be used as termination rules in adaptive test applications with shadow test (Choi et al., 2022). Since the LT and O-MST methods were fixed-length tests in this study, "fixed test length" (20, 30, or 40 depending on the condition) was used for all different adaptive testing as the termination rule.

Data Analysis

In this study, 50 replications were performed for each of the 24 different conditions. The relationships between the true and estimated ability parameters obtained from different adaptive test designs for each condition were interpreted by calculating the Pearson Correlation coefficient, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) values. The formulas for correlation, RMSE, and MAE values are given below.

$$Correlation = \frac{\sum_{i=1}^{n}(\widehat{\theta_\iota} - \overline{\widehat{\theta_\iota}})(\theta_i - \overline{\theta_\iota})}{(n-1)S_{\widehat{\theta_\iota}}S_{\theta_i}} \tag{3}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{\theta_\iota} - \theta_i)^2}{n}} \tag{4}$$

$$MAE = \frac{\sum_{i=1}^{n}|\widehat{\theta_\iota} - \theta_i|}{n} \tag{5}$$

The θ_i in the formulas represents the true ability level of the participants, and $(\widehat{\theta_\iota})$ the estimated ability level. $(\overline{\theta_\iota})$ and $(\overline{\widehat{\theta_\iota}}\ )$ denote the mean of the true and estimated ability levels respectively, S_{θ_i} and $S_{\widehat{\theta_\iota}}$ denote the standard deviation of the true and estimated ability levels respectively, and n denotes the sample size.

Codes were written by the researchers in the R to calculate correlation, RMSE, and MAE values according to the conditions. In addition, RMSE and MAE graphs were created to compare the effectiveness of different adaptive testing approaches according to their ability ranges.

Results

In this section, the findings of the research are given. First of all, the results of the research were examined in a general framework according to all conditions; then the results obtained for each sub-problem of the research were presented under sub-headings. The correlation, RMSE, and MAE values calculated for each of the 24 simulation conditions examined in the study are shown in Table 4.

Table 4

Overall Results from All Conditions

Ability Estimation Method	Test Length	Adaptive Test Type	Correlation	RMSE	MAE
EAP	20	CAT	0.938	0.350	0.275
		3-Stage O-MST	0.935	0.358	0.283
		2-Stage O-MST	0.931	0.367	0.289
		LOFT	0.913	0.410	0.322
	30	CAT	0.956	0.297	0.234
		3-Stage O-MST	0.953	0.305	0.241
		2-Stage O-MST	0.951	0.313	0.246
		LOFT	0.935	0.357	0.277
	40	CAT	0.965	0.266	0.211
		3-Stage O-MST	0.962	0.274	0.216
		2-Stage O-MST	0.961	0.280	0.221
		LOFT	0.947	0.325	0.251
MLE	20	CAT	0.937	0.381	0.297
		3-Stage O-MST	0.932	0.398	0.309
		2-Stage O-MST	0.927	0.409	0.317
		LOFT	0.904	0.450	0.346
	30	CAT	0.955	0.318	0.25
		3-Stage O-MST	0.950	0.336	0.261
		2-Stage O-MST	0.947	0.346	0.268
		LOFT	0.928	0.395	0.301
	40	CAT	0.963	0.283	0.223
		3-Stage O-MST	0.961	0.295	0.229
		2-Stage O-MST	0.958	0.309	0.238
		LOFT	0.940	0.363	0.271

Note. CAT = Computerized Adaptive Testing, O-MST = On-the-fly Computerized Multistage Testing, LOFT = Linear On-The-Fly Test, EAP = Expected a Posteriori, MLE = Maximum Likelihood Estimation.

When Table 4 is examined, it is seen that EAP, one of the ability estimation methods, presents good measurement precision (high correlation and low RMSE-MAE) compared to MLE in all conditions. In all conditions, the measurement precision increases as the test length increases.

CAT provides the best measurement precision in all conditions. While CAT is followed by 3-Stage O-MST and 2-Stage O-MST, respectively, LOFT is seen to be in the last sequence in all conditions. In addition, it can be said that while CAT, 3-Stage O-MST, and 2-Stage O-MST have very similar measurement precision, measurement precision is significantly less because LOFT does not have an adaptation point.

In this section, the correlation, RMSE, and MAE values obtained by averaging from Table 4 and the answers to the three sub-research questions mentioned above were sought. In addition, RMSE and MAE graphs were drawn and interpreted according to their ability ranges.

Results on the First Problem

With regard to the first sub-research question, it was examined how measurement precision changed in different adaptive testing approaches with the shadow test. Findings related to this sub-research question are presented in Table 5.

Table 5

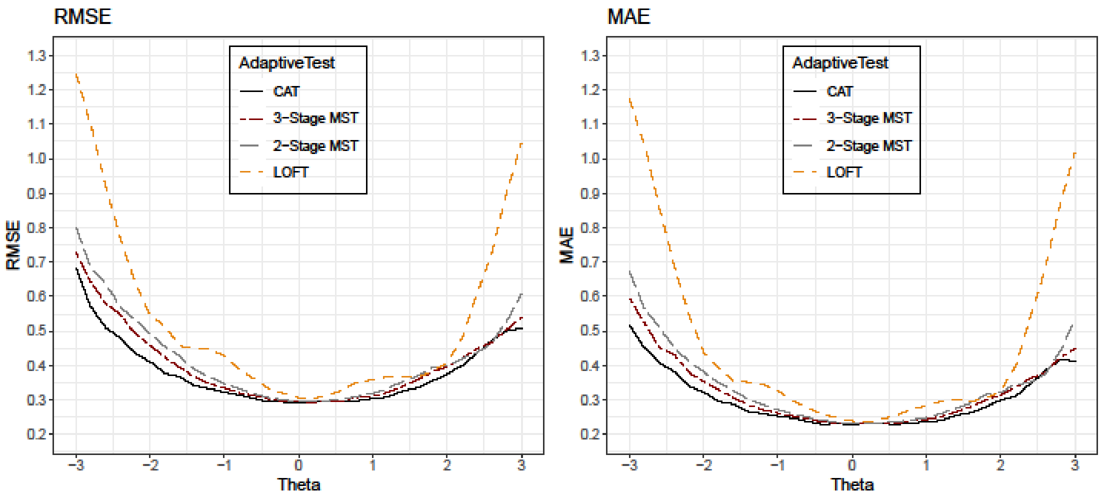
Results by Adaptive Testing Approach Condition

Adaptive Test Type	Correlation	RMSE	MAE
CAT	0.952	0.316	0.248
3-Stage O-MST	0.949	0.328	0.257
2-Stage O-MST	0.946	0.337	0.263
LOFT	0.928	0.383	0.295

As seen in Table 5, CAT shows better measurement precision (high correlation and low RMSE - MAE) than other adaptive tests. In terms of measurement precision, 3-Stage O-MST, 2-Stage O-MST, and LOFT come after CAT. In adaptive tests with the shadow test, it can be said that the measurement precision increases as the adaptation point increases. Figure 3 presents the RMSE and MAE values on the ability scale of different adaptive tests.

Figure 3

Findings on Measurement Precision According to Different Adaptive Testing Approaches



As seen in Figure 3, CAT presents better measurement precision than other adaptive testing approaches across the all ability scale in terms of both RMSE and MAE values. The 3 Stage3-Stage O-MST and 2-Stage O-MST approaches also appear to offer slightly worse but still good measurement precision than CAT. Although LOFT achieves almost as good measurement precision as other adaptive testing approaches around $\theta = 0$ ability level, its measurement precision decreases considerably towards extreme ability levels. Due to LOFT's test assembling at $\theta = 0$ ability level and the absence of an adaptation point, it is seen that the measurement precision of RMSE and MAE values at extreme ability levels decreases significantly compared to other adaptive testing approaches. Although there are not many individuals at extreme ability levels compared to the normal distribution, it is thought that the measurement precision of LOFT will decrease considerably in skewed or uniform ability distributions.

Results on the Second Problem

With the second sub-research question, it was examined how the measurement precision changed according to the test length of the different adaptive tests with shadow test. Findings related to the second sub-research question are presented in Table 6.

Table 6

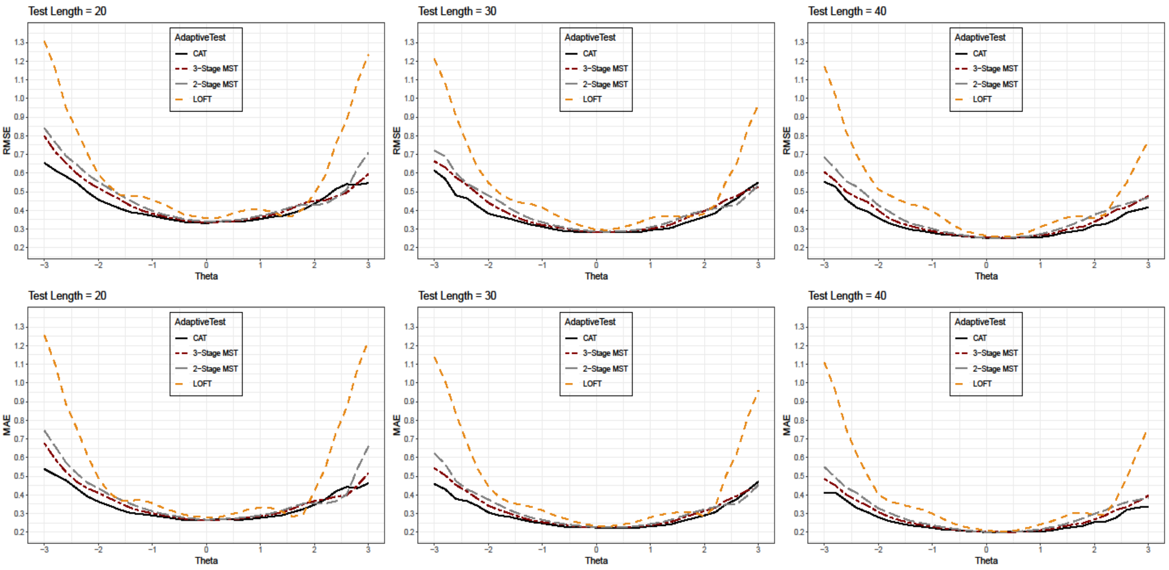
Results by Test Length

N	Correlation	RMSE	MAE
20	0.927	0.390	0.305
30	0.947	0.333	0.260
40	0.957	0.299	0.233

According to the correlation, RMSE, and MAE values, it is seen that the measurement precision increases as the test length increases. The difference in measurement accuracy between 20 and 30 test lengths (ΔCor = 0.020, ΔRMSE = 0.057 and ΔMAE = 0.045) is large, while the measurement precision between 30 and 40 test lengths (ΔCor = 0.010, ΔRMSE = 0.034 and ΔMAE = 0.026) less. This indicates that the measurement precision of 30 to 40 test lengths is more similar than that of 20 test lengths. Figure 4 presents the RMSE and MAE values on the ability scale of different test lengths.

Figure 4

Findings Concerning the Measurement Precision of Test Length by Ability Scale



As seen in Figure 4, both RMSE and MAE values decrease as the test length increases. In addition, as the test length increases, it is seen that the measurement precision at the extreme ability levels decreases more than at the middle ability levels.

Results on the Third Problem

With regard to the third sub-research question, it was examined how the measurement precision changed according to the ability estimation method of the different adaptive testing approaches with shadow test. Findings related to the third sub-research question are presented in Table 7.

Table 7

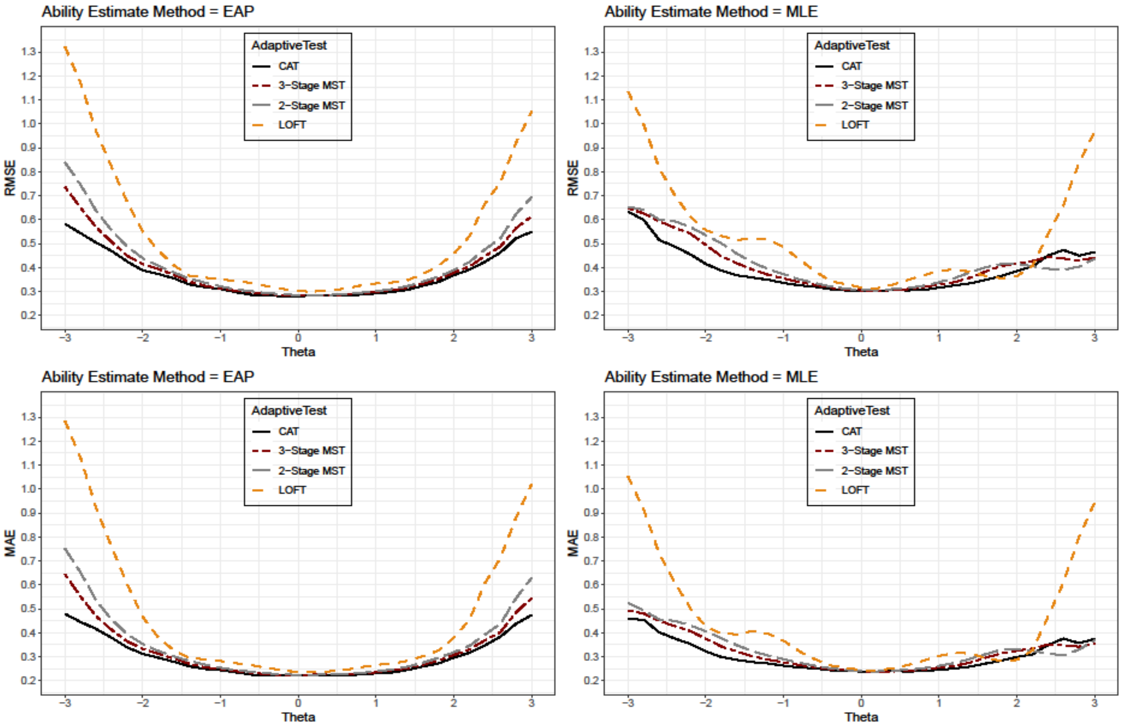
Findings by Ability Estimation Method

Ability Estimate Method	Correlation	RMSE	MAE
EAP	0.946	0.325	0.256
MLE	0.942	0.357	0.276

As seen in Table 7, the EAP method presents better measurement precision than the MLE method according to the correlation, RMSE, and MAE values. In Figure 5, RMSE and MAE values on the ability scale of different ability estimation methods are presented.

Figure 5

Measurement Precision Findings According to Different Ability Estimation Methods in Ability Scale



As seen in Figure 5, the EAP ability estimation method presents better measurement precision than the MLE method in the middle part of the ability scale. On the other hand, it can be said that MLE provides better measurement precision at the extremes of the ability scale. In addition, in terms of both RMSE and MAE, the graphs of the different adaptive testing approaches of the EAP method have a more uniform shape in the ability scale, while the graphs of the MLE method show a more fluctuating increase and decrease.

Discussion

Under the shadow test approach, different adaptive tests can be created with the freeze-refresh mechanism. With this freeze-refresh mechanism, which was first introduced by van der Linden and Diao (2014), adaptive tests such as hybrid-CAT, O-MST, and LOFT can be created since the adaptation points of the test can be adjusted. There are many studies in the literature that shadow tests work successfully under test specifications and constraints that make it difficult for the test algorithm to overcome (van der Linden & Veldkamp, 2004; Choi & Lim, 2022).

This study aims to compare four different adaptive testing approaches created with shadow test according to test length and ability estimation. When the different adaptive test approaches with shadow test are examined in terms of measurement precision, it has been concluded that CAT offers the best measurement precision. It can be stated that 3-Stage O-MST and 2-Stage O-MST follow the CAT, respectively, while LOFT perform worse than other methods in the aspect of measurement precision. Although CAT presents better measurement precision than 3-Stage O-MST and 2-Stage O-MST, it can be said that the RMSE and BIAS values of these three different adaptive testing approaches are quite similar. LOFT, on the other hand, produced worse results than these three approaches because there was no adaptation point.

Choi and van der Linden (2018), in their study on patient-reported outcomes (PRO) measurement, report that CAT offers better measurement accuracy than 3-Stage O-MST and LOFT, similar to the results of this study. In addition, this study states that CAT and 3-Stage O-MST produce very close results. Van der Linden and Diao (2014), in their study comparing different adaptive tests, reported that CAT and O-MST offer very close measurement precision, while fixed-LT offers worse measurement precision than these two adaptive tests. Similarly, Zheng and Chang (2015), in their study comparing CAT, O-MST, and fixed-MST, state that CAT and O-MST offer very similar measurement precision. Comparing the measurement precision with respect to different points of the ability scale, CAT offers very similarly good measurement precision on the 2-Stage O-MST and 3-Stage O-MST ability scales.

Choi et al. (2016), similar to these findings, in shadow tests, it is stated that reassembling the test at each item position and reassembling it at certain item positions will yield nearly equivalent results. On the other hand, LOFT offers good measurement precision in the middle part of the ability scale, similar to other adaptive tests, while measurement precision decreases sharply at extreme ability levels. The reason for the lower measurement precision of the LOFT is the absence of an adaptation point. The results of Han and Guo (2014), van der Linden and Diao (2014), and Choi and van der Linden (2018) are similar to this finding of the study.

In this study, it was concluded that measurement precision increased as the test length increased in different adaptive test types. There are many studies in the literature that adaptive test length increases measurement precision (Weiss, 2004; Özdemir & Gelbal, 2022; Erdem-Kara & Dogan, 2022). Choi and Linden (2018), in their study comparing different adaptive tests with shadow test, states that the 12-item test length offers better measurement precision than 6 items. Xiao and Bulut (2022), in their study examining O-MST, stated that similar to the findings of this study, 60-item length offers better measurement precision than 30 items. The two most important arguments of Computerized Adaptive testing are to reduce test length and increase measurement precision. Therefore, the test length should be short. At the same time, increasing the test length after a certain length will not improve the measurement precision at the desired level due to the "law of diminishing efficiency". In addition, due to the fatigue of the test taker, it may not reflect the real performance of the test taker. In this study, it was concluded that the efficiency of the measurement, which occurs when the test length is increased from 20 to 30, does not occur when it is increased from 30 to 40. It is thought that more research is needed to determine the optimal test length.

It has been concluded that the EAP ability estimation method presents better measurement precision than the MLE method in different adaptive tests with shadow tests. Sahin and Boztunc-Ozturk (2020), in their study on MST, it is seen that EAP performs better than MLE. Similarly, Han (2016) states that EAP performs better than MLE. When the results according to the ability scale are examined, it is seen in the graphs that while the EAP method presents very good measurement precision in the middle area of the ability scale compared to the MLE in all different adaptive tests, the MLE method presents better measurement precision than EAP in extreme ability scale. The findings of Han (2016) and Şahin and Boztunç-Öztürk (2020) support this conclusion. On the other hand, it can be stated that while the graphs of the ability estimation made with EAP follow a regular path across the all ability scale, the graphs of the ability estimation made with MLE follow a fluctuating path.

Computerized Adaptive Testing and its derivatives have been adopted and used in many large-scale applications over the years. MST has an increasing usage area, especially in recent years. In the PISA (Programme for International Student Assessment) administration implemented in 2018, one MST design was used only in the reading area, out of 3 main areas (Khorramdel et al., 2020). In PISA 2022, it is stated that MST design will be used in more than one area (NCES, 2019). In PIRLS, on the other hand, an MST design consisting of grouped items in 2021 was used (Mullis and Martin, 2019). TIMSS, on the other hand, is prepared to use MST design in both mathematics and science in the 2023 administration (Lin & Foy, 2021). The MST method used in these large-scale assessments is applications where modules and panels consisting of item groups are assembled before the exam administration. Test implementations where modules and panels are assembled before the test application are also called Fixed-MST (F-MST). In F-MST, the adaptation point is low as the test is assembled only according to certain ability levels. Therefore, there are many findings in the literature that the measurement precision of MST is lower than that of CAT (Patsula, 1999; Macken-Ruiz, 2008; Wang, 2017). O-MST, on the other hand, can be considered as a new approach to assembling the advantages of CAT and MST (Zheng and Chang, 2015). As seen in this study, the O-MST approach presents very similar measurement precision to CAT, although the number of adaptations is less. In addition, O-MST has advantages such as presenting items in groups, including items with common stems and passing, skipping, and returning between items, similar to F-MST. In the second and later stages of F-MST, certain ability levels are determined as adaptation points (for example, -1, 0, 1). In O-MST, just like in CAT, every point of the ability level is an adaptation point. This feature of O-MST provides an advantage over F-MST in terms of measurement precision (Han, 2016; van der Linden & Diao, 2014). Given the stated O-MST's advantages, O-MST is promising for international large-scale assessments.

LOFT, on the other hand, presents very similar measurement precision to both CAT and O-MST in the middle area of the ability scale. At extreme ability levels, it differs sharply from both CAT and O-MST by offering rather poor measurement precision. LOFT can be used for diagnostic assessments or to make pass-fail decisions for students with a cut-off point at the midpoint of the ability scale. However, it can be said that its use in exams with cut-off points at extreme ability levels or high-stakes exams will have disadvantages compared to other adaptive tests. In addition, LOFT creates unique linear test forms for each test taker. Therefore, since LOFT does not have any adaptation points, a computer application may not be required. The test forms created with LOFT are applied to the students even in the classroom environment and can be scored after the implementation.

Finally, we offer some practical recommendations. It can be pointed out that a 2 or 3-stage O-MST can be used instead of CAT with some compromise in measurement accuracy. In this way, the advantages of MST can also be utilized. If the scores to be obtained from the test are to be assessed with a cut-off score at extreme ability levels, CAT should be preferred. EAP method can be preferred instead of MAP as an ability estimation method. In terms of test length, shorter tests had lower measurement accuracy, while increasing test length did not linearly increase measurement accuracy. Therefore, it is important to determine the optimal test length in adaptive tests by considering the purpose, content, and measurement accuracy of the test together.

Limitations and Future Studies

This research has some limitations. These limitations can guide researchers in future research. In this study, the fixed test length rule was used as the termination rule. Studies that examine different termination rules can be designed. On the other hand, the MFI method was used as the item selection method in all conditions. The study can be reconsidered with different item selection methods. CAT, O-MST, and LOFT are considered Adaptive Testing Approaches. Hybrid-CAT approaches created by mixing CAT and O-MST can be considered (for more information, see. Choi & van der Linden, 2018). In this study, the item pool was generated by simulation. Working with real item pools is reproducible.

The ability distribution was obtained from the normal distribution. Results can be examined under different or skewed distributions. As an ability estimation method, EAP and MLE were compared. Comparisons can be made with different ability estimation methods such as MAP and MLEF.

Declarations

Author Contribution: Mahmut Sami YİĞİTER: conceptualization, investigation, methodology, data analysis, visualization, writing - review & editing. Nuri DOĞAN: conceptualization, methodology, supervision, writing - review & editing.

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

Ethical Approval: We declare that all ethical guidelines for authors have been followed by all authors. Ethical approval is not required as the data in this study were generated by a computer program.

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

References

Akhtar, H., Silfiasari, Vekety, B., & Kovacs, K. (2023). The effect of computerized adaptive testing on motivation and anxiety: A systematic review and meta-analysis. *Assessment*, *30*(5), 1379–1390. https://doi.org/10.1177/10731911221100995

Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores.*

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/bf02293801

Borgatto, A. F., Azevedo, C., Pinheiro, A., & Andrade, D. (2015). Comparison of ability estimation methods using IRT for tests with different degrees of difficulty. *Communications in Statistics-Simulation and Computation*, *44*(2), 474-488. https://doi.org/10.1080/03610918.2013.781630

Bulut, O., & Sünbül, Ö. (2017). Monte Carlo Simulation Studies in Item Response Theory with the R Programming Language. *Journal of Measurement and Evaluation in Education and Psychology*, *8*(3), 266-287. https://doi.org/10.21031/epod.305821

Chang, H.-H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 211–222. https://doi.org/10.1177/01466219922031338

Choi, S. W., & Lim, S. (2022). Adaptive test assembly with a mix of set-based and discrete items. *Behaviormetrika*, *49*(2), 231-254. https://doi.org/10.1007/s41237-021-00148-6

Choi, S. W., & van der Linden, W. J. (2018). Ensuring content validity of patient-reported outcomes: a shadow-test approach to their adaptive measurement. *Quality of Life Research*, *27*(7), 1683-1693. https://doi.org/10.1007/s11136-017-1650-1

Choi, S. W., Lim, S., & van der Linden, W. J. (2022). TestDesign: an optimal test design approach to constructing fixed and adaptive tests in R. *Behaviormetrika*, *49*(2), 191-229. https://doi.org/10.1007/s41237-021-00145-9

Choi, S. W., Moellering, K. T., Li, J., & van der Linden, W. J. (2016). Optimal reassembly of shadow tests in CAT. *Applied psychological measurement*, *40*(7), 469-485. https://doi.org/10.1177/0146621616654597

Çoban, E. (2020). *Bilgisayar temelli bireyselleştirilmiş test yaklaşımlarının Türkiye'deki merkezi dil sınavlarında uygulanabilirliğinin araştırılması*. Yayınlanmamış Doktora Tezi. Ankara Üniversitesi

Demir, S., & Atar, B. (2021). Investigation of classification accuracy, test length and measurement precision at computerized adaptive classification tests. *Journal of Measurement and Evaluation in Education and Psychology*, *12*(1), 15–27. https://doi.org/10.21031/epod.787865

Ebenbeck, N. (2023). Computerized adaptive testing in inclusive education. Universität Regensburg. https://doi.org/10.5283/EPUB.54551

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

Erdem Kara, B., & Doğan, N. (2022). The effect of ratio of items indicating differential item functioning on computer adaptive and multi-stage tests. *International Journal of Assessment Tools in Education*, *9*(3), 682–696. https://doi.org/10.21449/ijate.1105769

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, *35*(2), 36-49.

Gökçe, S., & Glas, C. A. W. (2018). Can TIMSS mathematics assessments be implemented as a computerized adaptive test? *Journal of Measurement and Evaluation in Education and Psychology*, *9*(4), 422–436. https://doi.org/10.21031/epod.487351

Gündeğer, C., & Doğan, N. (2018). Bireyselleştirilmiş Bilgisayarlı Sınıflama Testi Kriterlerinin Test Etkililiği ve Ölçme Kesinliği Açısından Karşılaştırılması. *Journal of Measurement and Evaluation in Education and Psychology*, *9*(2), 161–177. https://doi.org/10.21031/epod.401077

Han, K. T. (2016). Maximum likelihood score estimation method with fences for short-length tests and computerized adaptive tests. *Applied Psychological Measurement*, *40*(4), 289–301. https://doi.org/10.1177/0146621616631317

Han, K. T., & Guo, F. (2014). Multistage testing by shaping modules on the fly. *Computerized multistage testing: Theory and applications*, 119-133.

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101–125. https://doi.org/10.1177/014662169602000201

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

Huang, Y.-M., Lin, Y.-T., & Cheng, S. C. (2009). An adaptive testing system for supporting versatile educational assessment. *Computers & Education*, *52*(1), 53–67. https://doi.org/10.1016/j.compedu.2008.06.007

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement*, *39*(3), 167–188. https://doi.org/10.1177/0146621614554650

Khorramdel, L., Pokropek, A., Joo, S. H., Kirsch, I., & Halderman, L. (2020). Examining gender DIF and gender differences in the PISA 2018 reading literacy scale: A partial invariance approach. *Psychological Test and Assessment Modeling*, *62*(2), 179-231.

Kim, H., & Plake, B. (1993). *Monte Carlo simulation comparison of two-stage testing and computer adaptive testing*. Unpublished doctoral dissertation, University of Nebraska, Lincoln.

Kirsch, I., & Lennon, M. L. (2017). PIAAC: a new design for a new era. *Large-Scale Assessments in Education*, *5*(1), 1-22. https://doi.org/10.1186/s40536-017-0046-6

Macken-Ruiz, C. L. (2008). *A comparison of multi-stage and computerized adaptive tests based on the generalized partial credit model.* Unpublished doctoral dissertation, University of Texas at Austin

Mooney, C. Z. (1997). *Monte carlo simulation*. Sage.

Mullis, I. V., & Martin, M. O. (2019). *PIRLS 2021 Assessment Frameworks*. International Association for the Evaluation of Educational Achievement. Herengracht 487, Amsterdam, 1017 BT, The Netherlands.

National Center for Education Statistics (NCES). (2019). *Program for International Student Assessment 2022 (PISA 2022) Main Study Recruitment and Field Test*.

Özdemir, B., & Gelbal, S. (2022). Measuring language ability of students with compensatory multidimensional CAT: A post-hoc simulation study. *Education and Information Technologies*, *27*(5), 6273–6294. https://doi.org/10.1007/s10639-021-10853-0

Patsula, L. N. (1999). *A comparison of computerized-adaptive testing and multi-stage testing*. Unpublished doctoral dissertation, University of Massachusetts at Amherst.

Raborn, A., & Sari, H. (2021). Mixed Adaptive Multistage Testing: A New Approach. Journal of measurement and evaluation in education and psychology, *12*(4), 358–373. https://doi.org/10.21031/epod.871014

Şahin, M. G., & Boztunç Öztürk, N. (2019). Analyzing the maximum likelihood score estimation method with fences in ca-MST. *International Journal of Assessment Tools in Education*, *6*(4), 555–567.. https://doi.org/10.21449/ijate.634091

Samejima, F. (1977). A method of estimating item characteristic functions using the maximum likelihood estimate of ability. *Psychometrika*, 42(2), 163-191.

Schnipke, D. L. & Reese, L. M. (1999). A comparison of testlet-based test designs for computerized adaptive testing (Law School Admissions Council Computerized Testing Report 97-01). Newtown, PA: Law School Admission Council.

Sigal, M. J., & Chalmers, R. P. (2016). Play it again: Teaching statistics with Monte Carlo simulation. *Journal of Statistics Education: An International Journal on the Teaching and Learning of Statistics*, *24*(3), 136–156. https://doi.org/10.1080/10691898.2016.1246953

Stafford, R. E., Runyon, C. R., Casabianca, J. M., & Dodd, B. G. (2019). Comparing computer adaptive testing stopping rules under the generalized partial-credit model. *Behavior research methods*, 51(3), 1305-1320. https://doi.org/10.3758/s13428-018-1068-x

Theussl, S., Hornik, K., Buchta, C., Schwendinger, F., Schuchardt, H., & Theussl, M. S. (2019). Package 'Rglpk'. GitHub, Inc., San Francisco, CA, USA, Tech. Rep. 0.6-4.

van der Linden WJ, Diao Q (2014). *Using a universal shadow-test assembler with multistage testing*. In: Yan D, von Davier AA, Lewis C (eds) Computerized multistage testing: theory and applications. CRC Press, New York, 101–118

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*(2), 201–216. https://doi.org/10.1007/bf02294775

van der Linden, W. J. (2009). *Constrained adaptive testing with shadow tests.* Elements of adaptive testing (pp. 31-55). Springer, New York, NY.

van der Linden, W. J. (2010). *Elements of adaptive testing* (Vol. 10, pp. 978-0). C. A. Glas (Ed.). New York, NY: Springer.

van der Linden, W. J. (2022). Review of the shadow-test approach to adaptive testing. *Behaviormetrika*, *49*(2), 169-190. https://doi.org/10.1007/s41237-021-00150-y

van der Linden, W. J., & Chang, H. H. (2003). Implementing content constraints in alpha-stratified adaptive testing using a shadow test approach. *Applied Psychological Measurement, 27*(2), 107-120. https://doi.org/10.1177/0146621602250531

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29*(3), 273-291. https://doi.org/10.3102/10769986029003273

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics: A Quarterly Publication Sponsored by the American Educational Research Association and the American Statistical Association*, *22*(2), 203–226. https://doi.org/10.3102/10769986022002203

Wainer, H. (1990). An Adaptive Algebra Test: A Testlet-Based, Hierarchically-Structured Test with Validity-Based Scoring. Technical Report No. 90-92.

Wang, K. (2017). *A fair comparison of the performance of computerized adaptive testing and multistage adaptive testing* (Unpublished Doctoral Dissertation). Michigan State University.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(2), 109–135. https://doi.org/10.1111/j.1745-3984.1998.tb00530.x

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(2), 109–135. https://doi.org/10.1111/j.1745-3984.1998.tb00530.x

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427-450.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37*(2), 70-84.

Xiao, J., & Bulut, O. (2022). Item Selection with Collaborative Filtering in On-The-Fly Multistage Adaptive Testing. *Applied Psychological Measurement*, 01466216221124089.

Yiğiter, M. S., & Dogan, N. (2023). Computerized multistage testing: Principles, designs and practices with R. *Measurement: Interdisciplinary Research and Perspectives, 21*(4), 254–277. https://doi.org/10.1080/15366367.2022.2158017

Yin, L., & Foy, P. (2021). TIMSS 2023 Assessment Design. TIMSS 2023 Assessment Frameworks, 71.

Zheng, Y., & Chang, H.-H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, *39*(2), 104–118. https://doi.org/10.1177/0146621614544519

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(4); 413-439

A Comparison of the Classification Performances of the DINO Model, Artificial Neural Networks and Non-Parametric Cognitive Diagnosis

Emine YAVUZ* Hakan Yavuz ATAR**

Abstract

The purpose of this study was to compare the attribute (ACR) and pattern-level (PCR) classification rates of the Deterministic-Input, Noisy-Or Gate (DINO) model, Artificial Neural Networks (ANNs), and Non-Parametric Cognitive Diagnosis (NPCD) on simulation datasets. As a comparison condition, the number of attributes, sample size, the number of items, and missing data rate were chosen. A further purpose was to examine the similarities between the classification rates of the DINO model, ANNs, and NPCD on the PISA 2015 collaborative problem-solving (CPS) datasets in various numbers of attributes and sample sizes. For the study, simulation datasets were generated on the basis of the complex Q matrix structures and the DINO model. The conditions for the sample size factor for the real datasets were determined by simple random selection among the participants in the PISA 2015 administration. As a result, it was found that there was a similarity between the DINO model and NPCD classification rates in both simulation and real datasets. In addition, regarding the increase in sample size in both simulation and real datasets, no consistency was found in the increase or decrease of the classification rates of ANNs and NPCD and the similarities of these rates.

Keywords: diagnostic assessment, non-parametric cognitive diagnosis, artificial neural network, DINO, PISA

Introduction

In recent years, the importance of assessment and evaluation of twenty-first-century skills has increased. In parallel with this, the "No child left behind" (2001) act has caused a shift in assessment methods from summative to more diagnostic and formative. In this context, cognitive diagnosis models have started to be used in modeling and evaluating the complex structures of twenty-first-century skills and provide students with detailed feedback formed within the framework of diagnostic and formative assessment. The placement of students into attribute classes in cognitive diagnostics is generally performed through parametric Cognitive Diagnosis Models (CDMs).

When CDMs were first being developed, CDMs estimation algorithms were not publicly available (Chiu et al., 2017), and they usually required large samples as well as involving complex computational calculation procedures that could only be carried out through relatively expensive software, which restricted the use of CDM applications extensively (Chiu & Douglas, 2013; Chiu & Köhn, 2019). In addition, CDM analysis results were found to yield biased results in some situations with complex structures (i.e., a high number of attributes) or fewer items (Shu et al., 2013; Shuying, 2016). Hence, researchers have embarked on a quest to find a non-parametric method to be used in cognitive diagnosis. Furthermore, the need for the development and use of non-parametric cognitive diagnostic techniques has recently increased because these models were found to provide promising results with smaller

*Asst. Prof. Dr.,Erciyes University, Faculty of Education, Kayseri-Türkiye, emineyavuz@erciyes.edu.tr, ORCID: 0000-0002-1991-1416
** Prof. Dr., Gazi University, Faculty of Education, Ankara-Türkiye, hakanyavuzatar@gmail.com, ORCID ID: 0000-0001-5372-1926

To cite this article:
Yavuz, E. & Atar, H.Y. (2023). A Comparison of the classification performances of the DINO model, artificial neural networks and non-parametric cognitive diagnosis. *Journal of Measurement and Evaluation in Education and Psychology, 14*(4), 413-439. https://doi.org/10.21031/epod.1246752

Received: 2.02.2023
Accepted: 20.10.2023

samples and a high number of attributes (Paulsen & Valdivia, 2022). Therefore, this study has focused on the non-parametric perspective in cognitive diagnosis.

Cognitive diagnosis involves non-parametric methods such as the Attribute Hierarchy Method (AHM), clustering techniques (such as k-means and hierarchical agglomeration), Artificial Neural Networks (ANNs), and Non-Parametric Cognitive Diagnosis (NPCD). Since there was no hierarchy among the structure attributes measured in the real dataset, the AHM was not used in this study. Furthermore, clustering techniques that have an exploratory perspective were not included, as the study was carried out in the context of cognitive diagnosis with a confirmatory perspective. In this regard, the present study focused on the NPCD and ANN models used in cognitive diagnostic classification. These methods are similar to CDMs as they require knowledge about the data structure (compensatory or non-compensatory) and Q matrix for the analysis, but they differ from each other in that they make use of different perspectives to classify students according to their attribute profiles.

Non- Parametric Cognitive Diagnosis (NPCD)

Students are categorized in NPCD by comparing their observed response vectors to the ideal response patterns (Chiu & Douglas, 2013). Observed response vectors are the vectors of students' dichotomous responses to the test, and the observed response of a student i is usually represented by y_i. Ideal response patterns are the item response patterns that are expected to occur as a result of the comparison between the attributes that students master and the attributes required for students to respond to the items correctly. It is necessary to know the compensatory state (compensatory or non-compensatory) of the data structure for estimating the ideal response patterns. For the DINO model, which is one of the compensatory models, the ideal response of the student i to item j is calculated by using equation 1:

$$\omega_{ij} = 1-\prod_{k=1}^{K}(1-\alpha_{ik})^{q_{jk}} \qquad (1)$$

In equation 1, α_{ik} is whether the student i has attribute k or not, and q_{jk} is whether the presence of item j requires attribute k. A test with k^{th} attribute involves m=1,..,2^k attribute profiles; therefore, all ideal responses possible for α_i can be stated as ω_m=ω_1, ..., ω_2^k. Since ω_i is determined by α_i, the distance between the observed (y_i) and ideal response patterns of the student i with attribute α_m profile can be represented as d(y_i, α_m). The estimated attribute pattern ($\hat{\alpha}_i$) in NPCD is the attribute pattern that minimizes the distance between all ideal item response patterns and the observed response patterns of student i (Chiu & Douglas, 2013). In other words, $\hat{\alpha}_i$ can be defined as the attribute pattern of the ideal response pattern, which is the most proximate or similar to the observed item response pattern among all ideal item response patterns (Chiu et al., 2017).

$$\hat{\alpha}_i = \arg\min D(y_i, \alpha_m) \quad m \in (1,2,\dots,2^k) \qquad (2)$$

NPCDs estimate the attribute classes by comparing the observed item response patterns with each of the ideal response profiles of the possible 2^k attribute classes (Chiu et al., 2017). Various distance measures (Hamming, weighted Hamming, and penalized Hamming) can be used to measure the similarity between two vectors. (Chiu & Douglas, 2013). It has been determined that the weighted Hamming distance is more efficient, reduces the number of links between possible ideal response patterns, and yields higher classification accuracy results (Paulsen, 2019). Therefore, weighted Hamming was used as the distance measure in this study. The Hamming distance and the weighted Hamming distance for the compensatory data are calculated with equations 3 and 4, respectively:

$$d_h(y_i,\ \alpha_m) = \sum_{j=1}^{J} \left| y_{ij} - \omega_{mj} \right| \qquad (3)$$

$$d_{\mathrm{wh}}(y_i,\ \alpha_m) = \sum_{j=1}^{J} \frac{1}{\overline{p_j}(1-\overline{p_j})} \left| y_{ij} - \omega_{mj} \right| \qquad (4)$$

Multilayer Perceptron Artificial Neural Networks (MLP-ANN)

These are the models inspired by the structures and functions of neurons in the human brain. After Gierl et al. proposed to use ANNs together with AHM in 2007, ANNs started to be used within the scope of CDM. ANNs are employed in the context of CDM through two different learning paradigms: supervised and unsupervised learning paradigms. We utilized the supervised learning paradigm since it is compatible with the confirmatory perspective of CDMs.

The multilayer perceptron ANN (MLP-ANN) can be defined as a parallel processing architecture (Garson, 1998; Gierl et al., 2007, 2008) that receives stimuli with input units and carries these to the output unit with latent units. An input layer, at least one latent layer, and an output layer are typically present in these ANNs. A variety of neurons with various roles make up each layer. The latent layer(s) in the network helps to model the effects of the interaction of input neurons on output neurons. In the MLP-ANN, neurons in a particular layer cannot interact with each other directly, but they can only connect with neurons in the adjacent layers. These connections are called weights. The process of estimating the weights in an ANN or correlating the input layer with the output layer is called ‘ANN training’ or ‘ANN learning’. The training of the network is carried out iteratively in such a way that each iteration trains the network to minimize the difference (error) between the expected and observed attribute values for an estimated response pattern. In this iterative method, weight estimations are initialized with arbitrary values drawn from the ordinary normal distribution (the default approach of the R package neuralnet). In this study, weight back-tracking was used for the weight estimations in the framework of the resilient back-propagation approach. After the training of the ANN is completed, the cognitive diagnosis is terminated by analyzing a new dataset in which the inputs are known, but the outputs are not known.

Within the scope of CDMs, each input layer neuron of ANNs represents an item that constitutes the test, and the responses given to those items are used as input. The neurons of the output layer are interpreted as the attributes that the predetermined test is intended to measure. The expected response patterns derived from the Q matrix, or the collection of ideal response vectors produced from the Q matrix, are the inputs utilized to train the ANN. The results are the pertinent attribute profiles based on the validity of the Q matrix (Cui et al., 2016; 2017). Training of the network refers to the establishment of the connection between these ideal response vectors and the relevant attribute profiles. The training process continues until the neural network learns the connections. After the network has completed the learning process, students’ responses to the test items are put into the network as input for analysis. It is possible to summarize the mathematical form of an ANN training with the following steps. If it is assumed that there is a three-layer MLP-ANN where i refers to input neurons, j to latent neurons, and k to output neurons, the weighted sum of all input neurons is obtained in the first step.

$$a_j = \sum_{j=1}^{J} W_{ji} X_i \qquad (5)$$

$a_{j,}$ is the weighted sum for latent neuron j, W_{ji} is the weight of the connection from input neuron i to latent neuron j, and X_i is the value of input neuron i. This sum is converted to the following form by the activation function f(·) to calculate the latent neuron value:

$$h_j = \mathrm{f}\left(a_i\right) = f\left(\sum_{j=1}^{I} W_{\mathrm{ji}} X_i\right) \qquad (6)$$

Since the logistic/sigmoid function is frequently preferred as the activation function in the context of cognitive diagnosis (Guo et al., 2017), the logistic function was used in this study in the same way to determine the weight of two connections. When the latent neuron values have been obtained, the output neuron values must be calculated. The output neurons are calculated in a way that is similar to the calculation of the latent neurons.

The association of the input neurons with the output neurons through two successive conversions continues until the error function is reduced to the minimum or to a predetermined value or until it is fixed. In other words, the values of the connection weights that minimize the error function are selected for the estimation of the model parameters. In this study, the cross-entropy value, which is an error calculation function, was examined, and the number of hidden layers and neurons was determined. Since only MLP-ANN was studied within the scope of this research, ANN refers to MLP-ANN in this study.

There are some studies in which classification rates of ANNs and NPCDs are examined together under various conditions (see McCoy & Willse, 2014; Paulsen, 2019; Paulsen & Valdivia, 2022) in the related literature. The current research has some similarities and differences with the study of McCoy and Willse (2014) and Paulsen (2019). The similarities include the binary coding of attributes (0-1) and the multivariate normal distribution of attributes in simulation data. Since the data in the study of McCoy and Willse (2014) and Paulsen (2019) were created based on the Deterministic-Input, Noisy-And Gate (DINA) model, the findings obtained from the studies can be generalized to non-compensatory data structures. The use of both the real and DINO-based simulation datasets and the analysis of the classification performances of compensatory models based on the data structures are the features that distinguish the present study from the previous studies. The current study also differed in the number of attributes used in data analysis. While McCoy and Willse's (2014) and Paulsen's (2019) studies had a maximum of eight attributes, the current study examined a maximum of 11 attributes. Additionally, in the studies of McCoy and Willse (2014) and Paulsen (2019), item discrimination was evaluated as high or low, whereas in the current study, item discrimination was handled at a moderate level, based on the recommendation of the previous studies (see Guo et al., 2017; Shuying, 2016). Finally, the current research differs from the previous studies in that it examines the missing data effect in the field of non-parametric cognitive diagnostics and the effect of NPCD and ANN on classification rates. It is expected that comparing the findings of McCoy and Willse (2014) and Paulsen (2019) with the findings obtained from the simulation datasets generated based on real and compensatory models in this study will provide a holistic perspective regarding the effects of various factors on classification in non-parametric cognitive diagnosis.

Finally, this study is considered important as it provides a basis for further studies to be conducted in this field, especially on the use of NPCD and ANNs in evaluating students' twenty-first-century skills and comparing the classification performances of these methods. In light of this, the aims of this research are to compare the attribute and pattern-level classification rates of the DINO models, ANN, and NPCD in various settings in DINO-based simulation datasets and then to look at the similarities of the classification rates of the DINO models, ANN, and NPCD in the PISA 2015 CPS dataset. For this purpose, the study sought answers to the following questions:

- Do the attribute and pattern-level classification rates of the DINO models, ANN, and NPCD in simulation data differ according to the number of attributes (3, 5, and 7), sample sizes (30, 100, and 500), the number of items (15, 30, and 45) and the missing data rates (0, .05, and .10)?
- What is the similarity of the classification rates of the DINO models, ANN, and NPCD in the real data (PISA 2015 CPS) according to the number of attributes (3, 7, and 11) and sample sizes (30, 100, and 500)?

Methods

Research Design

Since this research aims to determine the classification performances of the DINO model, ANN, and NPCD under different settings, this study was designed as a simulation study. Simulation studies are frequently used to assess the performance of a specific statistical model or to predict the results of a given situation.

The Population

The PISA 2015 CPS administration data was utilized in the study due to the convenience of obtaining the Q matrix. A total of 414,498 students from 52 countries participated in the CPS assessment. Since Form 93 and Form 96 in the PISA 2015 CPS administration included more items representing the CPS structure, Form 93 and Form 96 were used in the study. Furthermore, the items scored as polytomous previously were rescored as dichotomous to maintain the content validity. After removing the missing data, 18,170 students from 43 countries were used in the data analysis. Using a simple random sampling procedure in IBM SPSS 26.0, sample sizes of 30, 200, and 500 students were sampled from the population of 18,170 students.

Research Procedure

The conditions of the simulation and the real datasets are explained in detail in the following sections.

Obtaining Simulation Datasets

In order to ensure the compatibility of the simulation datasets with the structure of the real datasets, simulation datasets were generated based on the DINO model, which is a frequently used compensatory model. The first step of the data generation process was the determination of the Q matrix structures. In the creation of the Q matrices, it was ensured that each attribute was measured by the same number of items and that the items measuring more than one attribute were also equal in number (see de la Torre, 2008; de la Torre & Douglas, 2008; Rupp & Templin, 2008a). An item was allowed to measure a maximum of three attributes in each Q matrix. In addition, the increase in the complexity of all the Q matrix structures used in the study indicated an increase in the number of attributes. Finally, all the Q matrices were completed (Chiu et al., 2009). Within the scope of the research, firstly, a 15-item Q matrix was created, then this structure was used twice to create the 30-item Q matrix and three times to create the 45-item Q matrix. The 15-item Q matrices of the simulation datasets are reported in Table 11 in the appendices.

After the determination of the Q matrix structures, features of the attributes were identified. Multivariate normal distribution was used in generating the attributes of the simulation data for describing the realistic situations in which the attribute patterns were not equally distributed and the attributes were correlated. The correlation value between the attributes was determined to be $\rho = .5$ (see Chiu & Douglas, 2013; McCoy & Willse, 2014).

After the distributions of attributes and the level of correlation between the attributes were determined, the item parameters (s and g parameters) were identified. It was found that the s and g parameter values were between .33 and .35 when the number of attributes in the real datasets used in the study (PISA 2015 CPS competency) was 3, 7, and 11. In order to make the structure of the simulation datasets as similar as possible to the structure of the real datasets and not to exceed the reliability of classification accuracy, s and g parameter values in all conditions were determined as U [0, .4] in uniform distribution.

The factors addressed in the study were the number of attributes (3, 5, and 7), sample sizes (30, 100, and 500), number of items (15, 30, and 45), and missing data rate (0, 5%, and 10%). These factors and their conditions were selected based on the literature review due to their frequent use in the simulation and real data research and the rich information they provide to the implementers and the readers. Within the scope of the study, the comparison condition of 3x3x3x3x3=243 was created in the simulation datasets, and 100 replications were carried out for each comparison condition of the simulation data.

Data collection tools for real datasets and data collection

In this study, the PISA 2015 CPS data were used due to the accessibility of the information about the Q matrix as the real datasets. The PISA 2015 CPS competency is a structure that is formed by the combination of collaboration skills and PISA 2012 individual problem-solving process skills. The skills constituting the PISA 2015 CPS competency and details about the skill(s) measured with these items were described in OECD reports (2017a; 2017b). Q matrices were primarily created by making use of these reports based on expert opinions, and the real datasets were analyzed. The Q matrices of the first 17 items are reported in Table 12 in the appendices. After Q matrix validation was performed with the PVAF method (de la Torre & Chiu, 2015), the analyses of the real datasets were performed again, and the findings obtained from both types of the Q matrices were interpreted together. The factors used in the analysis of the real datasets were the number of attributes (3, 7, and 11) and the sample sizes (30, 100, and 500). Furthermore, the comparison condition of 3x3x3=27 was created in the real datasets.

Data Analysis

We used R-Studio for the data generation and the validation and analyses of the Q matrices of the real datasets, and IBM SPSS 26.0 software (IBM, 2019) for the factorial ANOVA (analysis of variance) in the study. Moreover, various packages were used for different purposes: GDINA 2.8.0 (Ma et al., 2020) was used for the validation of the Q matrices of the real data; the packages of CDM 7.4-19 (Robitzsch et al., 2019) for DINO analysis; the packages of NPCD 1.0-11 (Zheng et al., 2019) for the NPCD analysis; the packages of neuralnet 1.44.2 (Fritsch et al., 2019) for ANN analysis; the packages of missForest 1.4 (Stekhoven, 2016) for the missing data creation, and TestDataImputation 1.1 (Dai et al., 2019) for missing data imputation.

The simulation datasets were analyzed by following the procedures of NPCD, DINO model, and ANN analyses to answer the first research question, while the classification performances were determined in various conditions. For the simulation dataset, the classification performances of the NPCD, DINO model, and ANN were investigated through ACR and PCR. The equations of ACR and PCR are presented below.

$$ACR \; = \frac{1}{NxK} \sum_{i=1}^{N} \sum_{i=1}^{K} I[\widehat{\alpha_{\imath k}} = \alpha_{ik}] \qquad (7)$$

$$PCR \; = \frac{1}{N} \sum_{i=1}^{N} I[\widehat{\alpha_{\imath}} = \alpha_{i}] \qquad (8)$$

where N is the sample, K is the number of attributes, i is a student, α_{ik} is the k. attribute that the student i. actually has, and $\hat{\alpha}_{ik}$ is the k. attribute that the student i. is estimated to have.

After the ACRs and PCRs were obtained for the DINO model, ANN, and NPCD, factorial ANOVAs were carried out, and the effect sizes were interpreted according to Cohen (1988). Cohen (1988) classified effect sizes as small (.20), medium (.50), and large (.80).

The analyses of the same real datasets (PSA 2015 CPS) were performed by making use of two different Q matrices that were created based on the technical reports and the PVAF method to address the second research question. ACRs and PCRs of the DINO model ANN and NPCD could not be calculated due to the fact that the real attribute classes of the students were not known in the real datasets. As a result, equations 11 and 12 were modified in accordance with the real data analyses, and the similarity of the attribute (SACR) and pattern-level classification rates (SPCR) of the DINO model, ANN, and NPCD were obtained. In this regard, α_{ik} was replaced with α value estimated by one of the compared models and $\hat{\alpha}_{ik}$ with α value estimated by the other model.

Results

1.Comparison of the ACRs and PCRs of the DINO model, ANN, and NPCD in the simulation data based on the number of attributes (3, 5, and 7), sample sizes (30, 100, and 500), number of items (15, 30, and 45) and missing data rate (0, .05, and .10)

The ACRs and PCRs of the DINO model, ANN, and NPCD under study conditions are given in Table 13 in the appendices.

1.1. Comparison of the ACRs and PCRs of the DINO modes, ANN, and NPCD according to the number of attributes (3, 5, and 7)

The results of the ACRs of the DINO model, ANN, and NPCD, and the comparison of the number of attributes (3, 5, and 7) factors are presented in Table 1. It can be seen in Table 1 that the interaction between the number of attributes and the models showed a statistically significant difference (p<.01) and that this interaction had a medium effect (η_p^2 = .39).

Table 1

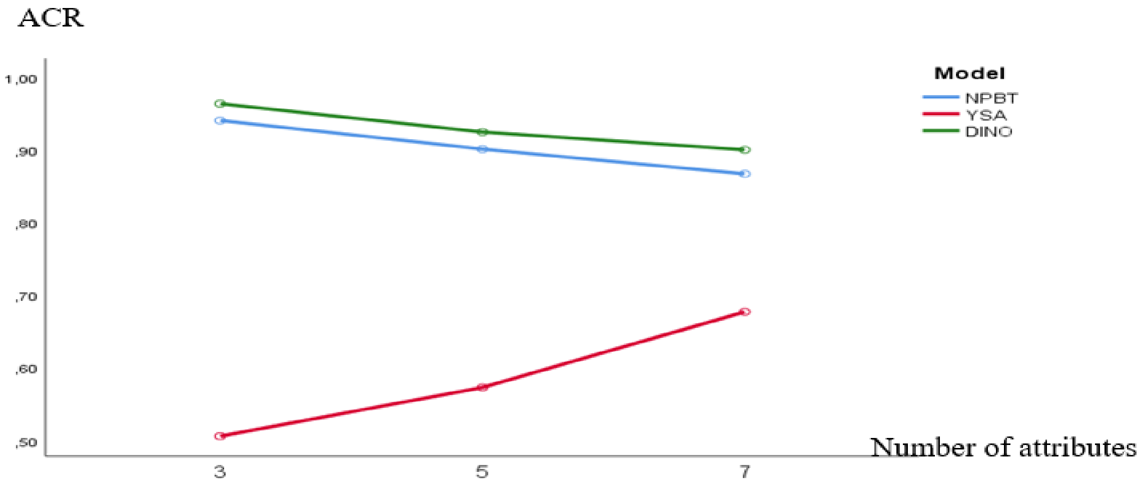
Factorial ANOVA Results of the DINO Model, ANN, and NPCD's ACRs according to the Number of Attributes

Source of variation	Sum of squares	*df*	Mean square	*F*	*p*	η_p^2
Number of attributes	1.018	2	0.509	153.127	0.000	0.012
Models	595.986	2	297.993	89683.310	0.000	0.881
Number of attributes* Models	52.177	4	13.044	3925.781	0.000	0.393
Error	80.712	24291	0.003			

Figure 1 shows that as the number of attributes increased, the ACRs of the DINO model and NPCD decreased, whereas the ACRs of the ANN increased. In addition, it was observed that the DINO model had the highest average ACR and that ANN had the lowest average ACR under all conditions of the number of attributes.

Figure 1

The ACRs of the DINO Model, ANN, and NPCD according to the number of attributes



The results of the PCRs of the DINO model, ANN, and NPCD, and the comparison of the number of attributes (3, 5, and 7) factor are presented in Table 2. Table 2 demonstrates that the interaction between the number of attributes of the PCRs and the models showed a statistically significant difference (p<.01) and that this interaction had a small effect (η_p^2 = .29).

Table 2

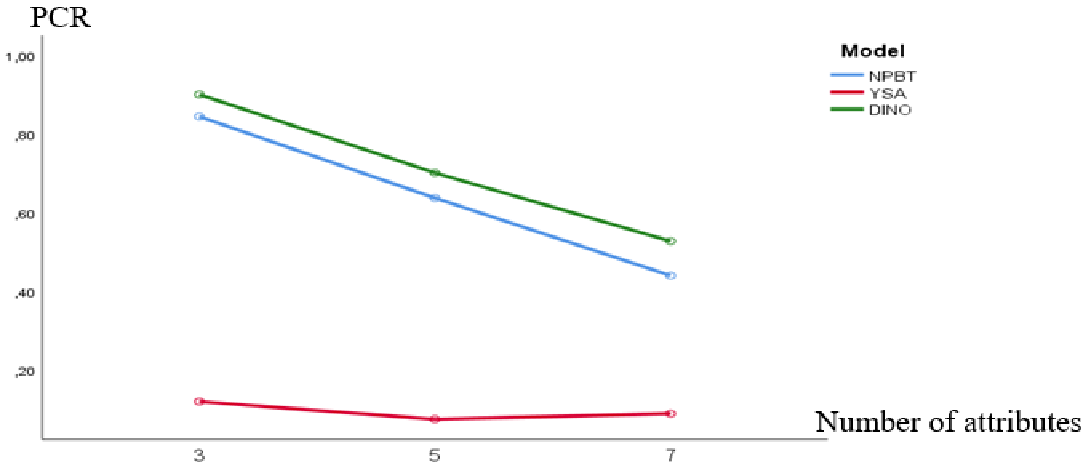
Factorial ANOVA Results of the DINO model, ANN, and NPCD's PCRs according to the Number of Attributes

Source of variation	Sum of squares	*df*	Mean square	*F*	*p*	η_p^2
Number of attributes	295.924	2	147.962	12421.698	0.000	0.506
Models	1847.983	2	923.991	77570.806	0.000	0.865
Number of attributes* Models	116.889	4	29.222	2453.265	0.000	0.288
Error	289.344	24291	0.012			

Figure 2 shows that the PCRs of the DINO model, ANN, and NPCD decreased as the number of attributes increased. It can also be seen that the DINO model had the highest average ACR and that ANN had the lowest average ACR under all conditions of the number of attributes.

Figure 2

The PCRs of the DINO Model, ANN, and NPCD according to the number of attributes



1.2. Comparison of the ACRs and PCRs of the DINO model, ANN, and NPCD according to the sample sizes (30, 100, and 500)

The results of the ACRs of the DINO model, ANN, and NPCD and the comparison of the sample sizes (30, 100, and 500) factor are presented in Table 3. Table 3 shows that the interaction between the sample sizes of the ACRs and the models was statistically significant (p<.01) and that this interaction had a small effect (η_p^2 = .002).

Table 3

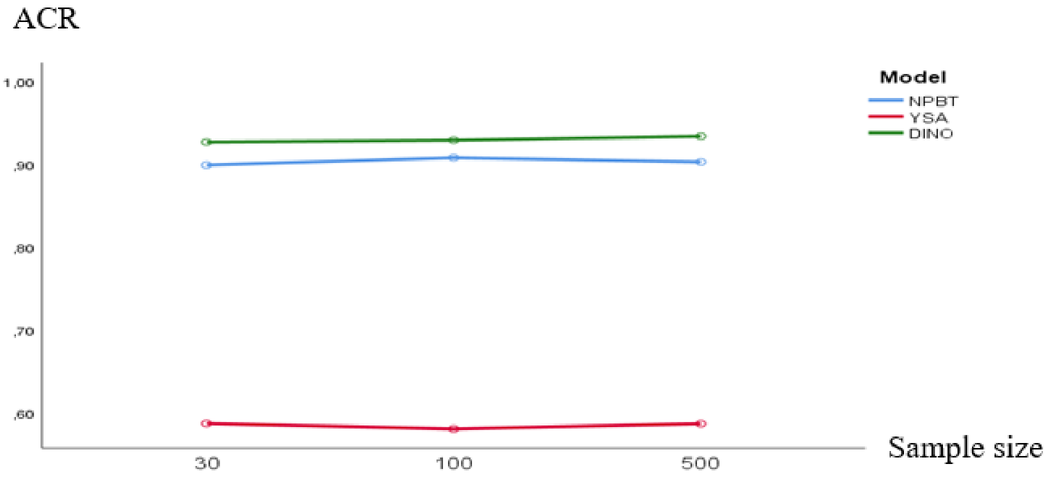
Factorial ANOVA Results of the DINO model, ANN, and NPCD's ACRs according to the Sample Sizes

Source of variation	Sum of squares	*df*	Mean square	*F*	*p*	η_p^2
Sample size	0.051	2	0.026	4.655	0.010	0.000
Models	595.986	2	297.993	54159.499	0.000	0.817
Sample size * Models	0.203	4	0.051	9.244	0.000	0.002
Error	133.652	24291	0.006			

Figure 3 shows that the ACRs of the NPCD, ANN, and DINO model did not change as the sample size increased. It can be seen that the DINO model had the highest average ACR and that ANN had the lowest average ACR under all conditions of the sample sizes.

Figure 3

ACRs of the DINO Model, ANN, and NPCD according to the sample sizes



The results of the PCRs of the DINO model, ANN, and NPCD and the comparison of the sample sizes (30, 100, and 500) factor are presented in Table 4. Table 4 shows that the interaction between the sample size of PCRs and the models showed a statistically significant difference (p<.01) and that this interaction had a small effect (η_p^2 = .003).

Table 4

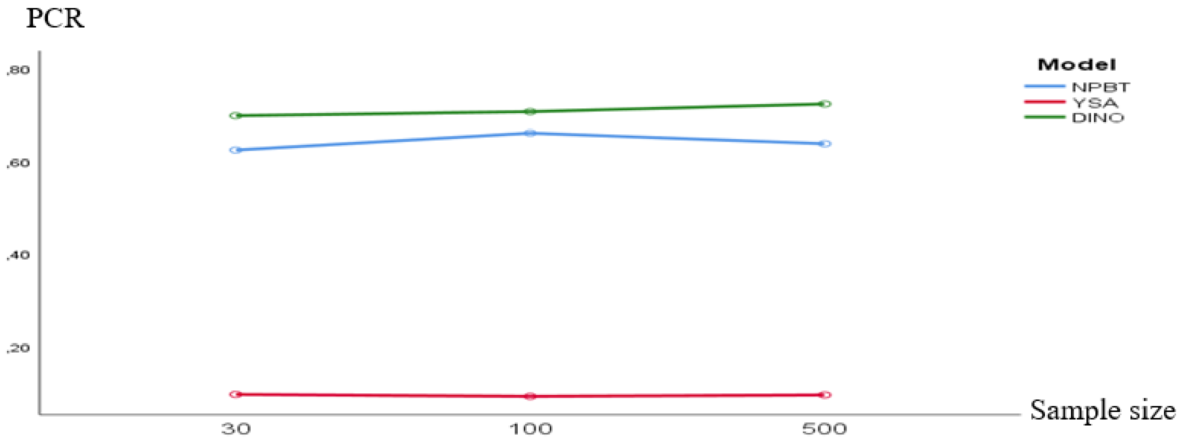
Factorial ANOVA Results of the DINO model, ANN, and NPCD's PCRs according to the Sample Sizes

Source of variation	Sum of squares	*df*	Mean square	*F*	*p*	η_p^2
Sample size	0.970	2	0.485	16.849	0.000	0.001
Models	1847.983	2	923.991	32094.844	0.000	0.725
Sample size* Models	1.864	4	0.466	16.188	0.000	0.003
Error	699.323	24291	0.029			

Figure 4 shows that when the sample sizes increased from 30 to 100, the PCRs of the DINO model and NPCD increased, whereas the PCRs of the ANN decreased. When the sample size was 500, it was observed that the PCRs of the NPCD decreased, the PCRs of the ANN showed no change, and the PCRs of the DINO model increased. Moreover, it was observed that the DINO model had the highest average PCR and that the ANN had the lowest average PCR under all conditions of the sample sizes.

Figure 4

The PCRs of the DINO Model, ANN, and NPCD according to sample sizes



1.3. Comparison of the ACRs and PCRs of the DINO model, ANN, and NPCD according to the number of items (15, 30, and 45)

The results of the ACRs of the DINO model, ANN, and NPCD and the comparison of the number of items (15, 30, and 45) factor are presented in Table 5. Table 5 shows that the interaction between the number of items of the ACRs and the models was statistically significant (p<.01) and that this interaction had a medium effect (η_p^2 = .38).

Table 5

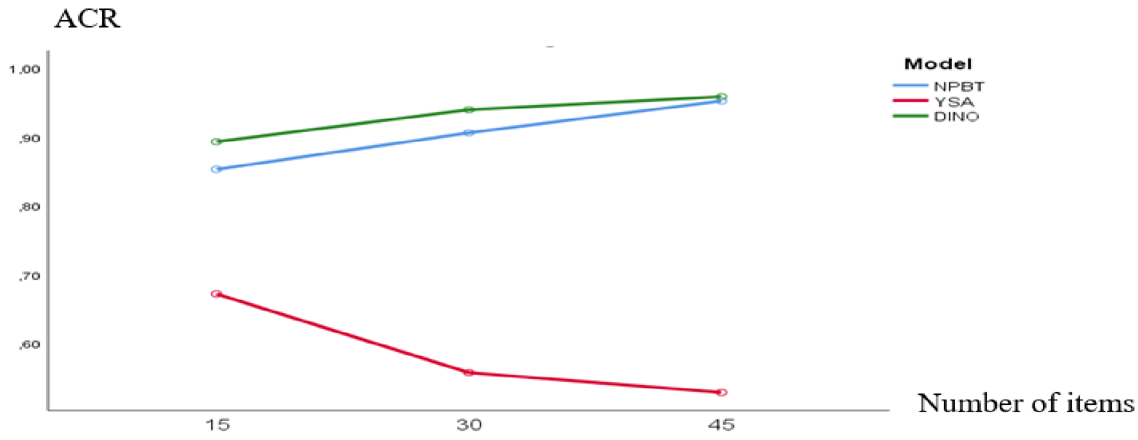
Factorial ANOVA Results of the DINO model, ANN, and NPCD's ACRs according to the Number of Items

Source of variation	Sum of squares	*df*	Mean square	*F*	*p*	η_p^2
Number of items	0.611	2	0.306	89.075	0.000	0.007
Models	595.986	2	297.993	86883.579	0.000	0.877
Number of items* Models	49.983	4	12.496	3643.283	0.000	0.375
Error	83.313	24291	0.003			

Figure 5 shows that as the number of items increased, the ACRs of the DINO model and NPCD increased, and the ACRs of the ANN decreased. When the number of items was 15 and 30, the DINO model had the highest average ACR, whereas the ANN had the lowest average ACR. Furthermore, when the number of items was 45, the DINO model and NPCD had the highest average ACR, whilst ANN had the lowest average ACR.

Figure 5

ACRs of the DINO Model, ANN, and NPCD according to the number of items



The results of the PCRs of the DINO model, ANN, and NPCD and the comparison of the number of items (15, 30, and 45) factor are presented in Table 6. Table 6 shows that the interaction between the number of items of the PCRs and the models was statistically significant (p<.01) and that this interaction had a small effect (η_p^2 = .19).

Table 6

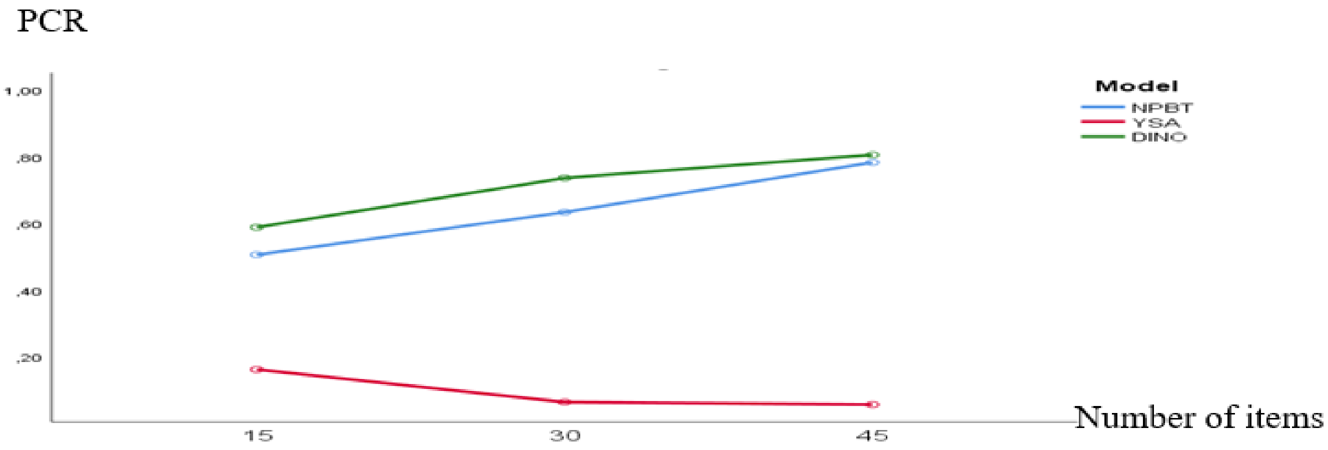
Factorial ANOVA Results of the DINO model, ANN, and NPCD's PCRs according to the Number of Items

Source of variation	Sum of squares	*df*	Mean square	*F*	*P*	η_p^2
Number of items	68.501	2	34.251	1621.845	0.000	0.118
Models	1847.983	2	923.991	43753.111	0.000	0.783
Number of items* Models	120.672	4	30.168	1428.522	0.000	0.190
Error	512.985	24291	0.021			

Figure 6 shows that when there was an increase in the number of items, the PCRs of the DINO model and NPCD increased, and the PCRs of the ANN decreased. It is also demonstrated that the DINO model had the highest average PCR and that ANN had the lowest average PCR under all conditions of the number of items.

Figure 6

PCRs of the DINO Model, ANN, and NPCD according to the number of items



1.4. Comparison of the ACRs and PCRs of the DINO Model, ANN, and NPCD according to the Missing Data Rates (0, .5, and .10)

The results of the ACRs of the DINO model, ANN, and NPCD and the comparison of the missing data rate (0, .05, and .10) factor are presented in Table 7. Table 7 shows that the interaction between the missing data rate and the models was statistically significant (p<.01) and that this interaction had a small effect (η_p^2 = .001).

Table 7

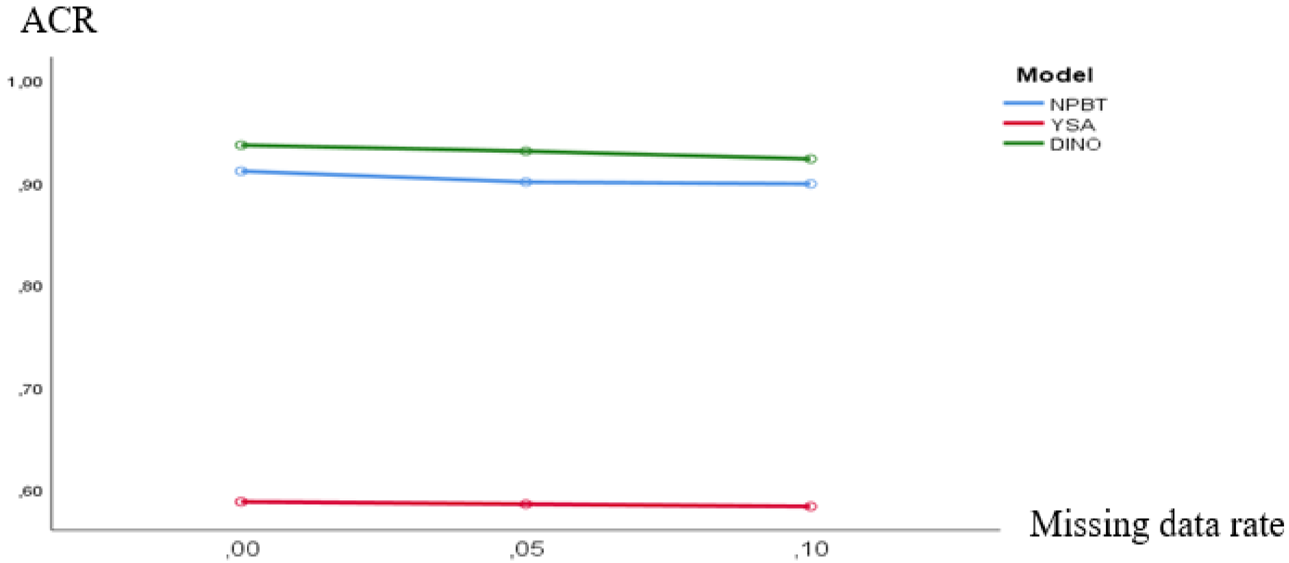
Factorial ANOVA Results of the DINO model, ANN, and NPCD's ACRs according to the Missing Data Rate

Source of variation	Sum of squares	*df*	Mean square	*F*	*p*	η_p^2
Missing data rate	0.424	2	0.212	38.614	0.000	0.003
Models	595.986	2	297.993	54266.062	0.000	0.817
Missing data rate * Models	0.093	4	0.023	4.235	0.002	0.001
Error	133.390	24291	0.005			

Figure 7 shows that the DINO model had the highest average ACRs compared with the ANN and NPCD, whereas the ANN had the lowest average ACRs under all conditions of the missing data rate.

Figure 7

The ACRs of the DINO Model, ANN, and NPCD according to the missing data rate



The results of the PCRs of the DINO model, ANN, and NPCD and the comparison of the missing data rate (0, .05, and .10) factor are presented in Table 8. Table 8 shows that the interaction between the missing data rate of the PCRs and the models was statistically significant (p<.01) and that this interaction had a small effect (η_p^2 = .001).

Table 8

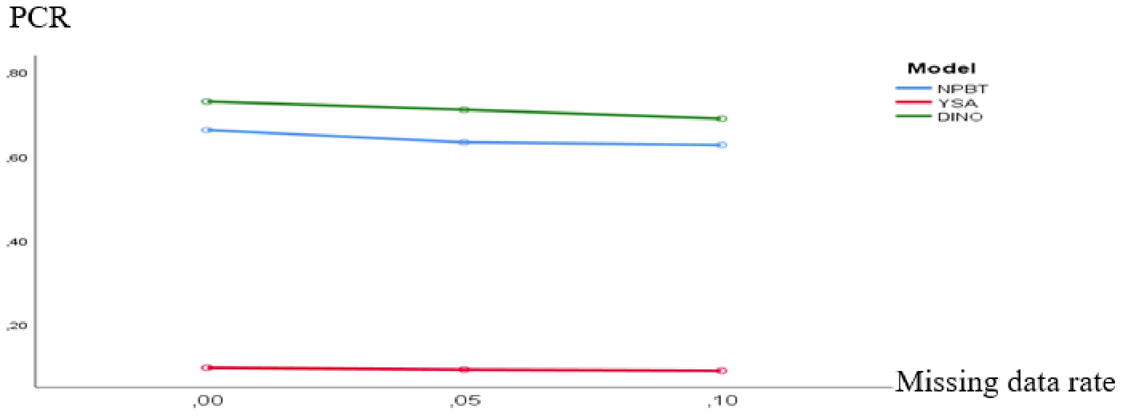
Factorial ANOVA Results of the DINO model, ANN, and NPCD's PCRs according to the Missing Data Rate

Source of variation	Sum of squares	*df*	Mean square	*F*	*p*	η_p^2
Missing data rate	3.230	2	1.615	56.219	0.000	0.005
Models	1847.983	2	923.991	32160.645	0.000	0.726
Missing data rate * Models	1.035	4	0.259	9.004	0.000	0.001
Error	697.893	24291	0.029			

Figure 8 shows that the DINO model had the highest PCRs compared with ANN and NPCD and that ANN had the lowest PCRs under all conditions of the missing data rate.

Figure 8

The PCRs of the DINO Model, ANN, and, NPCD according to the missing data rate



As a result of the factorial ANOVAs, the interaction effects for all conditions were found to be statistically significant. In all conditions, NPCD had slightly lower but comparable classification rates than the DINO model, while ANN always had lower rates than the NPCD and DINO model.

2. Similarities of the classification rates of the DINO model, ANN, and NPCD in the real dataset according to the number of attributes (3, 7, and 11) and sample sizes (30, 100, and 500)

2.1. Similarity of the classification rates of the DINO model, ANN, and NPCD according to the number of attributes (3, 7, and 11)

The SACRs and SPCRs of the DINA model,ANN, and NPCD based on different numbers of attributes (3,7, and 11) is presented in Table 9.

Table 9

Variation of SACR and SPCRs of the DINO Model, ANN, and NPCD according to the Number of Attributes in the Real Dataset

		Result of the Q matrix based on the technical reports						Result of the validated Q matrix					
		NPCD-ANN		NPCD-DINO		ANN-DINO		NPCD-ANN		NPCD-DINO		ANN-DINO	
N	*A*	SACR	SPCR	SACR	SPCR	SACR	SPCR	SACR	SPCR	SACR	SPCR	SACR	SPCR
30	3	.730	.471	.941	.875	.705	.472	.944	.867	.822	.533	.789	.467
	7	.714	.201	.808	.302	.721	.207	.786	.233	.771	.400	.805	.267
	11	.570	.002	.827	.070	.573	.003	.800	.033	.854	.267	.624	.00
100	3	.773	.532	.843	.684	.722	.481	.783	.470	.867	.710	.717	.380
	7	.691	.127	.776	.316	.658	.062	.741	.330	.771	.430	.819	.410
	11	.706	.038	.805	.087	.685	.014	.765	.090	.805	.080	.722	.080
500	3	.781	.515	.879	.664	.696	.426	.717	.436	.861	.668	.761	.496
	7	.682	.141	.777	.326	.679	.157	.790	.362	.860	.594	.824	.418
	11	.654	.026	.833	.123	.624	.018	.637	.030	.852	.334	.612	.018

Note: *A: Number of attributes, N: Sample size*

The SACRs and SPCRs obtained from the validated Q matrix were higher than the SACRs and SPCRs obtained from the Q matrix based on technical reports. In the findings obtained from Q matrices, the similarity between the SACRs and SPCRs of the DINO model and NPCD was high. It can be stated that as the number of attributes increased in the Q matrices, the SACRs between the ANN and NPCD and the SACRs between the NPCD and DINO model first decreased and then increased, whereas the SPCRs generally decreased.

2.2. Similarity of the classification rates of the DINO model, ANN, and NPCD according to the sample sizes (30, 100, and 500)

The SACRs and SPCRs of the DINO model, ANN, and NPCD according to different sample sizes (30, 100, and 500) are examined in Table 10.

Table 10

Variation of SACR and SPCRs of the DINO models, ANN and NPCD according to the Sample Sizes in the Real Dataset

		Result of the Q matrix based on the technical reports						Result of the validated Q matrix					
		NPCD-ANN		NPCD-DINO		ANN-DINO		NPCD-ANN		NPCD-DINO		ANN-DINO	
N	*A*	SACR	SPCR	SACR	SPCR	SACR	SPCR	SACR	SPCR	SACR	SPCR	SACR	SPCR
3	30	.730	.471	.941	.875	.705	.472	.944	.867	.822	.533	.789	.467
	100	.773	.532	.843	.684	.722	.481	.783	.470	.867	.710	.717	.380
	500	.781	.515	.879	.664	.696	.426	.717	.436	.861	.668	.761	.496
7	30	.714	.201	.808	.302	.721	.207	.786	.233	.771	.400	.805	.267
	100	.691	.127	.776	.316	.658	.062	.741	.330	.771	.430	.819	.410
	500	.682	.141	.777	.326	.679	.157	.790	.362	.860	.594	.824	.418
11	30	.570	.002	.827	.070	.573	.003	.800	.033	.854	.267	.624	.00
	100	.706	.038	.805	.087	.685	.014	.765	.090	.805	.080	.722	.080
	500	.654	.026	.833	.123	.624	.018	.637	.030	.852	.334	.612	.018

Note: A: Number of attributes, N: Sample size

This table shows that the SACRs and SPCRs obtained from the validated Q matrix were higher than the SACRs obtained from the Q matrix based on the technical reports. In the findings obtained from the Q matrices, no consistency was found regarding the increase or decrease of the SACR and SPCR values of NPCD-ANN, the NPCD-DINO model, and ANN-DINO model. In addition, it was observed that the SACRs and SPCRs between the DINO model and NPCD were more similar than those between ANN and NPCD and those between ANN and the DINO model in the findings obtained from the Q matrices.

Discussion

The aims of the study were to compare the attribute and pattern-level classification rates of the DINO model, ANN, and NPCD on the DINO-based simulation datasets based on various conditions and to examine the similarities between the classification rates of the DINO model, ANN, and NPCD on the PISA 2015 CPS dataset. In the current study, simulation datasets were generated similar to the structure of the real datasets in order to obtain comparable results from both datasets. With these aims in mind, the structure of the PISA 2015 CPS competency was examined, and it was found that there was no sequential or prerequisite relationship among the attributes, namely the problem-solving skills and collaboration skills, which constitute the PISA 2015 CPS competency. In other words, a student who has one or more problem-solving skills can solve a problem even if s/he does not have the other skills (Yavuz, 2014), which indicates that these skills are compensatory. Since all of the items are not shared in the technical reports on the PISA 2015 CPS competency, the extent to which each attribute contributes to the correct response cannot be determined for the items that require more than one attribute for the correct response. Therefore, the DINO model, which assumes that each attribute contributes equally to the correct response, was selected for the analysis of CPS competency and the generation of simulation datasets. In addition, it was found that one of the skills of the CPS competency, which consists of 12 skills, was not measured in the PISA 2015 CPS administration (see OECD, 2017). For this reason, the conditions of the attribute factor in the research were set as 3, 7, and 11. Moreover, since some conditions prevented the modeling, model-based data imputation methods could not be used to impute missing data. Instead, a two-way data imputation method was used in the current study. Therefore, the findings of this study were discussed within these limitations.

Discussion Of The Simulation Dataset

This study differs from other studies with ANN (see Cui, et al., 2016; Guo et al., 2017; McCoy & Willse, 2014; Paulsen, 2019; Paulsen & Valdivia, 2022; Shu et al., 2013) due to the DINO-based generation of simulation datasets. All the findings were discussed in light of the findings of previous studies conducted with the DINO and DINA models. As a result of the analyses, it was found that ANN always had lower rates than the DINO model and NPCD, whereas NPCD had slightly lower but comparable classification rates than the DINO model in all conditions. In parallel with this result, McCoy and Willse (2014) and Paulsen (2019), who studied the classification rates of the DINA model, ANN, and NPCD, found that the classification rates of NPCD and the DINA model were similar to each other whilst ANN consistently had lower classification rates in comparison with the DINA model and NPCD.

Shu et al. (2013) compared the classification rates of ANN, MCMC-ANN, JMLE-ANN, and the DINA model in simulation datasets with low item discrimination (s and g parameter values between .2 and .4) and a complex Q matrix structure. Shu et al. (2013) found that the DINA model cannot make estimations when the sample size is 50 or less in cases with four attributes and when the sample size is 150 or less in cases with six attributes. The studies in the related literature (Chiu et al., 2017; Paulsen, 2019; Paulsen & Valdivia, 2022) have also shown that the DINA model can also make estimations in small samples of 25 and 30 participants thanks to the increase in computational capacity and the use of the EM algorithm in parameter estimation of items. Likewise, it was also found in the current study that the DINO model could make estimations in the conditions of 30 participants, the smallest sample size.

In addition to the increase in the computational capacity and the use of the EM algorithm, the DINO model is thought to have higher classification rates than ANN and NPCD in small samples due to the fact that the simulation datasets were generated based on the DINO model and the balanced distribution of attributes and items constituting the Q matrix structures was taken into consideration in the data generation (see de la Torre, 2008; de la Torre et al., 2010; Rupp & Templin, 2008a). In the literature, there are studies with various results. In parallel with the results of the current study, Chiu and Douglas (2013) stated that the maximum likelihood estimation estimates better with the most suitable parametric model for the data structure; therefore, the DINO model can make better estimations than NPCD in some small DINO-based simulation samples. Similar to the results of the DINO model and NPCD, Cui et al. (2016) and Shu et al. (2013) also found that the DINA model had higher classification rates than ANN in DINA-based simulation datasets.

In the literature, there are studies showing that ANN and NPCD have higher and lower classification rates than the DINO or DINA models, depending on the conditions. For instance, McCoy and Willse (2014) studied DINA-based datasets and found that the classification rates of the DINA model were slightly higher than NPCD and considerably higher than ANN in some conditions. Chiu et al. (2017) stated that NPCD had higher classification rates than the DINA model when the sample size was 100 or less, whereas the DINA model had higher classification rates than NPCD when the sample size was 500. Furthermore, Paulsen (2019) found that NPCD had better classification rates than the DINA model when the sample size was 25, the number of items was small, and the item discrimination was low. Finally, Ma et al. (2020) found that NPCD had higher classification rates than the DINA model in samples of 30 and 50 and that the DINA model had higher classification rates than NPCD in samples of 200 and 500.

There are also studies in the literature showing that ANN and NPCD have higher classification rates than the DINA and DINO models in all conditions. For instance, Akbay (2016) examined the classification rates of the NPCD, DINO, and DINA models in various conditions by generating simulation datasets of 250, 500, and 1000 participants based on DINA and DINO. He found in his study that NPCD had higher classification rates than the DINO model in a DINO-based dataset, and NPCD had higher classification rates than the DINA model for a DINA-based dataset. In their DINA-based dataset with a sample of 5000 participants, Guo et al. (2017) found that ANN had a higher classification rate than the DINA model.

Discussion of the real dataset (PISA 2015 CPS)

In the current study, no consistency was found regarding the increase, decrease, or stability of the SACRs and SPCRs of the DINO model, ANN, and NPCD when the sample size increased (30, 100, and 500). This is an expected finding, considering that, in theory, the ACRs and PCRs of ANN and NPCD are not affected by sample size. No study was found in the literature that had investigated either the similarity of classification rates of the DINO model, ANN, and NPCD or the similarity of classification rates of the DINA model, ANN, and NPCD in the real datasets. Nevertheless, only three studies were found in the literature which estimated the similarity of classification rates of different models. Chiu and Douglas (2013) calculated the similarity of classification rates of NPCD and HODINA, while Chiu et al. (2017) calculated the similarity of classification rates of GNPCD and G-DINA. Lim and Drasgow (2017) calculated the similarity of classification rates of NPCD based on expert opinion. However, these studies were carried out with a single sample.

In conclusion, the similarity between the classification rates of the DINO model and NPCD was observed across both simulation and real datasets. In addition, no consistency was found regarding the increase or decrease in the classification rates of ANN and NPCD in simulation datasets and the similarities of these rates in real datasets when the sample size increased in both datasets. It was also observed that the variations in the classification rates and the similarity of these rates differed as the number of attributes in the simulation and the real datasets differed.

There are some limitations in this study. First, since not all of the real data sets were shared, it could not be determined which attribute contributed more to the correct answer of the items. Therefore, the DINO model, which is assumed to contribute equally to each attribute in answering the item correctly, was chosen as a parametric analysis. Second, the first Q matrices were created according to the technical reports (OECD, 2017a; 2017b) based on this reason. Third, model-based data imputation methods could not be used in the missing data imputation since some conditions prevented the establishment of the model; instead, the two-way data assignment method was used. There are some suggestions considering these limitations.

The findings of the current study have shown that the classification rates of the DINO model and NPCD were similar. It is thought that evaluating the results of the DINO model and NPCD together will increase the classification reliability and hence can contribute to the reliable assessments of students as well as their placement into correct attribute classes. For this reason, it is suggested that implementers use the DINO model and NPCD together if they are going to perform cognitive diagnoses in small samples.

The classification rates of the DINO model, ANN, and NPCD can be further investigated by changing the attributes, research factors, software, and packages of the simulation datasets used in this study. Based on the analysis of the classification rates of the DINO model, ANN, and NPCD, examining the PISA 2015 CPS construct with a three-attribute Q matrix was found to be more reliable. In OECD (2017a; 2017b) technical reports, these three attributes are described as competency areas of the "establishment and maintenance of a common understanding", the "identification of proper actions to solve the problem", and the "establishment and maintenance of team organization". Researchers who intend to investigate the PISA 2015 CPS construct by using different methods are recommended to conduct their investigation based on these three sub-competency areas.

Acknowledgments

This study was supported by the Scientific and Technological Research Council of Turkey under Grant 2228-B.

Declarations

Conflict of Interest: The authors of the article declare that they have no conflict of interest with any person or organization that may be a party to this study.

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

References

Akbay, L. (2016). Relative efficiency of the nonparametric approach on attribute classification for small sample cases. *Journal of European Education 6*(1), 43-59.

Chiu, C.-Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, 30, 225-250. doi:10.1007/s00357-013-9132-9

Chiu, C.-Y., Douglas, J., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*(4), 633-665.

Chiu, C.-Y., & Köhn, H.-F. (2019). Consistency theory for the general nonparametric classification method. *Psychometrika*, *84*(3), 830-845.

Chiu, C.-Y., Sun, Y., & Bian, Y. (2017). Cognitive diagnosis for small educational programs: The general nonparametric classification method. *Psychometrika*, *83*(2), 355-375.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Erlbaum.

Cui, Y., Gierl, M., & Guo, Q. (2016). Statistical classification for cognitive diagnostic assessment: An artificial neural network approach. *Educational Psychology: An International Journal of Experimental Educational Psychology*, *36*(6), 1065-1082.

Cui, Y., Gierl, M., & Guo, Q. (2017). The rule space and attribute hierarchy methods. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment frameworks, methodologies, and applications* (pp. 354-378). John Wiley & Sons.

Dai, S. (2017). *Investigation of missing responses in implementation of cognitive diagnostic models.* (Doctoral dissertation). Indiana University, Boston.

Dai, S., Wang, X., & Svetina, D. (2019, March). Test Data Imputation [software package in R]. https://cran.r-project.org/web/packages/TestDataImputation/index.html

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

de la Torre, J., & Chiu, C.-Y. (2015). A general method of empirical Q-matrix validation. *Psychometrika*, *81*(2), 253-273.

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*(4), 595-624.

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement*, *47*(2), 227-249.

Fritsch, S., Guenther, F., Wright, M. N., Suling, M., & Mueller, S. M. (2019, February). Neuralnet [software package in R]. https://cran.r-project.org/web/packages/neuralnet/neuralnet.pdf

Garson, G. D. (1998). *Neural networks: An introductory guide for social scientists*. Sage.

Gierl, M. J., Cui, Y., & Hunka, S. (2007, April). *Using connectionist models to evaluate examinees' response patterns on tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL, USA.

Gierl, M. J., Cui, Y., & Hunka, S. (2008). Using connectionist models to evaluate examinees' response patterns on tests. *Journal of Modern Applied Statistical Methods*, 7, 234-245. doi:10.22237/jmasm/1209615480

Guo, Q., Cutumisu, M., Cui, Y. (2017). *A neural network approach to estimate student skill mastery in cognitive diagnostic assessments*. Poster presented at the 10th International Conference on Educational Data Mining, Wuhan, Hubei Province, in China.

He, Q., von Davier, M., Greiff, S., Steinhauer, E. W., & Borysewicz, P. B. (2017). Collaborative problem-solving measures in the Programme for International Student Assessment (PISA). In A. A. von Davier, M. Zhu & P. C. Kyllonen (Eds.), *Methodology of educational measurement and assessment: Innovative assessment of collaboration* (pp. 95-112). Springer.

IBM Corp. (2019). IBM SPSS Statistics for Windows (Version 22.0) [Computer software]. https://www.ibm.com/tr-tr/analytics/spss-statistics-software

Lim, Y. S., & Drasgow, F. (2017): Nonparametric calibration of ıtem-by-attribute matrix in cognitive diagnosis. *Multivariate Behavioral Research, 52*(5), 562-575.

Ma, C., de la Torre, J., & Xu, G. (2020). *Bridging Parametric and Nonparametric Methods in Cognitive Diagnosis*. Retrieved from arXiv:2006.15409

Ma, W., de la Torre, J., Sorrel, M. & Jiang, Z. (2020, May). GDINA [software package in R]. https://cran.r-project.org/web/packages/GDINA/index.html

McCoy, T., & Willse, J. (2014, April). *Accuracy of neural network versus nonparametric approaches in diagnostic classification*. Paper presented at the National Council on Measurement in Education, Washington, DC, USA.

No Child Left Behind Act of 2001, Pub. L. No. 107-110. Retrieved from http://thomas.loc.gov/

OECD (2017a). *PISA 2015 collaborative problem-solving framework*. France. https://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf

OECD (2017b). *PISA 2015 results (Volume V): Collaborative problem solving*. France. https://www.oecd.org/publications/pisa-2015-results-volume-v-9789264285521-en.htm#:~:text=PISA%202015%20Results%20(Volume%20V)%3A%20Collaborative%20Problem%20Solving%2C%20is,try%20to%20solve%20a%20problem

Paulsen, J. (2019). *Examining cognitive diagnostic modeling in small sample contexts*. (Doctoral dissertation). Indiana University, Boston.

Paulsen, J. & Valdivia, D. S. (2022). Examining cognitive diagnostic modeling in classroom assessment conditions. *Journal of Experimental Education, 90*(4), 916-933. doi:10.1080/00220973.2021.1891008

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2019, September). CDM [software package in R]. https://cran.r-project.org/web/packages/CDM/index.html

Rupp, A. A., & Templin, J. (2008a). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*(1), 78-96.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement*. Guilford.

Shu, Z., Henson, R., & Willse, J. (2013). *Using neural network analysis to define methods of DINA model estimation for small sample sizes*. *Journal of Classification*, 30, 173-194. doi:10.1007/s00357-013-9134-7

Shuying, S. (2016). *Nonparametric diagnostic classification analysis for testlet based tests.* (Doctoral dissertation). The University of North Carolina at Greensboro, USA.

Stekhoven, D. J. (2016, August). missForest [software package in R]. https://cran.r-project.org/web/packages/missForest/missForest.pdf

Sünbül, S. (2018). The impact of different missing data handling methods on DINA model. *International Journal of Evaluation and Research in Education, 7*(1), 77-86.

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

Wang, M.-J. (2015). *Computer-based assessment of collaborative problem-solving for intermediate elementary students with one on one, human-to-agent approach.* (Master's dissertation). National Taichung University, Taiwan.

Wang, S., & Douglas, J. (2015). Consistency of nonparametric classification in cognitive diagnosis. *Psychometrika, 80*(1), 85-100.

Yavuz, E. (2014). *Determining the problem solving process skills of the primary education pre-service mathematics teachers as defined in PISA*. Gazi University, Ankara.

Zheng, Y., Chiu, C.-Y., & Douglas, J. A. (2019, November). NPCD [software package in R]. https://cran.r-project.org/web/packages/NPCD/NPCD.pd

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(4); 413-439

Appendices

Table 11

Q Matrices Used in Generating Simulation Datasets

Item no	The Q matrix with three attributes			The Q matrix with five attributes					The Q matrix with seven attributes						
1	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0
2	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0
3	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0
4	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0
5	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0
6	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0
7	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1
8	1	1	0	1	1	0	0	0	1	1	0	0	0	0	0
9	0	1	1	0	0	1	1	0	0	1	1	0	0	0	0
10	1	0	1	1	1	0	0	0	0	0	1	1	0	0	0
11	0	1	1	0	1	1	0	0	0	0	0	1	1	0	0

*Asst. Prof. Dr.,Erciyes University, Faculty of Education, Kayseri-Türkiye, emineyavuz@erciyes.edu.tr, ORCID: 0000-0002-1991-1416
** Prof. Dr., Gazi University, Faculty of Education, Ankara-Türkiye, hakanyavuzatar@gmail.com, ORCID ID: 0000-0001-5372-1926

To cite this article:
Yavuz, E. & Atar, H.,Y. (2023). A Comparison of the classification performances of the DINO model, artificial neural networks and non-parametric cognitive diagnosis. *Journal of Measurement and Evaluation in Education and Psychology, 14*(4), 413-439. https://doi.org/10.21031/epod.1246752

Received: 2.02.2023
Accepted: 20.10.2023

12	1	0	1	0	0	1	1	0	0	0	0	0	1	1	0
13	1	1	0	0	0	0	1	1	0	0	0	0	0	1	1
14	1	1	1	1	0	0	1	1	1	0	0	1	0	0	1
15	1	1	1	1	0	1	0	1	1	1	1	0	0	0	0

Table 12

Q Matrices for Top 17 Items for Real Data

Item no	The Q matrix with three attributes			The Q matrix with seven attributes							The Q matrix with eleven attributes										
	1	2	3	A	B	C	D	1	2	3	A1	A2	B1	B2	B3	C1	C2	C3	D1	D2	D3
1	0	1	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
2	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
3	0	1	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
4	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
5	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
6	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
7	0	0	1	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0

8	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
9	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
10	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1
11	0	1	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
12	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
13	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0
14	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
15	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
16	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
17	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0

Table 13

The ACRs and PCRs of NPCD, ANN, and the DINO Model under Study Conditions

			Classification Rates of NPCD						Classification Rates of ANN						Classification Rates of DINO					
			ACRs			PCRs			ACRs			PCRs			ACRs			PCRs		
			Missing Date Rate			Missing Date Rate			Missing Date Rate			Missing Date Rate			Missing Date Rate			Missing Date Rate		
A	*N*	*I*	0%	5%	10%	0%	5%	10%	0%	5%	10%	0%	5%	10%	0%	5%	10%	0%	5%	10%
3	30	15	.930	.903	.883	.810	.751	.707	.622	.580	.577	.214	.178	.167	.945	.933	.928	.850	.819	.806
		30	.956	.949	.941	.879	.862	.84	.429	.513	.496	.074	.093	.111	.978	.979	.967	.934	.938	.900
		45	.978	.979	.979	.938	.939	.943	.480	.476	.467	.107	.111	.092	.989	.984	.969	.966	.951	.911
	100	15	.877	.875	.902	.708	.699	.752	.508	.579	.576	.115	.168	.160	.941	.933	.926	.84	.818	.799
		30	.957	.949	.954	.883	.861	.873	.56	.447	.429	.161	.065	.073	.981	.983	.972	.948	.953	.920
		45	.971	.983	.968	.919	.950	.910	.484	.478	.475	.108	.104	.099	.989	.987	.978	.966	.962	.937
	500	15	.907	.869	.924	.761	.692	.801	.575	.512	.574	.163	.114	.162	.947	.940	.929	.852	.833	.805
		30	.945	.954	.944	.854	.875	.849	.430	.530	.432	.076	.115	.073	.974	.974	.968	.925	.927	.910
		45	.992	.982	.97	.977	.947	.915	.472	.482	.479	.100	.099	.101	.988	.985	.982	.963	.956	.948
5	30	15	.842	.796	.847	.471	.380	.477	.654	.684	.681	.127	.165	.148	.895	.885	.880	.588	.568	.539

		30	.883	.885	.899	.573	.572	.608	.536	.549	.591	.040	.044	.072	.936	.928	.926	.728	.704	.703
		45	.946	.948	.944	.775	.783	.765	.523	.484	.482	.041	.024	.022	.957	.943	.961	.803	.746	.829
	100	15	.881	.862	.853	.554	.526	.510	.668	.644	.654	.136	.119	.127	.886	.877	.869	.562	.527	.521
		30	.922	.91	.901	.686	.665	.619	.540	.540	.524	.044	.041	.033	.943	.939	.928	.764	.745	.728
		45	.977	.952	.943	.897	.794	.767	.488	.487	.485	.025	.025	.028	.967	.954	.947	.848	.801	.777
	500	15	.878	.849	.876	.560	.478	.552	.680	.699	.688	.152	.169	.156	.901	.891	.882	.607	.586	.559
		30	.909	.917	.898	.643	.684	.621	.535	.524	.595	.041	.036	.069	.944	.942	.931	.765	.756	.718
		45	.961	.940	.936	.828	.757	.742	.522	.514	.511	.041	.040	.037	.965	.961	.955	.847	.828	.809
7	30	15	.816	.818	.808	.285	.286	.288	.788	.761	.757	.193	.160	.152	.854	.854	.843	.375	.367	.347
		30	.844	.878	.87	.357	.446	.413	.632	.654	.629	.046	.048	.039	.909	.904	.892	.557	.525	.484
		45	.923	.925	.92	.606	.610	.581	.617	.565	.644	.036	.012	.052	.938	.934	.929	.665	.648	.620
	100	15	.823	.838	.793	.306	.358	.260	.795	.782	.757	.214	.188	.156	.868	.863	.979	.393	.388	.356
		30	.907	.874	.835	.538	.415	.355	.662	.668	.661	.058	.060	.057	.917	.909	.900	.575	.547	.517
		45	.957	.936	.929	.758	.662	.631	.570	.679	.562	.020	.077	.015	.940	.933	.922	.674	.647	.602
	500	15	.805	.777	.797	.271	.234	.257	.793	.786	.760	.203	.195	.158	.874	.865	.857	.415	.395	.38
		30	.894	.858	.832	.495	.384	.336	.670	.628	.636	.061	.038	.041	.923	.916	.907	.603	.579	.541
		45	.930	.916	.930	.630	.565	.630	.643	.566	.633	.051	.017	.044	.949	.944	.937	.712	.688	.658

Note: A: Number of attributes. N: Sample size. I: Number of items

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(4); 440-454

Comparing Estimated and Real Item Difficulty Using Multi-Facet Rasch Analysis *

Ayfer SAYIN **   Sebahat GÖREN ***

Abstract

This study aimed to compare estimated item difficulty based on expert opinion with real item difficulty based on data by utilizing Rasch analysis. For security reasons, some high-stakes tests are not pre-tested and item difficulty is estimated by teachers in classroom assessments, so it is necessary to examine the extent to which experts make accurate predictions. In this study, we developed a 12-item test in the field of measurement and evaluation similar to those used in the Public Personnel Selection Exam. Item difficulty was estimated and compared separately based on 1165 student responses and the opinions of 12 experts. A multi-facet Rasch analysis was conducted to examine the effects of raters on the test scores. The study revealed that the experts had a good ability to estimate item difficulty for items of moderate difficulty. However, they tended to underestimate item difficulty for items.

Keywords: test development, item difficulty, subject matter experts, multi-facet Rasch

Introduction

Item difficulty, a crucial factor for educational assessments and personalized learning resource recommendations, is a concept that measures how difficult a test item is for a given group of test takers. For the exams prepared for these assessments to be effective, item difficulties need to be adjusted. Especially in standardized tests that are used to differentiate between students with different abilities, there is a need to use items of different difficulties: easy, moderate, and hard. This requires writing test items that meet certain quality standards to ensure that achievement on each item is linked to overall test performance, but it is challenging to write an item at a certain difficulty (Yaneva et al., 2020). For this, the item writer needs to have a good experience of the factors that affect item difficulty. In standardized tests, the difficulty of items should be estimated after the item is written and before it is administered. Besides automatic estimation methods, historically, this has been done through expert validation or pre-tests. Pre-testing is resource-intensive and requires considerable time and effort (Lin et al., 2019). Moreover, piloting for classroom assessments is costly and pre-testing is not preferred for safety reasons in high-stakes tests.

Recently, there has been increasing interest in using new methods, such as neural networks, machine learning, exoplanet item response theory, etc. to predict item difficulty (He et al., 2021; Qiu et al., 2019; Yaneva et al., 2020), and here we present findings on predictors of test item difficulty. Chon and Shin (2010) identified potential predictors of test item difficulty, such as response time and paragraph length based on related research and data collected from the College Scholastic Aptitude Test (CSAT). Beinborn et al. (2014) developed a model for the cloze-test difficulty that includes four dimensions: solution difficulty, candidate ambiguity, gap dependency, and paragraph difficulty. The results suggest that all four dimensions contribute to the overall difficulty of the C-test. Stadler et al. (2016) state that item difficulty can be accurately predicted using six key item characteristics, including the use and

* A part of this study was presented at 7th International Congress on Measurement and Evaluation in Education and Psychology. Gazi University, Ankara-Türkiye.

** Assoc. Prof. Dr., Gazi University, Faculty of Education, Ankara-Türkiye, ayfersayin@gazi.edu.tr, ORCID ID: 0000-0003-1357-5674

*** Research Assistant. PhD., Hacettepe University, Faculty of Education, Ankara-Türkiye, sebahatgoren@gmail.com, ORCID ID: 0000-0002-6453-3258

To cite this article:
Sayın, A. & Gören, S. (2023). Comparing estimated and real item difficulty using multi-facet rasch analysis. *Journal of Measurement and Evaluation in Education and Psychology, 14*(4), 440-454. https://doi.org/10.21031/epod.1310893
Received: 7.06.2023
Accepted: 20.10.2023

number of self-dynamics, the number of input and output variables, the number of input and output variables not related to other variables, and the total number of relationships between all variables. Toyama (2021) found that several features of a passage, including sentence length, word frequency, syntactic simplicity, and temporality have a significant impact on comprehension difficulty. However, all these methods require the inclusion of predictor variables that predict item difficulty in the models they develop to predict item difficulty. However, it is not possible to talk about a variable that affects item difficulty. For example, while many studies have found that longer items tend to be more difficult (Fergadiotis et al., 2019; Lin et al., 2021; Pandarova et al., 2019; Yaneva et al., 2020), it has also been observed that longer items can be easier or item length does not affect item difficulty (Sano, 2015; Toyama, 2021). These findings showed that the difficulty model developed for one test may not be valid for other tests. In addition, in some types of modeling, difficulty features were identified, extracted, and presented as rules by experts (Beinborn et al., 2014; Grivokostopoulou et al., 2014; Perikos et al., 2016; Perkins et al., 1995). Therefore, it is crucial to determine the predictor variables to be used in the construction of models for item difficulties, and in this case, it is important to evaluate the accuracy of experts' difficulty predictions. Hence, expert opinions are often used for estimating item difficulty, but these estimates may differ from students' actual experience (Impara & Plake, 1998).

Expert estimation uncertainty may be due to a variety of factors involved in the cognitive process required to answer a question, as well as the tendency of test creators to overestimate student performance. Moreover, there is no standard that expert estimates accurately reflect item difficulty (Kurdi et al., 2021). Research has examined expert estimates and the cognitive operations involved in test items, but there is no guidance on what experts focus on when making difficulty estimates. Therefore, improving the accuracy of expert estimates of test difficulty requires a better understanding of the relationship between expert estimates and item difficulty (Hamamoto Filho et al., 2020). Attali et al. (2014) found that judges were successful in ranking multiple items in terms of difficulty, this ranking remained consistent among judges and across content areas of the Scholastic Aptitude Test [SAT]. Similarly, experts' ability to estimate item difficulty varied across different studies, with some showing good accuracy (Enright et al., 1993; Le Hebel et al., 2019; Lumley et al., 2012) and others showing limited predictive power (Kibble & Johnson, 2011; Sydorenko, 2011). For this reason, further studies should be conducted to determine the factors underlying the item difficulty of the experts because the predictions of the experts and item difficulty are not only important in the test development process, but also in the interpretation of test scores.

Item difficulty is crucial in setting the standard cut-off for passing or failing an exam. The Angoff method, which involves judges estimating the percentage of average examinees who will answer each test item correctly, is a commonly used criterion-referenced approach in determining the standard cut-off (Afrashteh, 2021; Wyse, 2020). The Angoff method is used to determine the final cut-off score by calculating the average of estimates made by referees for each item. This method is commonly used in high-stakes exams, such as medical exams, as it places a high value on expert opinions (Clauser et al., 2017; Impara & Plake, 1998; Kardong-Edgren & Mulcock, 2016; Wyse, 2018; Yim & Shin, 2020). In the current study, the emphasis was placed on assessing measurement and evaluation items that are similar to those found in the Public Personnel Selection Examination-[PPSE]

Many countries use selection and placement tests forteacher candidates. For example, The Praxis® exams are used to evaluate academic and subject-specific knowledge in the USA, according to the Educational Testing Service[ETS] (Praxis, 2022). It is worth noting that some states with significant teacher populations, such as California, New York, Texas, and Florida, have their own separate licensing exams (Gitomer & Qi, 2010). The Australian Institute for Teaching and School Leadership (AITSL) administers a range of assessments for teacher candidates, including the National Literacy and Numeracy Test for Initial Teacher Education students (AITSL, 2022); The Teaching Council of New Zealand requires all teacher candidates to pass the New Zealand Teachers Council Literacy and Numeracy Professional Skills Test (Ell, 2021). PPSE(Turkish KPSS) in Turkey includes a version specifically for individuals seeking to become teachers in the public school system. This test focuses on education-related subjects, such as pedagogy, educational psychology, and teaching methodologies

(OSYM, 2022). In the teacher certification exam for public institutions, there are 12 items related to assessment and evaluation. The difficulty of these items and the test as a whole is determined through expert opinions.

The aim of this research is to compare the accuracy of expert opinions in estimating item difficulty with real item difficulty based on data, particularly for high-stakes tests. This research lies in providing insights into the accuracy of expert estimates and identifying potential biases that may influence test scores. This information can be useful for improving the reliability and validity of high-stakes tests and ensuring that they accurately measure the knowledge and skills of test-takers.

Methods

Research Model

In this study, the relational survey design, which is a quantitative research method, has been used to demonstrate the relationship between multiple variables without intervention (Büyüköztürk et al.,2020).

Participants

Data were collected from two groups: pre-service teachers and experts who estimate the difficulty of the items. As summarized in Table 1, the first group of participants in the study included 1165 pre-service teachers who were in their third or fourth year of study at a faculty of education in a university. They all took a 14-week course on assessment and evaluation. The second group of participants comprised 12 experts who have either more than five years (5+) or less than five years (0-5) in the field of measurement and evaluation. They were also employed as instructors in education faculties, teaching courses related to assessment and evaluation. Experts who have more than five years of experience are familiar with both the course content and the participant group as they teach the students' courses, and experts who have less than five years of experience are acquainted with both the course content and the participant group as they assist the measurement and evaluation courses at the same universities which the data were collected. While the experts who have more than five years of experience prepare the exams themselves, the experts who have less than five years of experience help to prepare these exams as they assist the courses, and all experts determine the difficulty of the exams themselves. Since the difficulty of the PPSE is determined according to expert opinion, the difficulty of the achievement test developed in this study was also determined based on expert opinion.

Table 1.

Sample Descriptive Statistics

Participants 1 Pre-service teachers			Participants 2 Subject matter experts		
Characteristic	f	%	Characteristic	f	%
Gender			Gender		
Female	752	64.5	Female	10	83.3
Male	413	35.5	Male	2	16.7
Grade			Experiment		
3rdgrade	657	56.4	0-5 years	7	58.3
4th grade	508	43.6	5+ years	5	41.7

Table 2.
Sample Descriptive Statistics (Continued)

Participants 1 Pre-service teachers			-
Characteristic	f	%	
Department			
Turkish and Social Sciences Education	531	45,6	
Foreign Languages Education	252	21,6	
Primary Education	189	16,2	
Mathematics and Science Education	98	8,4	
Special Education	95	8,2	

Instrument

Achievement Test

In this study, we developed a 12-item test in the field of measurement and evaluation similar to those used in thePPSE.. Teacher candidates who apply to the PPSE to be appointed to the Ministry of National Education teacher positions are also required to take the Educational Sciences Test. Among the eight subtests within the Educational Sciences Test, the measurement and evaluation subtest accounts for approximately 6% of the total (OSYM Guide, 2023). In other words, there are 12 items from measurement and evaluation in the PPSE Educational Sciences Test, which consists of 80 items in total. In the achievement test developed in this study, the expert-subject effect of item difficulty perception was also tried to be determined by creating 12 items in five subjects including the most frequently asked subjects (alternative test tools, traditional test tools, item statistics, test statistics, interpretion of test scores). In addition, the fact that the items in the PPSE were prepared according to the university course contents supports the information that the items are appropriate for the course contents.

Prior to conducting factor analysis for the test's construct validity, the researchers examined whether the correlation matrix was suitable for factor analysis. Based on the results of the KMO and Bartlett's sphericity test (KMO=0.82, Bartlett's test=3039.97), exploratory factor analysis (EFA) was conducted. The parallel analysis showed that the difficulty of the responses for the items loaded a single dimension (see Appendix A). The results of the analysis revealed that the items explained 42% of the variance in the students' responses. Additionally, the reliability of the test was found to be 0.80 based on Cronbach's alpha coefficient. The factor loading values obtained from the EFA ranged between 0.484 and 0.792 (see Appendix B). Factor loading values greater than 0.30 for each factor indicate that the items serve the dimension well (Tabachnick & Fidell, 2013).

Expert Opinion Form

The experts were asked to estimate the difficulty of each item in the test. An expert opinion form was used in this process. In this form, the participants were first asked whether they had detailed information about the assessment items in PPSE, and those who answered "yes" to this item were included in the study. Then, the factors affecting the degree of difficulty were explained in detail in the form. Information was given about the semester averages of the participant group and the PPSE success ranking of those who graduated from the same department. Considering this information, 12 measurement and evaluation experts estimated item difficulty by rating each one on a scale of 1 to 5, where 1=very hard, 2=hard, 3=medium, 4=easy, and 5=very easy.

Data Analysis

The study analyzed data from 1165 students using exploratory factor analysis and the Rasch IRT model. The "ltm" package in R studio was used for exploratory factor analysis, and the Rasch IRT model was analyzed using the "TAM" package in R studio. In addition, multi-faceted Rasch analysis was conducted using the Minifac (Facets) package program to analyze data collected from 12 raters who rated 12 items. The study examined four sources of variability, including raters, items, item facets, and rater experience, and assessed model-data fit by examining standardized residual values. The results showed that there were seven values (0.48%) within the ±2 interval and 2 values (0.14%) within the ±3 interval, indicating acceptable model-data fitting. The data met all the assumptions, allowing for the analyses to be conducted.

Results

Real item difficulty (n=1165 students)

According to the results of Rasch IRT analysis, the difficulty parameters for each item are presented in Table 3. As it can be seen the item difficulty parameters in the test vary between -1.645 and 0.899. Furthermore, the test comprises items of varying difficulty levels: easy items (evidenced by negative coefficients), those of medium difficulty (coefficients near zero), and difficult items (marked by positive coefficients). The knowledge function in Figure 1 is a graphical representation of how well the test discriminates individuals with different ability levels. It shows that the test information function has a peak around 0 on the ability axis, indicating that the test is most informative for individuals with ability levels around 0. This means that the test is most accurate in discriminating students whose ability is close to the average ability level required for the test. The results of the Rasch analysis suggest that each item provides useful information about the difficulty parameters and the ability of the test to discriminate between individuals with different ability levels.

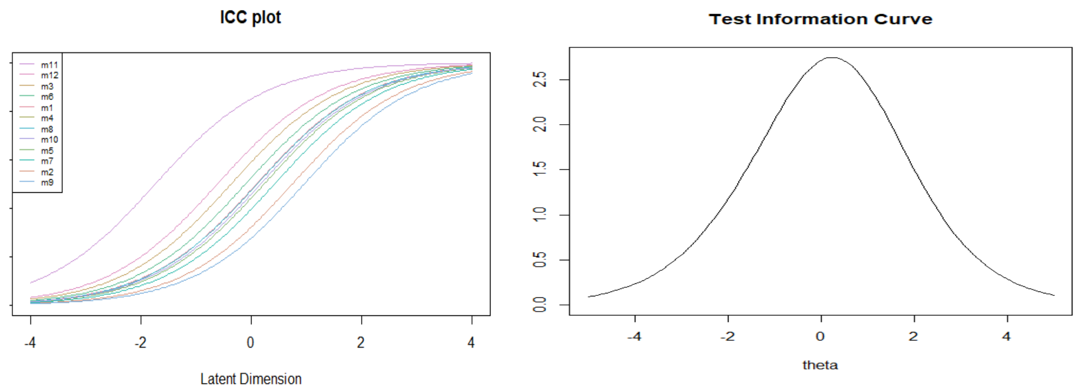
Table 3.

Results of the Rasch Analysis

Item	Difficulty value
I1	0.266
I2	0.899
I3	-0.198
I4	0.280
I5	0.405
I6	0.068
I7	0.591
I8	0.280
I9	1.101
I10	0.342
I11	-1.645
I12	-0.455
Model Summary:	
log.Lik -8167.34	
AIC 16361	
BIC 16426	

Figure 1.

ICC plot and Test Information Curve



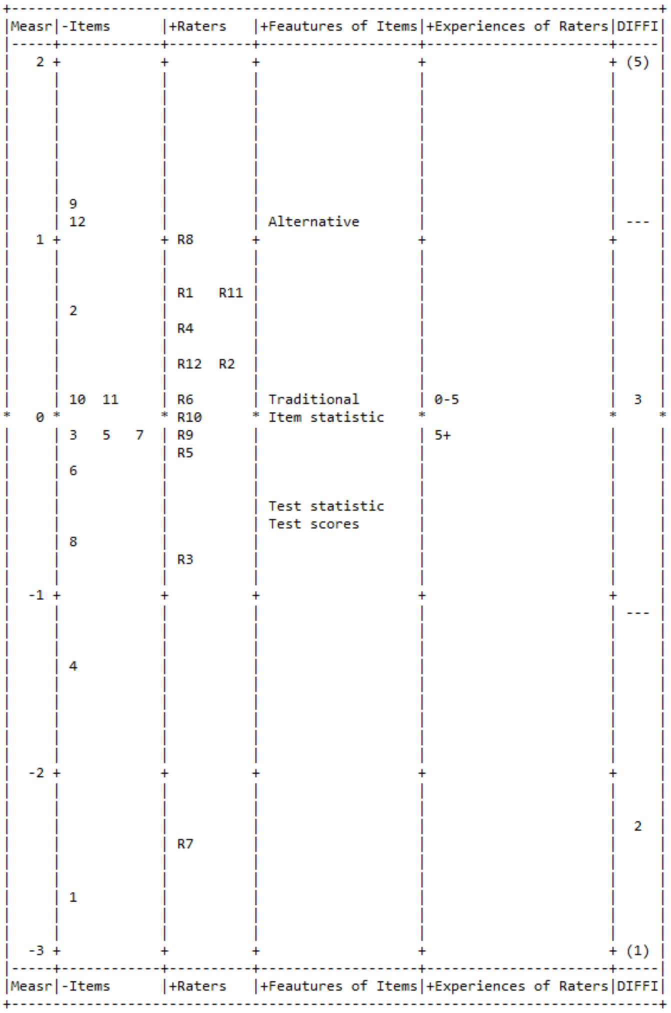
Prediction item difficulty (n=12 expert)

A multi-facet Rasch analysis was conducted to examine the effects of raters on the test scores. The four facets identified in this study were 12 items , 12 raters, four features of items (alternative,traditional, item statistic,test statistic, interpretation of test scores) and two experiences of raters (0-5, 5+ years). The item difficulty of 12 items are determined through the opinions of experts.12 measurement and evaluation experts estimated item difficulty by rating each one on a scale of 1 (very hard) to 5(very easy).

Figure 2 shows the distribution of items, raters, and item features on the same logit scale. The logit map gives general information about the facets and this measure allows for a comparability among variable sources in the study. In this distribution, the item facet is ranked from the most difficult to easiest, the rater facet is ranked from the most generous rater to the strictest rater and features of the item facet is ranked from the easiest subject to the most difficult, from top to bottom.The analysis revealed that I9 was the most difficult item and I1 was the easiest item. Among the raters, , R8 was the most lenient, while R7 was the strictest. Furthermore, items related to test scores and test statistics were found to be difficult, while items related to alternative topics were found to be easy. The multi-facet Rasch analysis is useful for examining the effects of raters on test scores, as it allows for the examination of multiple sources of variation and provides insights into the specific factors that affect the difficulty of test items.

Figure 2.

Logit Map of the Variables in the Study



```
+------------------------------------------------------------------------------+
|Measr|-Items      |+Raters   |+Feautures of Items|+Experiences of Raters|DIFFI|
|-----+------------+----------+-------------------+----------------------+-----|
|   2 +            +          +                   +                      + (5) |
|     |            |          |                   |                      |     |
|     |            |          |                   |                      |     |
|     |            |          |                   |                      |     |
|     |            |          |                   |                      |     |
|     |            |          |                   |                      |     |
|     |            |          |                   |                      |     |
|     |            |          |                   |                      |     |
|     | 9          |          |                   |                      |     |
|     | 12         |          | Alternative       |                      | --- |
|   1 +            + R8       +                   +                      +     |
|     |            |          |                   |                      |     |
|     |            |          |                   |                      |     |
|     |            | R1   R11 |                   |                      |     |
|     | 2          |          |                   |                      |     |
|     |            | R4       |                   |                      |     |
|     |            |          |                   |                      |     |
|     |            | R12  R2  |                   |                      |     |
|     |            |          |                   |                      |     |
|     | 10  11     | R6       | Traditional       | 0-5                  |  3  |
*   0 *            * R10      * Item statistic    *                      *     *
|     | 3   5   7  | R9       |                   | 5+                   |     |
|     |            | R5       |                   |                      |     |
|     | 6          |          |                   |                      |     |
|     |            |          |                   |                      |     |
|     |            |          | Test statistic    |                      |     |
|     |            |          | Test scores       |                      |     |
|     | 8          |          |                   |                      |     |
|     |            | R3       |                   |                      |     |
|     |            |          |                   |                      |     |
|  -1 +            +          +                   +                      +     |
|     |            |          |                   |                      | --- |
|     |            |          |                   |                      |     |
|     |            |          |                   |                      |     |
|     | 4          |          |                   |                      |     |
|     |            |          |                   |                      |     |
|     |            |          |                   |                      |     |
|     |            |          |                   |                      |     |
|     |            |          |                   |                      |     |
|     |            |          |                   |                      |     |
|  -2 +            +          +                   +                      +     |
|     |            |          |                   |                      |     |
|     |            |          |                   |                      |     |
|     |            |          |                   |                      |  2  |
|     |            | R7       |                   |                      |     |
|     |            |          |                   |                      |     |
|     |            |          |                   |                      |     |
|     | 1          |          |                   |                      |     |
|     |            |          |                   |                      |     |
|     |            |          |                   |                      |     |
|  -3 +            +          +                   +                      + (1) |
|-----+------------+----------+-------------------+----------------------+-----|
|Measr|-Items      |+Raters   |+Feautures of Items|+Experiences of Raters|DIFFI|
+------------------------------------------------------------------------------+
```

Measurement Report for Item

The measurement report obtained from the multi-faceted Rasch analysis for the item facet is presented in Table 4. It is observed that the items were differentiated in terms of difficulty/easiness, and the highest and lowest logit values were found to be 1.22 and -2.69, respectively. The reliability index obtained from the Rasch analysis was also acceptable with a value of 0.82. Furthermore, the separation index of 3.22 indicates that the items were significantly different in terms of difficulty. However, it is concerning that only one item, I2, did not meet the criteria for both internal and external consistency. It may be necessary to revise or remove this item from the test to improve its reliability and validity.

Table 4.

Measurement Report for Item

Item	Logit	Std.error	Infit		Outfit	
			MnSq	Zst	MnSq	Zst
I9	1.22	0.46	1.45	1.1	1.34	0.8
I12	1.08	0.46	0.55	-1.2	0.54	-1.2
I2	0.61	0.43	0.23	-2.8	0.23	-2.8
I10	0.10	0.44	0.71	-0.7	0.72	-0.6
I11	0.10	0.44	1.05	0.2	1.06	0.2
I5	-0.08	0.44	1.65	1.5	1.65	1.5
I7	-0.08	0.44	1.00	0.1	1.03	0.2
I3	-0.13	0.44	1.77	1.7	1.79	1.7
I6	-0.27	0.44	0.52	-1.3	0.55	-1.2
I8	-0.70	0.44	0.56	-1.1	0.55	-1.2
I4	-1.45	0.44	1.84	1.8	1.89	1.9
I1	-2.69	0.54	0.73	-0.6	0.74	-0.5
Mean	0.00	0.45	1.00	0.00	1.01	0.00
SD	1.07	0.03	0.55	1.5	0.55	1.5

Model, Sample: RMSE = .45 Standard deviation = .97
Discrimination ratio=2.17 Discrimination index = 3.22
Discrimination index of reliability= 0.82
Model, Fixed (all same) chi square=52.9 df =11 p= .00
Model, Random (normal) chi square =9.3 df = 10 p= .50

Measurement Report for Rater

The measurement report resulting from the multi-facet Rasch analysis of the rater facet is displayed in Table 5.The estimated separation ratio, separation index, and separation index reliability for the scoring facet are found to be high. The R8-coded rater is found to be the most generous, while the R7-coded rater is the strictest. When the separation index of 2.71 and the reliability coefficient of 0.76 are evaluated together with the chi-square test result ($\chi 2(df)=41.9(11)$, $p=.00$) for the fixed effect, it is determined that there is a significant difference among the raters who score the item difficulty of the items in terms of their strictness/generosity. The very low agreement index between the raters (-0.001) indicates that there is no agreement among the raters.

Table 5.

Measurement Report for Rater

Rater	Logit	Std.error	Infit		Outfit	
			MnSq	Zst	MnSq	Zst
R8	1,03	0.44	1.93	1.9	1.87	1.8
R1	0,74	0.44	1.29	0.7	1.32	0.8
R11	0,74	0.44	0.52	-1.3	0.53	-1.3
R4	0,55	0.44	1.21	0.6	1.26	0.7
R2	0,25	0.44	0.67	-0.8	0.66	-0.8
R12	0,25	0.44	0.36	-2.0	0.37	-2.0
R6	0,06	0.44	0.82	-0.3	0.81	-0.3
R10	-0,03	0.44	0.96	0.0	0.96	0.0
R9	-0,13	0.44	0.40	-1.8	0.41	-1.8
R5	-0,23	0.44	1.26	0.7	1.24	0.7
R3	-0,81	0.44	1.09	0.3	1.04	0.2
R7	-2,41	0.48	1.72	1.6	1.63	1.4
Mean	0.00	0.44	1.02	0.00	1.01	0.00
SD	0.91	0.01	0.50	1.3	0.48	1.3

Table 6.

Measurement Report for Rater (Continued)

Model, Sample: RMSE = .44 Standard deviation = .76
Discrimination ratio=1.79 Discrimination index = 2.71
Discrimination index of reliability= 0.76
Model, Fixed (all same) chi square=41.9 df =11 *p*= .00
Model, Random (normal) chi square =9.00 df = 10 *p*= .53
Observed inter-rater agreement: 36.9 %
Expected inter-rater agreement: 37.1%
Kappa inter-rater reliability statistics: -0.001

Measurement Report for Sub-test of the Items

The measurement report obtained through multi-faceted Rasch analysis for the *Features of Items* facet is given in Table 7. It is observed that the separation ratio, separation index, and separation index reliability calculated for item characteristics are high. Accordingly, a significant difference was found in the item difficulty of items based on their characteristics (χ2(df)=18.1(4), p=0.00). A negative logit value indicates a low (difficult) score, while a positive logit value indicates a high (easy) score. Accordingly, items related to Interpretation of Test Scores and Test Statistic were found to be difficult, whereas items related to alternative topics were found to be easy.

Table 7.

Features of Items Measurement Report

Item	Logit	Std.error	Infit		Outfit	
			MnSq	Zst	MnSq	Zst
Alternative	1.07	0.34	1.28	1.0	1.19	0.7
Traditional	0.08	0.25	1.28	1.1	1.30	1.2
Item statistic	-0.05	0.30	0.76	-0.8	0.79	-0.7
Test statistic	-0.49	0.31	0.98	0.0	0.94	-0.1
Interpretation of Test Scores	-0.62	0.26	0.78	-0.9	0.77	-1.0
Mean	0.0	0.29	1.01	0.1	1.0	0.0
SD	0.67	0.04	0.26	1.0	0.24	1.0

Model, Sample: RMSE = .29 Standard deviation = .52
Discrimination ratio=2.03 Discrimination index = 3.04
Discrimination index of reliability= 0.80
Model, Fixed (all same) chi square=18.1 df =4 p= .00
Model, Random (normal) chi square =3.3 df = 3 p= .35

Measurement Report for Rater’s Experiment

The measurement report obtained through multi-faceted Rasch analysis for the rater’s experiment facet is given in Table 8. According to the analysis results in Table 8, which evaluates the item difficulty of the items, there was no differentiation according to the experience of the raters, as the discrimination index was 0.99 and the reliability coefficient was 0.19 with a chi-square test result of (χ2(df)=1.2(1), p=.27).

Table 8.

Measurement Report for Rater's Experiment

Item	Logit	Std.error	Infit		Outfit	
			MnSq	Zst	MnSq	Zst
0-5	0.14	0.18	0.96	-0.1	1.06	-0.2
5+	-0.14	0.18	1.06	0.3	0.94	0.4
Mean	0.00	0.45	1.01	0.00	1.01	0.1
SD	0.2	0.00	0.07	0.4	0.07	0.4

Model, Sample: RMSE = .18 Standard deviation = .00
Discrimination ratio=0.49 Discrimination index = 0.99
Discrimination index of reliability= 0.19
Model, Fixed (all same) chi square=1.2 df =1 p= .27

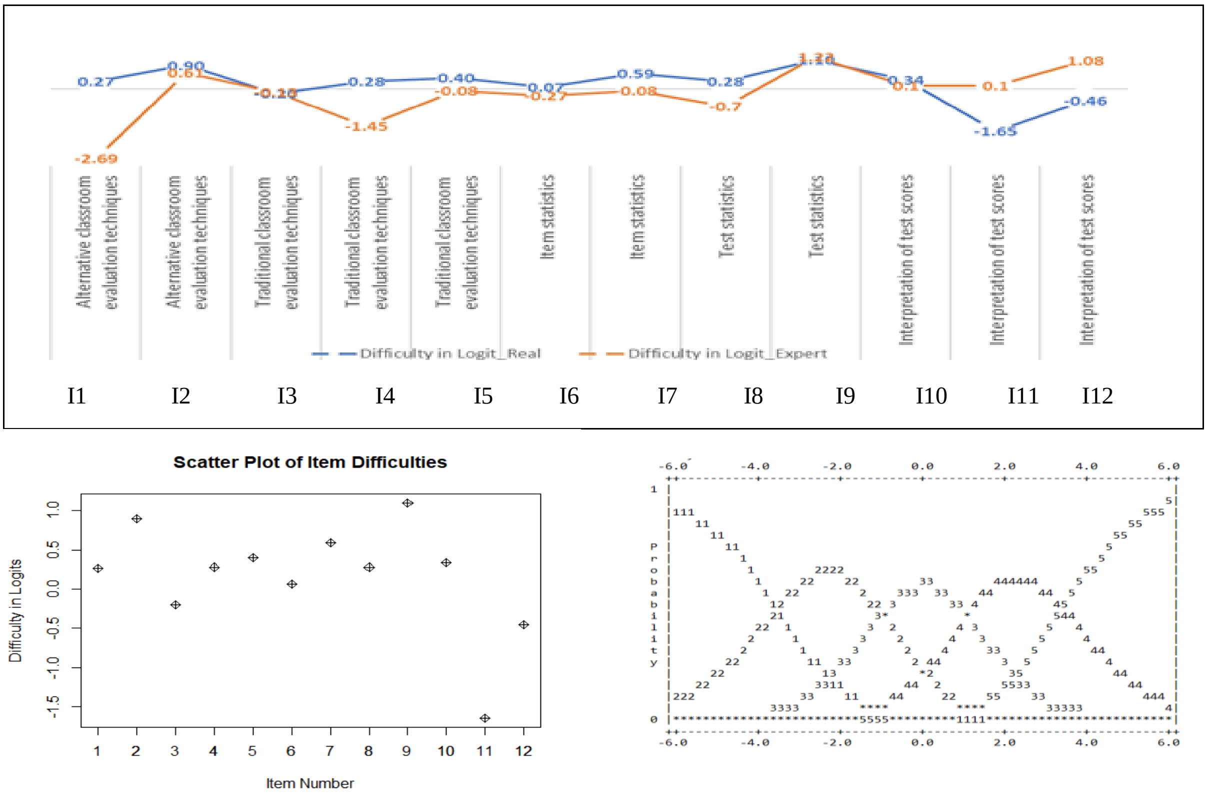
Compared real and prediction item difficulty

In accordance with the results gathered from 1165 students, the difficulties in logit values of the items in the measurement and evaluation test, which consisted of 12 items, were calculated using Rasch (Figure 3). Multi-facet Rasch analysis was used to estimate the difficulty in logit values for the estimations provided by 12 experts who participated in the study (Figure 3).

The experts predicted that the items were easier, except for the "Interpretation of test scores" subtest, where they estimated that 2 out of 3 items were actually more difficult.

Figure 3.

Estimated and Real Item Difficulty in Logit



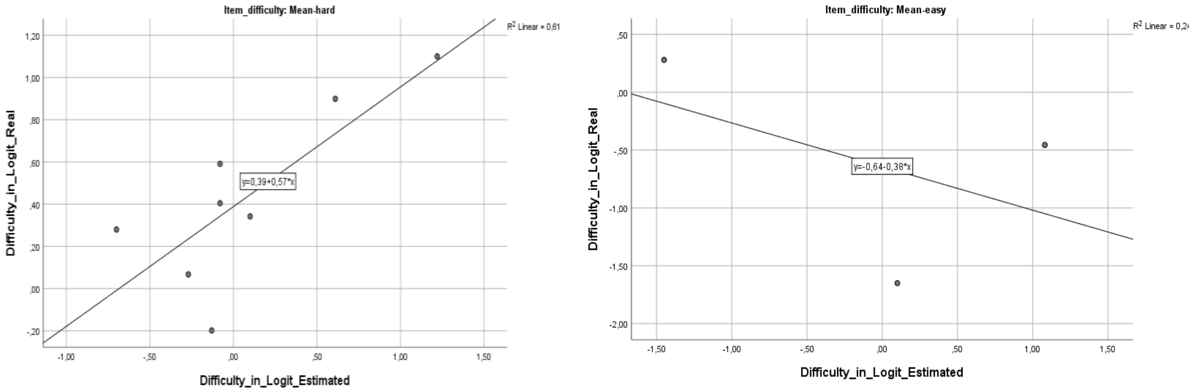
Difficulty in Logit_Real

Difficulty in Logit_Estimated

To examine the relationship between the experts' estimates and the real item difficulty index in detail, items with an item difficulty parameter close to or above 0.00 (moderate and hard) and those with a negative value (easy) were studied separately. These comparisons of the difficulty in logit values are presented in Figure 4. As can be seen in Figure 4, while experts made better predictions for moderate items, their difficulty predictions for easy items were not as accurate as for moderate or hard items.

Figure 4.

Comparing the Estimated and Real Item Difficulty in Logit (moderate-hard and easy items)



Discussion

Due to security reasons, some high-stakes tests are not pre-tested, and the item difficulty in-class assessments is estimated by teachers. While there has been increasing interest in using new methods to predict item difficulty, these methods all require the inclusion of predictor variables in the models they build, and the predictors are identified and represented as rules by experts. Furthermore, the difficulty model created for one test may not be applicable to other tests. It is crucial to identify the relevant predictors to create accurate difficulty models and to assess the reliability of the experts' difficulty assessments. The study aimed to compare the experts' estimated item difficulty with the real values based on the data of the high stakes test like the PPSE in Turkey. The present research was to evaluate the accuracy of the experts' difficulty assessments by creating a high-stakes test that resembles the certification exam, and then comparing the results to their estimated item difficulty. The teacher certification exam for public institutions consists of 12 assessment and evaluation-related items. The item difficulty of these items and the overall exam are determined through the opinions of experts.

The results of this study suggested that experts in the field of assessment and evaluation had some bias in predicting item difficulty. Previous studies by Enright et al. (1993) and Wauters et al. (2012) demonstrated a strong positive correlation between expert ratings from science educators and correct rate in forecasting item difficulty. Moreover, they found that there was no significant difference between expert ratings and true value comparisons. However, Lumley et al. (2012) found that experienced experts were able to consistently predict item difficulty on reading tests. Sydorenko (2011) suggested that experts' ability to predict item difficulties may vary depending on the type of test and the specific items assessed. Furthermore, Kibble and Johnson (2011) reported a statistically significant but weak correlation between the intended difficulty of test items and actual student scores.

In our study, it was seen that the experts were adequate in estimating the medium item difficulty. Le Hebel et al. (2019) analyzed the difficulty of science inquiry tasks based on both estimated and real values in relation to students' abilities. The study also examined how accurately teachers predict students' difficulty in answering Programme for International Student Assesment-[PISA]science questions. Like the previous study, Hamamoto Filho et al. (2020) found that the panel of experts' estimates of item difficulty had 54% correlation with real item difficulty. The study also found that items expected to be easy had significantly lower average difficulty than items expected to be moderate or difficult.

The average score of the students was calculated to be 46 out of 100, indicating that the test generally was of mean difficulty, while the experts estimated that the mean difficulty would be 51. The study found that the experts underestimated the difficulty of the test, with a particular bias towards underestimating items that were easy. Urhahne and Wijnia (2021) reviewed 10 studies that investigated the comparison between teachers' perceived and real difficulty of tasks. In 8 out of 10 studies, it was discovered that teachers often wrongly believed tasks were less challenging than they were, or expected students to perform better than they did. The studies were carried out in various academic subjects, including math, science, language arts, and a combination of language arts and science (Urhahne & Wijnia, 2021). Schult and Lindner (2018) found that teachers tend to underestimate the difficulty of items that require written answers.

The results of the study indicated that the accuracy of the experts' predictions varied across the subtests. The experts believed that the items in the test, excluding the "Interpretation of Test Scores" subtest, would be easier for the students. The accuracy of the experts in predicting the item difficulty in this subtest was lower compared to other topics. However, their predictions for the items in the "Item Statistics" subtopic were found to be more accurate.

The study's results showed that there were differences among experts' estimates of generosity-stinginess, but this variance was not associated with their years of experience. Thus, there is potential for improving the methodology and training used by experts to predict the item difficulty. Wauters et al. (2012) indicated that the inter-rater agreement for the estimation of the item difficulty by experts was good, with an ICC (Intraclass Correlation Coefficient) value of 0.68 for expert rating and for one-to-many comparison. Similarly, Attali et al. (2014) discovered that there was little variability in the quality of judgments across content areas and raters. This means that even new item writers who are not familiar with the items and not exposed to item statistics can perform similarly to more experienced SAT raters. The study implies that the ability to differentiate between the difficulties of the items is less related to test development experience and more linked to the specific difficulty scale used. While experts can assess the item difficulty, there can be variations in their evaluations, indicating room for improvement in their training and methodology. This information can aid in developing effective training programs for item writers and raters involved in test development.

In conclusion, the study highlights the importance of accurately predicting the item difficulty to ensure a fair and valid assessment of student performance. The findings suggest that further research is needed to improve the accuracy of expert estimations of item difficulty in high-stakes tests. The results of this study can be used to improve the accuracy of expert predictions and to refine the methods used for estimating item difficulty in the future. It also suggests the need for more objective and consistent methods that need attention to determine the predictors for predicting item difficulty, such as machine learning and item response theory, which can provide more reliable and accurate estimates of test difficulty. The results of this study can inform future research on item difficulty prediction and help improve the accuracy of expert opinions. It also indicates the need to create an item difficulty guide for item writers and moderators.

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 Gazi University dated 23.05.2023 and numbered E-77082166-604.01.02-673852.

References

Afrashteh, M. Y. (2021). Comparison of the validity of bookmark and Angoff standard setting methods in medical performance tests. *Bmc Medical Education*, *21*(1). https://doi.org/10.1186/s12909-020-02436-3

AITSL, A. I. f. T. a. S. L. (2022). *AITSL, Australian Professional Standards for Teachers*. https://www.aitsl.edu.au/tools-resources/resource/australian-professional-standards-for-teachers

Attali, Y., Saldivia, L., Jackson, C., Schuppan, F., & Wanamaker, W. (2014). Estimating item difficulty with comparative judgments. *ETS Research Report Series*, *2014*(2), 1-8. http://dx.doi.org/10.1002/ets2.12042

Beinborn, L., Zesch, T., & Gurevych, I. (2014). Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics*, *2*, 517-530. https://doi.org/10.1162/tacl_a_00200

Chon, Y. V., & Shin, T. (2010). Item difficulty predictors of a multiple-choice reading test. *English Teaching*, *65*(4), 257-282. http://journal.kate.or.kr/wp-content/uploads/2015/02/kate_65_4_11.pdf

Clauser, J. C., Hambleton, R. K., & Baldwin, P. (2017). The Effect of Rating Unfamiliar Items on Angoff Passing Scores.*Educational and Psychological Measurement*, *77*(6), 901-916. https://doi.org/10.1177/0013164416670983

Ell, F. (2021). Teacher education policy in Aotearoa New Zealand: Global trends meet local imperatives. In *Teacher Education Policy and Research: Global Perspectives* (pp. 113-128). Springer.

Enright, M. K., Allen, N., & Kim, M. I. (1993). A Complexity Analysis of Items from a Survey of Academic Achievement in the Life Sciences. *ETS Research Report Series*, *1993*(1), i-32. https://files.eric.ed.gov/fulltext/ED385595.pdf

Fergadiotis, G., Swiderski, A., & Hula, W. D. (2019). Predicting confrontation naming item difficulty. *Aphasiology*, *33*(6), 689-709. https://doi.org/10.1080/02687038.2018.1495310

Gitomer, D. H., & Qi, Y. (2010). Recent Trends in Mean Scores and Characteristics of Test-Takers on" Praxis II" Licensure Tests. *Office of Planning, Evaluation and Policy Development, US Department of Education*.

Grivokostopoulou, F., Hatzilygeroudis, I., & Perikos, I. (2014). Teaching assistance and automatic difficulty estimation in converting first order logic to clause form. *Artificial Intelligence Review*, *42*, 347-367. http://dx.doi.org/10.1007/s10462-013-9417-8

Hamamoto Filho, P. T., Silva, E., Ribeiro, Z. M. T., Hafner, M. d. L. M. B., Cecilio-Fernandes, D., & Bicudo, A. M. (2020). Relationships between Bloom's taxonomy, judges' estimation of item difficulty and psychometric properties of items from a progress test: a prospective observational study. *Sao Paulo Medical Journal*, *138*, 33-39. http://dx.doi.org/10.1590/1516-3180.2019.0459.R1.19112019

He, J., Peng, L., Sun, B., Yu, L. J., & Zhang, Y. H. (2021). Automatically Predict Question Difficulty for Reading Comprehension Exercises. *2021 Ieee 33rd International Conference on Tools with Artificial Intelligence (Ictai 2021)*, 1398-1402. https://doi.org/10.1109/Ictai52525.2021.00222

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, *35*(1), 69-81. https://psycnet.apa.org/doi/10.1111/j.1745-3984.1998.tb00528.x

Kardong-Edgren, S., & Mulcock, P. M. (2016). Angoff Method of Setting Cut Scores for High-Stakes Testing Foley Catheter Checkoff as an Exemplar. *Nurse Educator*, *41*(2), 80-82. https://doi.org/10.1097/Nne.0000000000000218

Kibble, J. D., & Johnson, T. (2011). Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations? *Advances in physiology education*, *35*(4), 396-401. https://doi.org/10.1152/advan.00062.2011

Kurdi, G., Leo, J., Matentzoglu, N., Parsia, B., Sattler, U., Forge, S., Donato, G., & Dowling, W. (2021). A comparative study of methods for a priori prediction of MCQ difficulty. *Semantic Web*, *12*(3), 449-465. https://doi.org/10.3233/Sw-200390

Le Hebel, F., Tiberghien, A., Montpied, P., & Fontanieu, V. (2019). Teacher prediction of student difficulties while solving a science inquiry task: example of PISA science items. *International Journal of Science Education*, *41*(11), 1517-1540. https://doi.org/10.1080/09500693.2019.1615150

Lin, C.-S., Lu, Y.-L., & Lien, C.-J. (2021). Association between Test Item's Length, Difficulty, and Students' Perceptions: Machine Learning in Schools' Term Examinations. *Universal Journal of Educational Research*, *9*(6), 1323-1332. http://dx.doi.org/10.13189/ujer.2021.090622

Lin, L. H., Chang, T. H., & Hsu, F. Y. (2019). Automated Prediction of Item Difficulty in Reading Comprehension Using Long Short-Term Memory. *Proceedings of the 2019 International Conference on*

Asian Language Processing (IALP), Shanghai, China, 132-135. https://doi.org/10.1109/IALP48816.2019.9037716.

Linacre, J.M. (2014). A user's guide to FACETS Rasch-model computer programs. Retrieved from http://www.winsteps.com/a/facets-manual.pdf

Lumley, T., Routitsky, A., Mendelovits, J., & Ramalingam, D. (2012). A framework for predicting item difficulty in reading tests.

OSYM. (2022). *KPSS: Kamu Personel Seçme Sınavı*. https://www.osym.gov.tr/TR,23892/2022-kpss-lisans-genel-yetenek-genel-kultur-ve-egitim-bilimleri-oturumlarinin-temel-soru-kitapciklari-ve-cevap-anahtarlari-yayimlandi-31072022.html

Pandarova, I., Schmidt, T., Hartig, J., Boubekki, A., Jones, R. D., & Brefeld, U. (2019). Predicting the difficulty of exercise items for dynamic difficulty adaptation in adaptive language tutoring. *International Journal of Artificial Intelligence in Education*, *29*, 342-367. https://doi.org/10.1007/s40593-019-00180-4

Perikos, I., Grivokostopoulou, F., Kovas, K., & Hatzilygeroudis, I. (2016). Automatic estimation of exercises' item difficulty in a tutoring system for teaching the conversion of natural language into first-order logic. *Expert Systems*, *33*(6), 569-580. https://doi.org/10.1111/exsy.12182

Perkins, K., Gupta, L., & Tammana, R. (1995). Predicting item difficulty in a reading comprehension test with an artificial neural network. *Language testing*, *12*(1), 34-53.https://doi.org/10.1177/026553229501200103

Praxis, E. T. S. (2022). *ETS, The Praxis Tests*. https://www.ets.org/praxis

Qiu, Z. P., Wu, X., & Fan, W. (2019). Question difficulty prediction for multiple choice problems in medical exams. *Proceedings of the 28th Acm International Conference on Information & Knowledge Management (Cikm '19)*, 139-148. https://doi.org/10.1145/3357384.3358013

Sano, M. (2015). Automated capturing of psycho-linguistic features in reading assessment text. *Annual meeting of the National Council on Measurement in Education*, Chicago, IL,

Schult, J., & Lindner, M. A. (2018). Judgment Accuracy of German Elementary School Teachers: A Matter of Response Formats? *German Journal of Educational Psychology, 32*(1-2), 75-87. https://doi.org/10.1024/1010-0652/a000216

Stadler, M., Niepel, C., & Greiff, S. (2016). Easily too difficult: Estimating item difficulty in computer simulated microworlds. *Computers in Human Behavior*, *65*, 100-106. https://doi.org/10.1016/j.chb.2016.08.025

Sydorenko, T. (2011). Item writer judgments of item difficulty versus real item difficulty: A case study. *Language Assessment Quarterly*, *8*(1), 34-52. https://doi.org/10.1080/15434303.2010.536924

Toyama, Y. (2021). What makes reading difficult? An Investigation of the contributions of passage, task, and reader characteristics on comprehension performance. *Reading Research Quarterly*, *56*(4), 633-642. https://doi.org/10.1002/rrq.440

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review*, *32*, 100374. https://doi.org/10.1016/j.edurev.2020.100374

Wauters, K., Desmet, P., & Van Den Noortgate, W. (2012). Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education*, *58*(4), 1183-1193. https://doi.org/10.1016/j.compedu.2011.11.020

Wyse, A. E. (2018). Equating angoff standard-setting ratings with the rasch model.*Measurement-Interdisciplinary Research and Perspectives*, *16*(3), 181-194. https://doi.org/10.1080/15366367.2018.1483170

Wyse, A. E. (2020). Comparing cut scores from the angoff method and two variations of the hofstee and beuk methods. *Applied Measurement in Education*, *33*(2), 159-173. https://doi.org/10.1080/08957347.2020.1732385

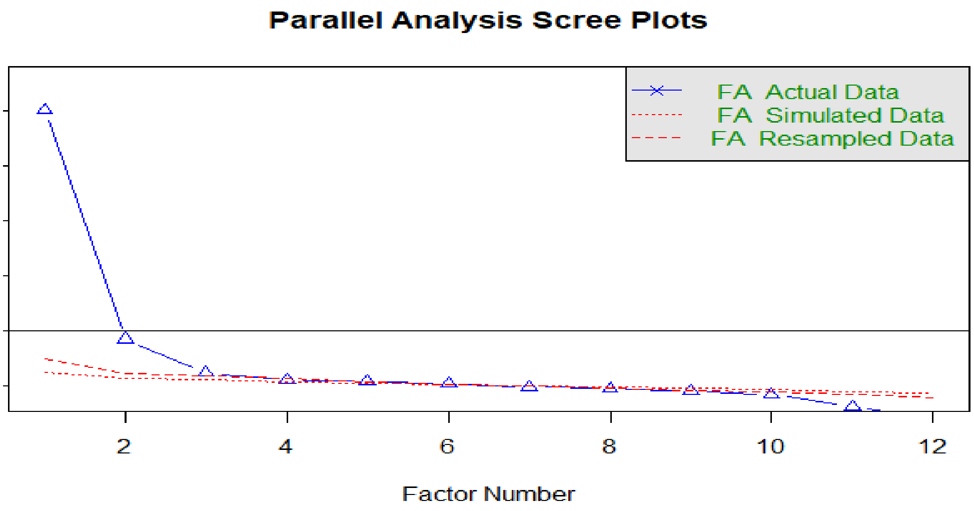
Yaneva, V., Ha, L. A., Baldwin, P., & Mee, J. (2020, May). Predicting item survival for multiple choice questions in a high-stakes medical exam. *Proceedings of the 12th International Conference on Language Resources and Evaluation (Lrec)*, 6812-6818. Marseille, France. https://aclanthology.org/2020.lrec-1.841.pdf

Yim, M. K., & Shin, S. J. (2020). Using the Angoff method to set a standard on mock exams for the Korean Nursing Licensing Examination. *Journal of Educational Evaluation for Health Professions*, *17*(4). https://doi.org/10.3352/jeehp.2020.17.14

Appendix

Appendix A.

Parallel Analysis Scree Plots



Appendix B.

Results of the EFA

Item	Factor loadings
I1	0.579
I2	0.621
I3	0.532
I4	0.604
I5	0.590
I6	0.623
I7	0.639
I8	0.859
I9	0.691
I10	0.792
I11	0.647
I12	0.484
Variance	%42
α	0.80

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(4); 455-471

The Impact of Item Preknowledge on Scaling and Equating: Item Response Theory True and Observed Score Equating Methods

Çiğdem AKIN ARIKAN*   Allan S. COHEN**

Abstract

Testing programs often reuse items due mainly to the difficulty and expense of creating new items. This poses potential problems to test item security if some or all test-takers have knowledge of the items prior to taking the test. In this study, simulated data are used to assess the effect of preknowledge on item response theory true and observed score equating. Root mean square error and bias were used for the recovery of equated scores and linking coefficients for scaling methods. The results of this study indicated that item preknowledge has a large effect on equated scores and linking coefficients. Furthermore, as the mean ability distribution of the group difference, the number of exposed items, and the number of examinees with item preknowledge increase, the bias and RMSE for equated scores and linking coefficients also increase. Additionally, IRT true score equating results in a higher bias and RMSE than IRT observed score equating. These findings suggest that item preknowledge has the potential to inflate equated scores, putting the validity of the test scores at risk.

Keywords: cheating, item preknowledge, equating, RMSE, bias

Introduction

Testing programs often reuse items due to the difficulty and expense of creating new items. When items are reused, the security of those items poses a potential problem. Major testing organizations such as the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) define test security as “protection of content of a test from unauthorized release or use, to protect the integrity of the test scores so they are valid for their intended use” (AERA/APA/NCME, 2014, p. 236). When the security of test items is violated through preknowledge by some or all individuals taking the test, the results do not provide a valid indication of the knowledge or ability of test-takers.

Preknowledge is a form of cheating (Cizek & Wollack, 2017; Lee, 2018). Test cheating reduces the reliability and validity of the test scores (Man et al., 2019). Researchers described cheating as an ethical error (Fly, 1995) and defined it as any actions that breaches the rules of tests (Cizek, 1999). Cheating among students has increased in part due to the increased speed and straightforwardness of communication. For example, as many as 95% of university students and 53% to 60% of high school students admit to having cheated on at least one test during their educational career (Josephson Institute, 2012; Wang et al., 2015). Reports showed that some teacher candidates used systematic cheating on teacher selection exams in Turkey in 2010 and 2011 (Demir & Arcagok, 2013). What is clear from examples such as these is that cheating is a problem at all educational levels. Because of the potentially serious impact of cheating on test results, its detection is critical.

Item preknowledge occurs when examinees obtain access to test items or their answers before taking the test (Foster, 2013; Gorney & Wollack, 2022). Cizek and Wollack (2017) argue, if any cheating occurs, the resulting test scores might not accurately reflect the actual knowledge of the individuals who cheat. This means that the test may need to measure the skill or ability being assessed accurately. In

* Assoc. Prof. Dr., Ordu University, Faculty of Education, Ordu-Türkiye, akincgdm@gmail.com, ORCID ID: 0000-0001-5255-8792
** Prof. Dr., University of Georgia, College of Education, Athens-USA, acohen@uga.edu, ORCID ID: 0000-0002-8776-9378

To cite this article:
Akın-Arıkan, Ç. & Cohen, A. (2023). The impact of item preknowledge on scaling and equating: Item response theory true and observed score equating methods. *Journal of Measurement and Evaluation in Education and Psychology, 14*(4), 455-471. https://doi.org/10.21031/epod.1199296

Received: 4.11.2022
Accepted: 25.10.2023

such a case, the comparability of scores across individuals and the validity of the test would be threatened (AERA et al., 2014; Lee, 2018), as exceptionally high scores might be due to prior knowledge rather than to the ability and preparedness of the examinee (Qian et al., 2016). In addition, preknowledge affects the interpretation of the test results of all individuals, not just those who had the unfair advantage of preknowledge. Further, it is the responsibility of test administrators to establish that no test-taker has an unfair advantage over others.

Using different test forms can help mitigate the effects of preknowledge, but it is then necessary to confirm that the different forms are of equal difficulty; this ensures that examinees are indifferent to which form they receive (Lord, 1980). The psychometric approach to ensuring different forms of a test are of comparable difficulty is to place these forms on the same score scale (Kolen & Brennan, 2014; von Davier et al., 2004). In that way, the scores of each form have the same interpretation.

Methods for placing test forms on the same scale are referred to as equating. If the test forms are to be equated, the first step is to decide on the equating design. In IRT-based equating, there are different designs, including random group, single group, and the non-equivalent groups with anchor item (NEAT) design. In this study, we investigate the effects of item preknowledge on equating results using the NEAT design. This design is flexible and can be used to equate multiple different test forms to a common scale. For simplicity, however, we present the following discussion in terms of equating two different forms of a test.

In the NEAT design, two groups of examinees each take one unique test form and one anchor test, in which the anchor test is the same for both groups. This anchor test is used to link the test forms to each other (Kolen & Brennan, 2014). One concern with using two different forms of a test is that the two groups of examinees may sometimes sit the exam at different times. In such a case, it is possible that information about some or all of the items on the test from the first testing group may be passed on to individuals in the other group. This creates a potential preknowledge situation for the examinees from the second group who receive the information. It is important to note that items in the anchor test are not typically identifiable as being on the anchor test. That is, examinees in the first group may not know which items are on the anchor test and which items are unique to the first test form, such that complete information may not necessarily be available to examinees from the other group. If some of the items passed along by people in the first group are anchor items, however, it could result in some members of the second group having inflated test scores. The scores of examinees with preknowledge would likely be inflated and would not accurately reflect the true abilities of these test-takers. Tan (2001) states that the results obtained from tests with individuals with preknowledge will not be valid for score-based decisions. Thus, item preknowledge can pose a serious problem in test security for test developers, test administrators, and users of test results (Pan & Wollack, 2021).

Item Response Theory

Item response theory (IRT) models are commonly utilized for test analysis and scoring (González & Wiberg, 2017). IRT models can be used for dichotomous and polytomous items. In this study, we focused on equating dichotomously scored items. IRT is used to obtain estimates of item and ability parameters. The Rasch model, one-parameter logistic (1PL) model, two-parameter logistic (2 PL) model, and three-parameter logistic (3 PL) model are among the IRT models frequently employed for analyzing dichotomously scored data. The 2 PL model, which is the model used in this study, is a generalized form of the 1PL model: the 1PL model has only item difficulty parameters, while the 2 PL model also includes item discrimination parameters. The 2 PL model takes the following mathematical form:

$$P(X_i = 1|\theta) = \frac{\exp[a_i(\theta - b_i)]}{1+\exp[a_i(\theta - b_i)]} \qquad (1)$$

where θ is the individual's level of ability, a_i is the item discrimination parameter for item i, and b_i is the item difficulty parameter for item i (de Ayala, 2009).

Item Response Theory (IRT) Based Equating

In large-scale test applications, including PISA and TIMSS etc., different test forms of similar content and difficulty are generally used. Using different test forms on various dates raises concerns about potential differences in difficulty. To address this, equating is employed to make scores interchangeable between test forms, ensuring their interchangeability (Kolen & Brennan, 2014). Equating methods can be classified as traditional, kernel, local, and IRT based (which is used in the present study) equating methods (González & Wiberg, 2017).

IRT-based equating methods are classified into the following two categories: true score equating and observed score equating. In IRT, equating takes place in three stages. These steps are, respectively, estimation of item parameters, calibration, and equating the test scores. Examinees who take different test forms are not considered equivalent, and the parameters of the test forms are not represented on the same IRT scale (Kolen & Brennan, 2014). Therefore, once the item parameters have been estimated with the appropriate IRT model, the item and ability parameters can be estimated using separate or concurrent calibration, which is the first stage in the test equating. The calibration of IRT scales aims to link the new and old forms together. Through the concurrent calibration, the parameters of test forms can be estimated together, and common items are assumed to have the same parameter values in both test forms. Separate estimating methods based on test characteristics curves were shown to be the most reliable in practice (Kolen & Brennan, 2014). As a result, separate calibration methods were used in the present study.

In NEAT, items are in common from one test to the other, which allows for test forms to be linked to a common scale. However, parameter estimates from different test forms may not be on the same scale, for which a linear transformation should be performed (González & Wiberg, 2017; Kolen & Brennan, 2014). The one test is chosen as the base scale, and then the common items are used to place item parameter estimates, examinee ability estimates, and estimated ability distributions on the base scale using separate calibration methods: mean/mean (MM), mean/sigma (MS), and characteristic curve methods that are Haebara (HB) and Stocking Lord (SL). The characteristic curve methods give more consistent results for dichotomous IRT models than the mean/sigma and the mean/mean methods (Kolen & Brennan, 2014). The MS and SL methods were used in this study. The MS method is preferred because it might be easily influenced by variations in item strength, whereas the SL method is preferred because it gives more consistent results. The separate calibration can be done in the NEAT design using orthogonal regression (e.g., Kane & Mroch, 2020). For this purpose, the linking coefficients of the regression, *A* (slope) and *B* (intercept), are used. The parameters of the anchor items are used for transforming the θ-scale of form X (new form-target) to the θ-scale of form Y (old form-base form). Typically, raw scores on the new form are equated to raw scores on the old form (Kolen & Brennan, 2014). Following calibration of the items, the resulting item and ability parameter estimates are used in the equating.

The IRT true score equating method is used to link the number of correct scores on the two forms. It is done by assuming that a given θ-related true score obtained with the base scale form is equivalent to the true score of the θ in the new form. IRT observed score equating, on the other hand, uses the observed-score distributions of the two test forms obtained using the given IRT model (Han et al., 1997). These are weighted for the two distributions using equipercentile equating in IRT observed score equating (Kolen & Brennan, 2014).

There are several key differences between these two equating methods. First, IRT observed score equating specifies the equating relationship for the observed scores, while IRT true score equating uses the true scores for equating (although these are not available in practice) (Kolen & Brennan, 2014). Additionally, IRT observed score equating is sample dependent, while IRT true score equating is sample invariant (Cook & Eignor, 1991; Han et al., 1997). However, IRT true score and IRT observed score equating methods are comparable in terms of errors.

Purpose of Study

The purpose of tests is to make valid decisions about individuals in accordance with a specific aim. In order to do this, tests are expected to reflect the true ability of the individuals accurately. In other words, it is expected that highly talented individuals will score well while less talented individuals will receive lower scores. However, if individuals who take the test also have preknowledge about one or more items, they will likely score higher on those items, reducing the validity of the test (Eckerly, 2017). In IRT, the probability of answering exposed items correctly decreases as the item difficulty increases (Zimmermann et al., 2016). It can reasonably be expected that the item discrimination parameters will also change. Furthermore, the ability estimates for individuals also change with this change in item parameters, and an increasingly negative effect on the performance of honest individuals will be observed relative to the performance of individuals who cheat.

Equating can be employed to correct for differences in test form difficulty. As noted above, preknowledge among some test takers can result in corruption of the equating of form difficulties. However, it is a concern that usual equating methods do not consider item preknowledge. Thus, standard equating methods used to correct group ability differences may exacerbate the inaccuracy of the equating. It is likely that the scores obtained from the equating of tests with preknowledge among some test takers will not accurately correct for form difficulties. Therefore, it is important to determine how the presence of exposed items affects the equating.

IRT equating is a useful methodology and has been used in test construction by several testing programs and companies (Skaggs & Lissitz, 1986). However, few studies have been presented on the effects of exposed items on test equating (Barri, 2013; Jurich et al., 2010; Jurich, 2011). It would be useful, therefore, to investigate the extent to which errors in test equating change as a result of preknowledge. Barri (2013) analyzed the impact of exposed anchor items on the equated scores obtained under Rasch IRT true score equating. As the number of exposed items increased, Barri found that test scores exhibited inflated results. Similarly, Jurich (2011) investigated the effects of cheating on equated scores using 3PL true score equating for five equating methods, including the SL approach, the MM, MS method, the HB method, and the fixed anchor method. Results indicated that cheating artificially equated scores for all five methods. More recently, Liu and Becker (2022) studied the impact of item exposure and preknowledge on 1PL model pre-equated item difficulty and ability estimates. Results showed that item exposure had a significant impact on item difficulty for exposed and nonexposed items. In the previous studies (e.g., Barri, 2013; Liu & Becker, 2022) examining the effect of item preknowledge on test equating, it was seen that the 2 PL model, which is also used in large-scale test applications, hasn’t been used. In addition, IRT observed score equating methods hasn’t been used. In our study, we build on the previous research knowledge base by investigating the effects of preknowledge on IRT true score and IRT observed score equating methods with MS and SL. More specifically, this study aims to examine the impact of the exposure of anchor items with the 2 PL model on IRT true score and IRT observed equating under the NEAT design.

Method

A Monte Carlo simulation study was conducted to investigate the impact of exposure of anchor items. Results were compared for IRT true score equating and IRT observed score equating.

Simulation Conditions

We investigated the effects of three conditions on equating errors: ability distribution, exposed anchor items, and the proportion of examinees with pre-knowledge. The sample size is 2000, the test length is 40, and both variables were handled as constants. The simulated data was equated using the NEAT design. NEAT, the most widely used equating design, is used in Turkey in the Program for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), and the Monitoring and Evaluation of Academic Skills (ABIDE) test administrations. Simulation conditions were listed in Table 1.

Table 1

Factors in the Simulation Design

Factor	Condition
Ability distribution (old & new forms)	(0, 1) & (.05, 1) (0, 1) & (-.2, 1.25) (0, 1) & (-1, 1)
The number of exposed anchor items	2 (20%), 6 (50%), & 10 (100%)
Proportion of examinees with preknowledge	5%, 10%, 30%, & 60%

Ability Distribution

Another important factor in equating is the ability distribution. Wang et al. (2008) suggested that a mean difference in ability of 0.05 to 0.10 is considered “relatively large,” while a difference of 0.25 is considered “very large.” The ability distribution of examinees who took the old form was generated using a standard normal distribution $\theta\sim N(0,1)$; however, the ability distribution of examinees who took the new form varied between conditions. In this study, three different ability distributions were analyzed: $\theta\sim N(0.05, 1.00)$ was chosen as relatively large, $\theta\sim N(-0.20, 1.25)$ was chosen as large, and $\theta\sim N(-1.00, 1.00)$ was chosen as an unacceptably large ability mean difference. A low-ability examinee had an estimated ability level of less than 0, whereas a high-ability examinee had an estimated ability level of greater than 0 (Zopluoglu, 2017). Additionally, it has been suggested that individuals of lower ability are more likely to cheat (Cizek & Wollack, 2017). Therefore, all but one of the groups simulated in this study had ability means of less than 0.

Exposed Anchor Items

Several studies have examined the effects of differing levels of preknowledge. For example, Jurich (2011) used conditions in which 5 of 10 and 10 of 10 anchor items were exposed. Barri (2013) had 2 of 10 anchor items exposed. However, it is important to note that some studies, such as Eckerly (2017), suggest that a high degree of item compromise is not typical of tests. In fact, should this occur, the validity of the scores would be seriously compromised. Bearing this in mind, the number of exposed anchor items in this study was set at 2 (20%), 6 (60%), and 10 (100%) out of 10 anchor items. In addition, a condition in which no items were exposed was included. In this way, the change that occurs as the number of items exposed increases is more clearly observed.

Proportion of Examinees with Preknowledge

A range of percentages of examinees with preknowledge have been reported in other studies. The percentages of examinees with preknowledge were determined by Barri (2013) as 5%, 10%, 15%, and 20%; Zopluoglu (2017) as 20%, 40%, and 60% and Lee (2018) as 10%, 20%, 50%, and 70%. Considering the proportions of participants who had preknowledge in other investigations, the following values were employed in this study: 0%, 5%, 10%, 30%, and 60%.

Examinees’ probability of correctly responding to an exposed anchor item must also be taken into account. Previous studies have used values of .50 (Jurich, 2011), 1.00 (Barri, 2013), and/or .90 (Belov, 2016; Lee, 2018; Sinharay, 2017). In this study, the probability of a correct response was set at .90. In non-exposed conditions, no coefficient was added for the probability of correct answers to the anchor items. In other terms, the response probability in the non- exposed condition was modeled by the 2 PL model. The non-exposed condition was used as a basis for the comparisons. This situation was created as a situation where anchor test items were not shared between the groups in the test application with the NEAT.

Data Generation

The dichotomous item response data under the 2 PL model were generated using R (R Core Team, 2021). In the previous research, the Rasch model (Barri, 2013), the 3PL model (Jurich, 2011), and the 1PL model (Liu and Becker, 2022) were used to examine exposed items on test equating. For this reason, the 2 PL model, which is also preferred in large-scale test applications such as PISA, was used in this study. The latent trait model (ltm) package (Rizopoulos, 2006) was used for item and ability parameter estimations. This package provides marginal maximum likelihood for item parameters estimation and expected *a posteriori* for ability parameters estimation. The NEAT design requires two different test forms (old and new) with an anchor test, and also two different groups taking one test.

Test length: Spence (1996) has suggested that tests should have a minimum length of 35 items for equating purposes, whereas Kolen and Brennan (2014) suggested a range of 30 to 40 items. For this particular study, both the old and new versions of the test were constructed with a total of 40 items. Angoff (1984) and Kolen and Brennan (2014) suggest that the number of anchor items should be 20% of the test. Based on these suggestions, the anchor test was set at 10 items. This study utilized a test length of 40, with 10 internal anchor items used for both forms; meeting all of the above criteria for test length.

Sample Size: Kolen and Brennan (2014) noted that random equating error is influenced by sample size and suggest a minimum sample size of 400 per test form for linear equating methods and of 1,500 for equipercentile equating (Harris, 1993; Kolen & Brennan, 2014). Spence (1996) similarly recommends a minimum sample size of 500 for accurate equating results. In this research, a sample size of 2,000 was used, which exceeds each of these recommendations.

Ability distributions: For each group, the ability distribution was sampled from a standard normal distribution for all conditions.

Item parameters: Item difficulties were generated using a random normal distribution with the rnorm function, and item discrimination parameters were generated using a random log-normal distribution with the rlnorm function. For the old form and the anchor test, item difficulties had a mean of 0.00 and a variance of 1.00; the new form had a mean of 0.05 and a variance of 1.00. It was thought that the difference of .05 represented two forms with similar difficulties. The item discrimination parameter had a mean of 0.30 and a variance of 0.20. Descriptive statistics for the generated parameters are given in Table 2.

Table 2

Descriptive Statistics of True Item Parameters

Descriptive statistics	*b* (Old Form)	*b* (New Form)	*a* (Old Form)	*a* (New Form)	*b* (Anchor test)	*a* (Anchor test)
Mean	0.08	0.12	1.48	1.38	-0.08	1.47
SD	0.93	1.09	0.36	0.30	1.44	0.39
Min.	-1.81	-1.78	0.93	1.03	-2.02	1.06
Max	1.91	2.42	2.09	2.08	2.12	2.09

SD: standard deviation

As shown in Table 2, the *b* parameters (i.e., item difficulty parameters) for the old form ranged from -1.81 to 1.91, with a mean of 0.08 and an SD of 0.93. The *b* parameters for the new form ranged from -1.78 to 2.42, with a mean of 0.12 and an SD of 1.09. This design includes a noticeable difference in the mean difficulties of the old and new forms, but the *a* parameter values for the two were similar. For anchor items, *b* parameters ranged from -2.02 to 2.12, and *a* parameters ranged from 1.06 to 2.09.

Scaling Methods

The *plink* package (Weeks, 2010) in R was used for scaling transformations with MS and SL and equating tests under IRT true and observed score test equating.

The MS method (Marco, 1977) is the scaling method that uses the means and standard deviations of the *b* parameters of the common items to estimate the linking coefficients: *slope (A)* and *intercept (B)* in IRT scale transformation. The mean of item parameters $\mu(b_J)$, $\mu(b_I)$, $\sigma(b_J)$, and $\sigma(b_I)$ are given below:

$$A = (\sigma(b_J)) \,/ (\sigma(b_I)) \tag{2}$$

$$B = \mu(b_J) \; - \; A\mu(b_I) \tag{3}$$

The other method used in this study was the SL characteristic curve method (Stocking & Lord, 1983), which is one of the most widely used IRT-based equating. This is done by applying summation to the parameter estimates before squaring. The SL equation can be given as follows:

$$SL_{diff}(\theta_i) = \left[\sum_{j:V} \theta_{Ji}; \hat{a}_{Ji}, \hat{b}_{Ji} + \hat{c}_{Ji} \sum_{j:V} p_{ij}\left(\theta_{Ji}; \frac{\hat{a}_{Ij}}{A}, A\,\hat{b}_{Ij} + B, \hat{c}_{Ij}\right)\right]^2 \qquad (4)$$

The A and B coefficients obtained by using the MS and SL methods can then be used to transform the θ and item parameter estimates to the base scale as follows:

$$\theta_{Ji} = A\theta_{Ii} + B \qquad (5)$$

$$a_{Ji} = \frac{a_{Ii}}{B} \qquad (6)$$

$$b_{Ji} = Ab_{Ii} + B \qquad (7)$$

After calibration, the equating step was performed using IRT true score (IRT-T) and observed score (IRT-O) equating methods.

In IRT true score equating, the old test, *X*(θ), and the new test, *Y*(θ), are regarded as equivalent for a given θ. The true score of θ_i: is indicated by τx^{-1}

$$\tau(X) = \tau(Y)(\tau x^{-1}) \qquad (8)$$

IRT true score equating has the following three steps, and each step is performed for all true scores (Kolen & Brennan, 2014):

1. Choose a true score from form X [$\tau(X)$].
2. Identify the θ_i *corresponding* to the true score.
3. Define the true score of form Y that corresponds to θ_i.

In IRT observed score equating, after the observed score distribution of each form is obtained using IRT models, the tests are equated using the equipercentile method. The IRT observed score equating method consists of four steps (Kolen & Brennan, 2014):

1. For forms X and Y, the distribution of observed scores is calculated using the compound binomial distribution for examinees of a given ability. This is done using the recursion formula.
2. The distribution of observed examinee scores at each ability is obtained using equations 9 through 12 for forms X and Y. The distributions are then added together.

$$f_1(x) = \sum_i f(x|\theta_i)\varphi_1(\theta_i) \qquad (9)$$

$$f_2(x) = \sum_i f(x|\theta_i)\varphi_2(\theta_i) \qquad (10)$$

$$g_1(x) = \sum_i g(y|\theta_i)\varphi_1(\theta_i) \qquad (11)$$

$$g_2(x) = \sum_i g(y|\theta_i)\varphi_2(\theta_i) \qquad (12)$$

3. For IRT observed score equating under the NEAT design involving two populations, an equating function is typically viewed as defining a single population. Thus, populations 1 and 2 must be equated to be able to treat them as a single population. Populations 1 and 2 are weighted by w_1 and w_2 to form a synthetic population where $w_1 + w_2 = 1$ and $w_1, w_2 \geq 0$. Synthetic weights are used to determine the distributions in the synthetic population.

$$f_s(x) = w_1 f_1(x) + w_2 f_2(x) \qquad (13)$$

$$g_s(y) = w_1 g_1(y) + w_2 g_2(y) \tag{14}$$

4. The traditional equipercentile method is used to obtain equated scores. For this study, synthetic population weights of .50 were used for both populations for all groups ($w_1 = w_2 = .50$) for IRT observed score equating. The use of equal weights means that both populations were treated as contributing equally to the synthetic population (Kolen & Brennan, 2014).

Recovery in IRT Equating

The aim of this research is to see how the ability distribution, number of exposed anchor items, and proportion of examinees having preknowledge affect test equating under the 2 PL model. Root mean square error (RMSE) and bias were calculated to evaluate the recovery of the equating scores and scaling coefficients for the slope and intercept.

The equations for calculating bias and RMSE are given below:

$$Bias(x) = \frac{1}{R}\sum_{j=1}^{n} \hat{e}_y(x) - e_y(x) \tag{15}$$

$$\text{RMSE (x)} = \sqrt{\frac{\sum_{j=1}^{n}((\hat{e}_y(x) - e_y(x))^2}{R}} \tag{16}$$

where R is the number of replications, $e_y(x)$ is the true value, and $\hat{e}_y(x)$is the estimated value of each replication. 100 replications of the data were generated for each combination of ability distributions, the number of exposed anchor items, and the proportion of examinees with preknowledge.

Results

This simulation study was conducted to evaluate the impact of exposed anchor items on IRT true score observed equating methods under the 2 PL model.

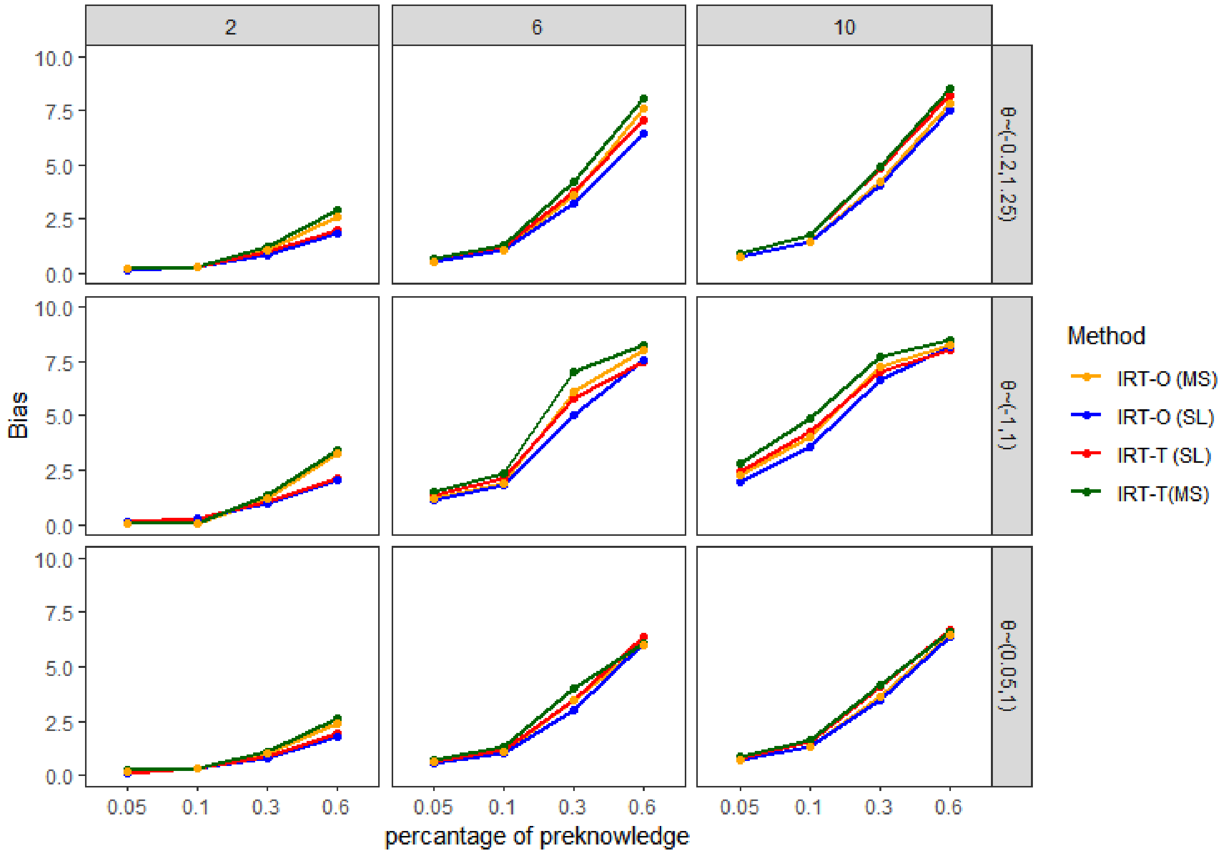
Bias and RMSE of Equated Scores

Figure 1 shows bias results for equating under different numbers of exposed anchor items and percentages of examinees with preknowledge for each of the different mean ability distributions. The bias under the nonexposed condition was close to zero (see Appendix 1) but positive in all exposure conditions. Positive bias indicates that the examinees with preknowledge produced higher scores than expected for the given condition. Bias increased slightly for each scaling method for the two exposed anchor items in both the true score and observed score equating methods compared to the condition with nonexposed item. The condition with two exposed items had a similar increasing pattern for ability distributions except for θ~*N*(-1,1). The MS scaling method showed a higher bias than the SL method when preknowledge was set at 60%. For the condition with two exposed anchor items, the largest amount of bias was found for the 30% condition, though bias was also observed for the 10% condition when the number of exposed items was six and 10.

The condition with 10 exposed items and 60% preknowledge resulted in larger, positively biased equated scores under IRT true score equating with the MS scaling method. As the number of exposed anchor items increased, bias also increased for both scaling methods under both equating methods. SL performed the best and produced the least bias for IRT observed score equating methods under the θ~*N*(0.05,1) ability condition. Both equating methods had similar amounts of bias under the θ~*N*(0.05,1) ability condition for all numbers of exposed items. For true score equating with the MS scaling method, the ability distribution θ~*N*(-1,1) produced the largest amount of bias except for the condition with two exposed anchor items. For observed score equating with SL scaling, the ability distributions θ~*N*(0.05,1) and θ~*N*(-0.2,1.25) produced the least bias.

Figure 1

Bias of Equated Score under IRT True and Observed Score Equating Methods with Different Scaling Methods



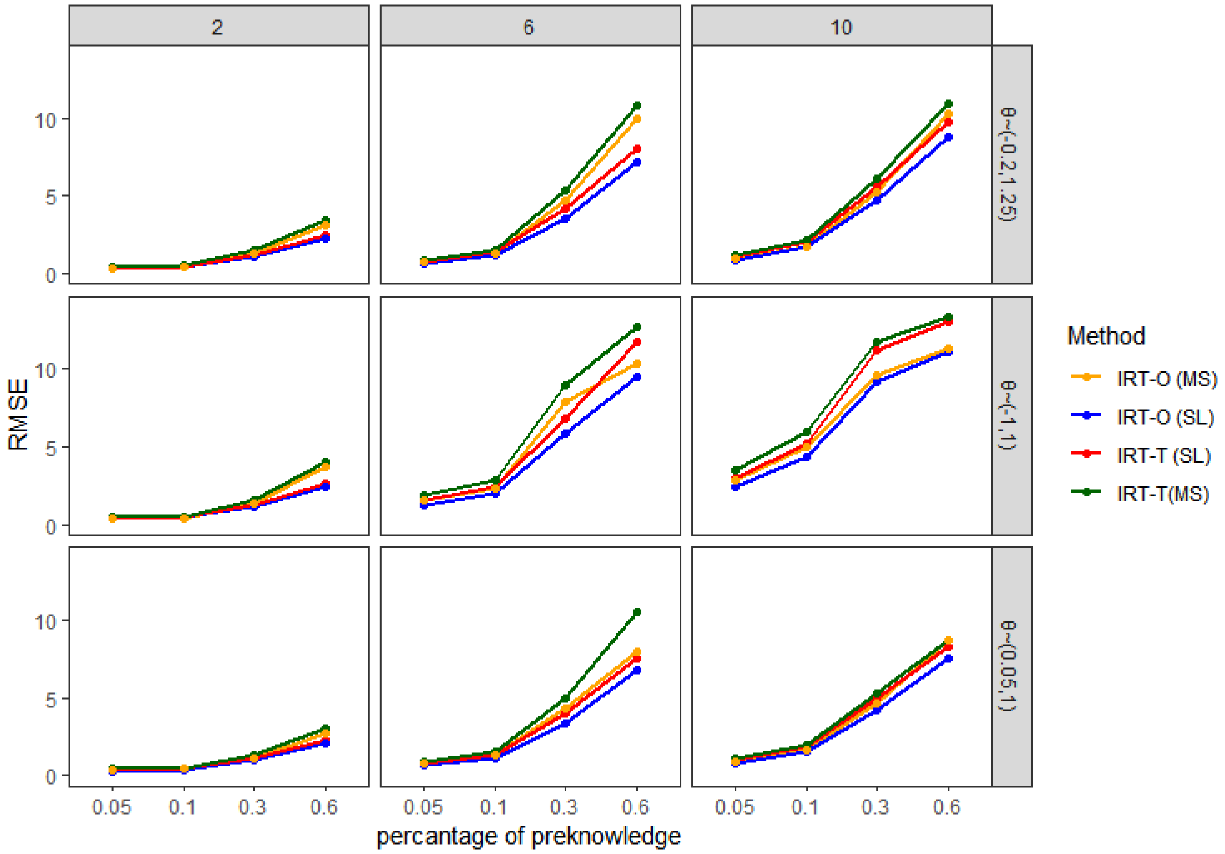
Overall, IRT true score equating produced higher levels of bias for all conditions than did IRT observed score equating. Additionally, the MS method produced more biased scores than the SL method. The largest amount of bias was observed when using the MS method under IRT true score equating, with individual scores being an average of 8.46 raw score points above the expected scores. It is also clear from our results that the ability distribution affected the estimated equating scores and that the number of exposed items affected the accuracy of equating.

Figure 2 shows the RMSE results for equating scores under different conditions. The RMSE results for both equating methods had the smallest values under the nonexposed conditions, whereas the largest value was obtained from the MS scaling method under IRT true score equating with the ability distribution $\theta \sim N(-1,1)$ (see Appendix 2).

In the condition with two exposed anchor items and 5% to 30% of examinees with preknowledge, the RMSE increased slightly for each scaling method. The RMSE for the MS scaling method was higher than that of the SL scaling method when the ability distribution was θ~*N*(-1,1), and the level of preknowledge was less than 60%. When the percentage of examinees with preknowledge increased from 5 to 10%, the RMSE increased. However, when preknowledge was increased from 10 to 30% and from 30 to 60%, the RMSE approximately doubled. This is especially noticeable for the MS scaling method with six exposed anchor items; the highest increase was from 30 to 60%. In addition, when the ability distribution was θ~*N*(-1.1), the IRT true score equating method with SL and MS exhibited a large RMSE with 60% preknowledge. When the ability distribution was θ~*N*(-0.2.1.25), the MS scaling method had a high RMSE for both equating methods. When there were 10 exposed anchor items with 5 to 10% preknowledge, the RMSE increased slightly for each scaling method when the ability distributions were θ~*N*(0.05,1) and θ~*N*(-0.2.1.25); however, the RMSE was particularly high, when the preknowledge of examinees was between 10 and 60% for θ~*N*(-0.2.1.25). In addition, when the mean ability distribution of the groups was different [θ~*N*(-0.2.1.25) vs. θ~*N*(-1.1)], the discrepancy between IRT true and observed score equating methods increased. This differentiation was most obvious when preknowledge was set at 60%. Another finding is that when the mean of the ability distribution was negative, the number of exposed items was six, and preknowledge was set at 60% (or when the number of exposed items was 10), IRT true score equating with the SL and MS scaling methods gave a higher RMSE than did IRT observed score equating.

Figure 2

RMSE of Equated Score under IRT True and Observed Score Equating Methods with Different Scaling Methods



As the number of exposed anchor items increased, the RMSE also increased under all conditions. The largest RMSE value was obtained from the ability distribution θ~*N*(-1.1) using the MS scaling method under IRT true score equating; in contrast, the smallest RMSE was produced from the ability distribution θ~*N*(0.05,1) with the SL scaling method under IRT observed score equating.

Bias and RMSE of Slope and Intercepts of the Scaling Methods

Table 3 shows that bias under nonexposed conditions was nearly zero for both the slope and the intercept. The bias of the slope was low when the percentage of preknowledge was 5 or 10%; however, this bias increased when the percentage of preknowledge reached 30% and then 60%. On the other hand, the bias of the intercept was still close to that of the nonexposed conditions when the percentage of preknowledge was 5 or 10% with two exposed items. As the percentage of preknowledge and the number of exposed items increased, the bias also increased. Additionally, both scaling methods underestimated the slope when two items were exposed. In contrast, as the number of exposed items increased, the slope was overestimated for all ability distributions when six or 10 items were exposed. The intercept, on the other hand, was overestimated by both scaling methods for nearly all conditions, and the recovery of the slope was more accurate than that of the intercept in almost all conditions. The results show that the recovery of the slope and intercept was affected by the ability distribution: the most biased estimate of the slope was obtained using ability distribution of θ~*N*(-1,1), 60% preknowledge, and 10 exposed items under the MS scaling method. The MS method had a larger bias than the SL for all conditions, except with 10 exposed items and 60% preknowledge when the mean ability distribution was θ~*N*(-1,1).

Table 3
Bias of Slope and Intercept

		Ability Distribution											
		θ~*N*(0.05,1)				θ~*N*(-0.2,1.25)				θ~*N*(-1,1)			
	Scaling methods	SL		MS		SL		MS		SL		MS	
Item Pre.	Perc. of Pre.	A	B	A	B	A	B	A	B	A	B	A	B
none		0.01	-0.01	0.01	-0.01	-0.01	0.01	-0.01	0.00	0.01	0.02	0.02	0.03
2	5%	0.02	-0.03	0.01	-0.04	0.00	-0.02	-0.01	-0.02	0.03	0.02	0.02	0.04
	10%	0.03	-0.07	0.01	-0.07	0.01	-0.05	-0.01	-0.04	0.05	-0.01	0.04	0.03
	30%	0.05	-0.20	0.02	-0.20	0.04	-0.16	-0.03	-0.16	0.09	-0.18	0.05	-0.18
	60%	0.11	-0.45	0.11	-0.55	0.14	-0.38	-0.12	-0.48	0.22	-0.46	0.22	-0.69
6	5%	-0.02	-0.12	-0.04	-0.12	-0.01	-0.09	-0.04	-0.08	-0.05	-0.17	-0.11	-0.17
	10%	-0.03	-0.21	-0.07	-0.20	-0.01	-0.17	-0.07	-0.15	-0.05	-0.31	-0.17	-0.29
	30%	-0.10	-0.62	-0.28	-0.60	-0.03	-0.55	-0.25	-0.51	-0.24	-0.92	-0.51	-0.98
	60%	-0.11	-1.14	-0.59	-1.33	0.03	-1.20	-0.50	-1.26	-0.48	-1.63	-0.54	-1.72
10	5%	-0.04	-0.13	-0.05	-0.14	-0.03	-0.11	-0.05	-0.11	-0.16	-0.30	-0.21	-0.33
	10%	-0.08	-0.25	-0.11	-0.25	-0.05	-0.23	-0.09	-0.21	-0.28	-0.55	-0.34	-0.60
	30%	-0.25	-0.66	-0.32	-0.65	-0.15	-0.65	-0.25	-0.62	-0.65	-1.11	-0.62	-1.23
	60%	-0.50	-1.15	-0.56	-1.16	-0.27	-1.30	-0.45	-1.27	-0.81	-1.80	-0.78	-1.87

A: slope; B: intercept; Perc.of Pre: percentage of preknowledge

Table 4
RMSE for Slope and Intercept

		Ability Distribution											
		θ~*N*(0.05,1)				θ~*N*(-0.2,1.25)				θ~*N*(-1,1)			
Scaling methods		SL		MS		SL		MS		SL		MS	
Item Pre.	Perc. of Pre.	A	B	A	B	A	B	A	B	A	B	A	B
none		0.02	0.02	0.03	0.03	0.02	0.02	0.02	0.02	0.03	0.03	0.04	0.05
2	5%	0.04	0.04	0.03	0.05	0.02	0.03	0.03	0.03	0.04	0.04	0.05	0.05
	10%	0.04	0.08	0.03	0.08	0.02	0.05	0.03	0.05	0.06	0.05	0.06	0.06
	30%	0.06	0.20	0.04	0.20	0.05	0.16	0.04	0.17	0.10	0.18	0.08	0.19
	60%	0.11	0.45	0.12	0.55	0.14	0.38	0.13	0.49	0.22	0.47	0.22	0.70
6	5%	0.03	0.12	0.05	0.13	0.02	0.09	0.05	0.08	0.05	0.17	0.12	0.17
	10%	0.04	0.21	0.09	0.20	0.02	0.18	0.08	0.15	0.06	0.31	0.18	0.29
	30%	0.11	0.62	0.29	0.61	0.04	0.56	0.26	0.51	0.25	0.93	0.51	0.99
	60%	0.12	1.14	0.59	1.14	0.04	1.25	0.50	1.26	0.50	1.64	0.54	1.74
10	5%	0.05	0.13	0.07	0.14	0.04	0.11	0.06	0.11	0.16	0.30	0.21	0.34
	10%	0.08	0.25	0.13	0.26	0.05	0.23	0.09	0.23	0.29	0.55	0.35	0.61
	30%	0.25	0.66	0.33	0.66	0.15	0.65	0.26	0.63	0.65	1.12	0.63	1.24
	60%	0.50	1.16	0.59	1.17	0.28	1.31	0.49	1.28	0.81	1.81	0.79	1.88

A: slope; B: intercept; Perc.of Pre: percentage of preknowledge

Table 4 shows the RMSE for the slope and intercept under various conditions. The RMSE under nonexposed conditions was nearly zero for both the slope and the intercept. At 5 and 10% preknowledge, the RMSE was close to that of the nonexposed conditions; however, the RMSE increased for 30% preknowledge, and increased again for 60%. These results suggest that the recovery of the slope was affected by the ability distribution. The largest RMSE of the slope was obtained when the ability distribution was θ~*N*(-1,1), preknowledge was 30% or 60%, and 10 items were exposed under the SL scaling method. As with the slope, the RMSE of the intercept increased when the percentage of preknowledge and number of exposed items increased. Under all conditions, however, the recovery of the slope was more accurate than that of the intercept. The SL and MS scaling methods had a similar RMSE for all conditions, except when preknowledge was set at 60%. The RMSE was close to that of nonexposed conditions for 5 and 10% preknowledge with two exposed items, though it was larger for 30% and 60% preknowledge. In addition, these results suggest that the conditions chosen influenced the estimation of the intercept, especially the number of exposed items and mean ability distribution.

Conclusion and Discussion

In this study, we aimed to investigate the effect of the anchor item preknowledge on equated scores and scaling coefficients under IRT true and IRT observed score equating. This was premised on the idea that the validity of inferences based on test scores becomes questionable if individuals have preknowledge of the anchor items on tests.

The results of this study suggest that as the number of exposed items and percentage of examinees with item preknowledge increase, bias also increases. As all bias observed in this study was positive, the equated scores were estimated with higher values than the true (i.e., generating) values. The amount of bias differed based on the scaling and equating methods used. Results obtained from MS exhibited a larger bias than those obtained from SL. Our finding that the SL method provides more accurate and stable equating results than the MS method is in line with previous research (Kim & Cohen, 1992; Kim & Kolen, 2006; Kim & Lee, 2006). It can be explained that MS, which requires simple summary statistics (Kim, 2004), has higher errors because it is more sensitive to the variation of the estimates of the b parameter, and as a result, the slope and the intercept values may become unstable. Furthermore, bias and RMSE increased with the number of exposed items and percentage of preknowledge, which is also consistent with previous research (Barri, 2012; Chen, 2021; Jurich, 2011; Kopp & Jones, 2020).

Both linking methods had higher bias values for IRT equating than the nonexposed condition. However, results for 5 and 10% preknowledge for the two exposed items condition had bias values close to those of the nonexposed condition. One possible reason for this may be that the MS method considers item parameters separately while the SL method considers them simultaneously (Kolen & Brennan, 2014; Tian, 2011). The MS method is more directly affected by variation in the item difficulty parameter since the scaling coefficients depend on the item difficulty parameter, and item preknowledge increases the probability of a correct answer. This result is consistent with the findings of Lee and Becker (2022), who reported that as the percentage of examinees with preknowledge increases, the variance of the item difficulty parameter estimates increases for exposed conditions. Finally, consistent with previous research (Barri, 2012; Jurich, 2011), when the ability distributions were similar or equivalent, the bias and RMSE were lower and thus appeared to have a more minor effect on equated scores.

IRT true score equating exhibits a larger bias and RMSE does IRT observed score equating for both scaling methods. However, bias and RMSE values for both equating methods were similar in the nonexposed condition. Our findings are consistent with Tao and Cao’s (2016) findings, which showed that IRT observed score equating outperforms the IRT true score equating. However, others found that IRT observed score equating are more stable compared to IRT true score equating (Han et al., 1997). Due to different results on equating methods in the related literature, there is no consensus on the best method. IRT observed score equating uses synthetic weights, while IRT true score equating uses the true score to equate through an ability parameter (Ogasawara, 2003). As a result, the presence of exposed items changes the probability of a correct answer, resulting in higher scores and thereby higher bias and RMSE values. In the present study, we utilized equal synthetic weights for groups, which may have affected the difference between the equated scores of the two groups.

Higher levels of bias and RMSE were observed as item exposure and percentage of knowledge increased for both scaling methods, which is consistent with the previous research (Barri, 2012; Chen, 2021; Jurich, 2011). The scaling coefficient *A* was overestimated by both scaling methods for the condition with two exposed items and underestimated for the conditions with six and ten exposed items. In other words, the number of exposed items appeared to affect the estimation of coefficient *A*. These findings were consistent with Barri (2012) and Chen (2021) for the condition with two exposed items and with Jurich (2011) in the sense that coefficient *A* was underestimated by both estimation methods for the conditions with six and ten exposed items. The reason for these differences in the effect on scaling coefficient *A* may be that the variance of the difficulty parameter changes as the difficulty of the exposed items decreases. Jurich (2011) suggests that while the probability of a correct answer will increase as the number of exposed items increases, this may also lead to a decrease in item discrimination. This situation may also cause an underestimation of coefficient *A*. On the other hand, scaling constant *B* was underestimated under all conditions, which is consistent with Barri (2011) but not with other studies (e.g., Jurich, 2011). This disagreement may be due, in part, to the way in which item preknowledge was simulated. Barri (2012) simulated item preknowledge by adding 1 to the probability of a correct answer, but Jurich (2011) added .5. In this study, item preknowledge was simulated by adding .9. In addition, scaling coefficient A was more accurate than scaling coefficient B for both linking methods. Thus, as the number of exposed items, the percentage of examinees with item preknowledge, and the differences in mean ability increased, bias and RMSE values increased for both scaling coefficients.

Ability also had an impact on linking and equating results when item exposure occurred, which is consistent with the previous research (Barri, 2012; Chen, 2021; Jurich, 2011). As the difference in mean ability increased, bias and RMSE also increased. This effect can be seen in low-ability examinees, who correctly answer exposed items at a higher rate than higher-ability examinees (Barri, 2012).

We found that the choice of scaling or equating methods may be of less importance when items are exposed. However, the generalizability of our findings needs to be critically evaluated. Some examinees exhibit a greater change in equated scores when they have preknowledge of the anchor items. In fact, this situation affects not only examinees with preknowledge but also the decisions made about all examinees who take the test (Jurich, 2011). For this reason, the effects of exposure should be examined before equating. Otherwise, the validity of the decisions made may be open to question. The bias results from this study suggest that if the anchor items are exposed, it is most appropriate to exclude them from the test before equating, as item preknowledge affects the equated test scores.

In this study, we focused on determining the effect of item preknowledge on IRT test equating methods under the NEAT design. We acknowledge that several additional avenues for future research exist. First, in this study, equal synthetic weights were used. Future research might choose differing synthetic weights to determine their effect on equated scores. Second, examining the effect of item preknowledge on tests with mixed item formats would be useful. Third, although the NEAT design is frequently reported in the existing literature, other equating designs, such as the random group design and common-item equivalent groups design, could be used. Fourth, the variance of the item difficulty parameter estimates may differ in real data. Therefore, the effect of exposed items on the test equating can be examined in a real data set. Finally, future studies can extend our work of using IRT models to estimate parameters and equating by applying other methods (e.g., classical equating, Bayesian nonparametric, and kernel equating).

Declarations

Author Contribution: Cigdem Akin Arikan: Conceptualization, methodology, analysis, writing & editing, visualization. Allan S. Cohen: Conceptualization, writing-review & editing, supervision.

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

Ethical Approval: Ethical rules were followed in this research. Ethical approval is not required, because simulation data was used in this research.

Funding: This study was funded by the Scientific and Technological Research Council of Turkey (TÜBİTAK) BIDEB 2219.

References

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

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Educational Testing Service.

Barri, M. A. (2013). *The impact anchor item exposure on mean/sigma linking And IRT true score equating under the neat design* [Unpublished master's thesis]. University of Kansas.

Belov, D. I. (2016). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement*, *40*(2), 83-97. https://doi.org/10.1177/0146621615603

Chen, D. F. (2021). *Impact of item parameter drift on IRT linking methods* [Unpublished doctoral thesis]. The University of North Carolina.

Cizek, G. (1999). *Cheating on tests: how to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J., & Wollack, J. A. (Eds.). (2017). *Handbook of quantitative methods for detecting cheating on tests*. Routledge.

Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational measurement: Issues and practice*, *10*(3), 37-45. https://doi.org/10.1111/j.1745-3992.1991.tb00207.x

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

Demir, M. K., & Arcagok, S. (2013). Primary school teacher candidates' opinions on cheating in exams. *Erzincan University Faculty of Education Journal*, *15*(1), 148-165. Retrieved from https://dergipark.org.tr/en/pub/erziefd/issue/6010/80121

Eckerly, C. A. (2017). Detecting preknowledge and item compromise. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 101-123). Routledge.

Fly, B. J. (1995). *A study of ethical behaviour of students in graduate training programs in psychology* [Unpublished doctoral thesis]. University of Denver.

Foster, D. (2013). Security issues in technology-based testing. In J. A. Wollack and J. J. Fremer , Eds., *Handbook of test security* (pp. 39–83). Routledge

Gorney, K., & Wollack, J. A. (2022). Generating models for item preknowledge. *Journal of Educational Measurement*, *59*(1), 22-42. https://doi.org/10.1111/jedm.12309

Han, T., Kolen, M., & Pohlmann, J. (1997). A comparison among IRT true and observed-score equatings and traditional equipercentile equating. *Applied Measurement in Education*, *10*(2), 105-121. https://doi.org/10.1207/s15324818ame1002_1

Harris, D. J. (1993, April). *Practical issues in equating* [Paper presentation]. American Educational Research Association, Atlanta, Georgia, USA.

Josephson Institute (2012). *Josephson Institute's 2012 report card on the ethics of American youth*. Los Angeles, CA. Retrieved from http://charactercounts.org/programs/reportcard/2012/index.html.

Jurich, D. P. (2011). *The impact of cheating on IRT equating under the non-equivalent anchor test design* [Unpublished master's thesis]. James Madison University.

Jurich, D. P., Goodman, J. T., & Becker, K. A. (2010). *Assessment of various equating methods: Impact on the pass-fail status of cheaters and non-cheaters*. In Poster presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Kane, M. T., & Mroch, A. A. (2020). Orthogonal Regression, the Cleary Criterion, and Lord's Paradox: Asking the Right Questions. *ETS Research Report Series*, *2020*(1), 1-24. https://doi.org/10.1002/ets2.12298

Kim, S. (2004). *Unidimensional IRT scale linking procedures for mixed-format tests and their robustness to multidimensionality [*Doctoral dissertation]. Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3129309)

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed format tests. *Applied Measurement in Education*, *19*, 357–381.

Kim, S. H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of educational measurement*, *29*(1), 51-66.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices.* 3rd Edn. Springer

Kopp, J. P., & Jones, A. T. (2020). Impact of item parameter drift on Rasch scale stability in small samples over multiple administrations. *Applied Measurement in Education*, *33*(1), 24-33.

Liu, J., & Becker, K. (2022). The Impact of cheating on score comparability via pool-based IRT pre-equating. *Journal of Educational Measurement*, *59*(2), 208-230. https://doi.org/10.1111/jedm.12321

Lee, S. Y. (2018). *A mixture model approach to detect examinees with item preknowledge* [Unpublished doctoral dissertation]. The University of Wisconsin-Madison.

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

Man, K., Harring, J. R., & Sinharay, S. (2019). Use of data mining methods to detect test fraud. *Journal of Educational Measurement*, *56*(2), 251-279. https://doi.org/10.1111/jedm.12208

Marco, G. L. (1977). Item Characteristic Curve Solutions to Three Intractable Testing Problems. *Journal of Educational Measurement*, *14* (2), 139-160.

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice*, *35*(1), 38-47. https://doi.org/10.1111/emip.12102

Pan, Y., & Wollack, J. A. (2021). An unsupervised-learning based approach to compromised items detection. *Journal of Educational Measurement*, *58*(3), 413-433. https://doi.org/10.1111/jedm.12299

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from https://www.R-project.org/

Rizopoulos, D. (2006). ltm: An R Package for latent variable modeling and item response analysis. *Journal of Statistical Software*, *17*(5), 1–25. https://doi.org/10.18637/jss.v017.i05

Sinharay, S. (2017). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics*, *42*(1), 46-68. https://doi.org/10.3102/1076998616673872

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research*, *56*(4), 495-529.

Spence, P. D. (1996). *The effect of multidimensionality on unidimensional equating with item response theory* [Unpublished doctoral dissertation]. University of Florida.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201-210.

Tan, Ş. (2001). Sınavlarda kopya çekmeyi önlemeye yönelik önlemler [Measures against cheating in exams]. *Education and Science*, *26*(122), 32-40.

Tao, W., & Cao, Y. (2016). An extension of IRT-based equating to the dichotomous testlet response theory model. *Applied Measurement in Education*, *29*(2), 108-121.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. Springer.

Wang, J., Tong, Y., Ling, M., Zhang, A., Hao, L., & Li, X. (2015). Analysis on test cheating and its solutions based on extenics and information technology. *Procedia Computer Science*, *55*, 1009-1014. https://doi.org/10.1016/j.procs.2015.07.1024

Wang, T., Lee, W., Brennan, R. L., & Kolen, M. J. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item nonequivalent groups design. *Applied Psychological Measurement*, *32*, 632-651. https://doi.org/10.1177/0146621608314943

Weeks, J. P. (2010). plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software*, *35*(12), 1–33. https://doi.org/10.18637/jss.v035.i12

Zimmermann, S., Klusmann, D., & Hampe, W. (2016). Are exam questions known in advance? Using local dependence to detect cheating. *PloS One*, *11*(12). https://doi.org/10.1371/journal.pone.0167545

Zopluoglu, C. (2017). Similarity, answer copying, and aberrance. Understanding the status Quo. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 25–46). Routledge.

Appendix 1

Table 1
The Bias of Equated Scores

		Ability Distribution											
		θ~*N*(0.05,1)				θ~*N*(-0.2,1.25)				θ~*N*(-1,1)			
Item Pre.	Percentage of preknowledge	SL		MS		SL		MS		SL		MS	
		IRT-T	IRT-O	IRT-T	IRT-O	IRT-T	IRT-O	IRT-T	IRT-O	IRT-T	IRT-O	IRT-T	IRT-O
	non	0.02	0.03	0.05	0.06	0.02	0.03	0.05	0.05	0.00	0.02	-0.04	-0.01
2	5%	0.13	0.13	0.22	0.21	0.19	0.17	0.21	0.19	0.11	0.14	0.04	0.05
	10%	0.32	0.29	0.32	0.29	0.33	0.30	0.33	0.29	0.31	0.30	0.07	0.08
	30%	0.89	0.82	1.10	0.98	0.96	0.87	1.22	1.07	1.04	0.98	1.35	1.18
	60%	1.91	1.79	2.64	2.43	2.01	1.86	2.88	2.61	2.11	2.06	3.45	3.24
6	5%	0.65	0.56	0.73	0.63	0.67	0.56	0.68	0.56	1.32	1.10	1.51	1.22
	10%	1.19	1.01	1.31	1.11	1.24	1.04	1.26	1.03	2.13	1.77	2.32	1.86
	30%	3.49	3.03	3.97	3.44	3.75	3.20	4.24	3.60	5.77	5.02	7.03	6.08
	60%	6.42	6.05	6.05	6.03	7.05	6.40	8.06	7.60	7.47	7.59	8.24	8.00
10	5%	0.79	0.68	0.87	0.74	0.90	0.75	0.92	0.76	2.41	1.98	2.82	2.29
	10%	1.52	1.30	1.60	1.35	1.74	1.46	1.72	1.42	4.28	3.54	4.84	4.00
	30%	4.05	3.50	4.18	3.61	4.79	4.09	4.89	4.17	7.02	6.65	7.70	7.28
	60%	6.68	6.39	6.61	6.43	8.23	7.49	8.46	7.84	7.99	8.14	8.46	8.27

Appendix 2

Table 2

The RMSE of Equated Scores

		Ability Distribution											
Item Pre.	Percentage of preknowledge	θ~*N*(0.05,1)				θ~*N*(-0.2,1.25)				θ~*N*(-1,1)			
		SL		MS		SL		MS		SL		MS	
		IRT-T	IRT-O	IRT-T	IRT-O	IRT-T	IRT-O	IRT-T	IRT-O	IRT-T	IRT-O	IRT-T	IRT-O
2	non	0.22	0.2	0.26	0.24	0.23	0.21	0.28	0.26	0.29	0.27	0.44	0.41
	5%	0.29	0.25	0.4	0.35	0.34	0.29	0.39	0.34	0.41	0.34	0.49	0.41
	10%	0.43	0.37	0.48	0.42	0.46	0.40	0.48	0.41	0.53	0.45	0.51	0.42
	30%	1.06	0.95	1.31	1.14	1.18	1.05	1.45	1.25	1.28	1.16	1.60	1.37
	60%	2.28	2.09	3.07	2.75	2.478	2.27	3.42	3.05	2.63	2.46	4.07	3.70
6	5%	0.78	0.65	0.9	0.77	0.779	0.64	0.84	0.70	1.58	1.25	1.90	1.52
	10%	1.35	1.13	1.58	1.34	1.4	1.14	1.52	1.28	2.40	1.97	2.83	2.35
	30%	3.95	3.39	4.94	4.32	4.203	3.56	5.35	4.68	6.85	5.84	9.01	7.86
	60%	7.53	6.78	10.54	7.95	8.056	7.20	10.80	10.00	11.81	9.49	12.70	10.39
10	5%	0.98	0.82	1.09	0.91	1.07	0.88	1.12	0.93	2.92	2.40	3.47	2.86
	10%	1.82	1.52	1.98	1.66	2.017	1.66	2.09	1.73	5.24	4.36	6.01	5.03
	30%	4.92	4.23	5.32	4.61	5.597	4.74	6.06	5.21	11.28	9.18	11.78	9.59
	60%	8.25	7.51	8.7	8.69	9.791	8.82	11.93	10.27	13.10	11.10	13.43	11.35

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(4); 472-491

The Factor Structure of the Satisfaction with Life Scale (SWLS): A Meta-Analytic Structural Equation Modeling

Sedat KANADLI* Nezaket Bilge UZUN**

Abstract

This study aims to determine the most appropriate factor structure for the life satisfaction scale by using the meta-analytic structural equation modeling (MASEM) approach. For this purpose, we extracted 41 correlation matrices from 33 primary studies (*N* = 49316) in accordance with the inclusion criteria. Results of the heterogeneity test indicated that the matrices were heterogeneous. Therefore, in the first step of MASEM, we created the total correlation matrix according to the random-effects model. At this stage, we determined that there was a large and statistically significant relationship between the scale items. In the second phase of MASEM, we established four models of SWLS (original single-factor model, modified single-factor model, two-factor model, and the first three-item model). As a result of the analysis, we determined that although the goodness-of-fit indices of the original single-factor model were at a "good" level, the model-data fit of the modified-single-factor model and the two-factor model was better. However, we determined that the modified-single factor model was the most appropriate one since there was a high correlation (r = .92, p < .01) between the factors in the two-factor model, and its divergent validity could not be ensured. We determined that the first three-item model is a saturated model. Therefore, it is not possible to compare the statistically obtained findings for this model.

Keywords: life satisfaction, SWLS, meta-analysis, two-stage structural equation modeling

Introduction

Variables play a critical role in the effort of trying to understand how they are causally associated, which is the main purpose of science in scientific research processes. In education, social sciences, and psychology, the characteristics that are subject to measurement are tried to be defined through indirect measurements. In these definitions, measurement tools are used to measure these variables. The psychometric qualities of the measurement tools used to measure these constructs, which are the subject of the measurement, have a critical role. One relevant construct in psychology is "life satisfaction ", which is one of the study subjects that are the focus of attention of researchers (Appleton & Song, 2008; Dağli & Baysal, 2016; Vassar, 2008). Life satisfaction is defined as a judgmental process in which individuals evaluate their quality of life according to their own criteria (Shin & Johnson, 1978). Therefore, the judgment of life satisfaction depends on the comparison of the conditions in which the individuals are living with the standards they think are appropriate (Diener et al., 1985). As a result of

* Assoc. Prof. Dr., Mersin University, Faculty of Education, Mersin-Türkiye, skanadli@mersin.edu.tr, ORCID ID: 0000-0002-0905-8677
** Assoc. Prof. Dr., Mersin University, Faculty of Education, Mersin-Türkiye, buzun@mersin.edu.tr, ORCID ID: 0000-0003-2293-4536

To cite this article:
Kanadlı, S. & Uzun N. B. (2023). The factor structure of the satisfaction with life scale (SWLS): A meta-analytic structural equation modeling. *Journal of Measurement and Evaluation in Education and Psychology, 14*(4), 472-491. https://doi.org/10.21031/epod.1221365

Received: 7.03.2023
Accepted: 20.11.2023

this comparison, as long as the conditions comply with the standards, the person reports a high level of life satisfaction (Pavot & Diener, 1993).

There are different scales in the literature to measure individuals' life satisfaction. These scales; Satisfaction with Life Scale (Diener et al., 1985), Quality of Life Inventory (Frisch, 1994), Riverside Life Satisfaction Scale (Margolis et al., 2018) can be given as examples. The most preferred of these scales is the Satisfaction with Life Scale (SWLS) developed by Diener et al. (1985). Thus, a Google Scholar search of Diener et al.'s (1985) study that introduced SWLS yielded 36582 citations until 2022, which provides information about the potential of the SWLS. This scale can be preferred in many large national surveys (e.g., General Social Survey, German Socio-Economic Panel, and World Value Survey) for its high reliability and validity instead of single-item measurements. Therefore, having a valid, concise, and easy-to-understand measurement tool is of great significance.

Diener et al. (1985) used the principal axis factoring method to explain the structure of the scale and obtained a one-factor, five-item structure that explained 66% of the variability. These items are; *(i)* In most ways my life is close to ideal, *(ii)* The conditions of my life are excellent, *(iii)* I am satisfied with my life, *(iv)* So far, I have gotten the important things I want in life, and *(v)* If I could live my life over, I would change almost nothing. The SWLS items are global rather than specific in nature, allowing respondents to weigh domains of their lives in terms of their own values. The Cronbach's alpha coefficient, showing the internal consistency of the measurements obtained in the scale development study, and the test-retest reliability coefficients obtained in two months related to its stability were reported as .87 and .82, respectively. Pavot and Diener (1993, 2008) reported in their review studies that the coefficient alpha for the SWLS ranged from .79 to .89, indicating that the scale has high internal consistency. In support of this, a meta-analysis of 60 studies that assessed SWLS reliability reported an average Cronbach alpha coefficient value of .78, with confidence intervals of 95%, ranging from .76 to .80 (Vassar, 2008). Similarly, Busseri (2018) calculated the reliability of the scale as .82, 95% CI [.75, .89] in his meta-analysis study. At the same time, the psychometric findings of the SWLS scale, which has been adapted to many cultures (France, Germany, Russia, Korea, Turkey, Portugal, and Romania), show that the scale is widely used to measure life satisfaction in psychological research.

The Debate on the Factor Structure of the SWLS

The purpose of this studyis to determine the most appropriate factor structure for the SWLS. For this reason three confirmatory factor analysis (CFA) models of the SWLS were evaluated and specified in the present study based on findings in previous research: Model 1: The original single-factor model with all SWLS items loading onto a single factor of life satisfaction (Pinto da Costa & Neto, 2019; Dahiya & Rangnekar, 2020; Dirzyte et al., 2021; Espejo et al., 2022; Gadermann et al., 2009; Galanakis et al., 2017; Garcia et al., 2021; Jovanovic, 2019; Lopez-Ortega et al., 2016; Marcu, 2013; Sachs, 2003; Sagar & Karim, 2014; Sancho et al., 2014; Silva et al., 2014; Wu et al., 2009; Wu & Wu, 2008). Model 2: The modified single-factor model, allowing for correlated errors between items 4 and 5 (Clench-Aas et al., 2011; Cazan, 2014; Dahiya & Rangnekar, 2020; Jovanovic, 2019; Mishra, 2019; Moksnes et al., 2013; Sachs 2003). Model 3: The two-factor model with two correlated factors; present satisfaction (items 1, 2, and 3) and past satisfaction (items 4 and 5) (Dahiya & Rangnekar, 2020; Hultell & Gustavsson, 2008; Jovanovic, 2019; Sachs 2003; Slocum-Gori et al., 2009; Wu et al., 2006).

Studies on the construct validity of the scale confirmed the one-factor structure (Arrindell et al., 1991; Atienza et al., 2000; Diener et al., 1985; Glaesmer et al., 2011; Swami & Chamorro-Premuzic, 2009). Although Model 1 was supported in some studies (Gouveia et al., 2009; Jovanović, 2016), the goodness-of-fit of this model was "poor" in other studies (Fabio & Gori, 2015; Wang et al., 2017). Pavot and Diener (1993) reported weak convergence of the last two items with the other three items based on item-total score correlations and factor loadings. In this respect, in other studies, Model 2 showed a better model-data fit than Model 1 (Clench-Aas et al., 2011; Dahiya & Rangnekar, 2020; Moksnes et al., 2014; Pavot & Diener, 2008; Sach, 2003). In some studies, Model 3 was used (Hultell & Gustavsson 2008; Wu & Yao, 2006). In these studies, in which a multifactorial structure is suggested, the first three items are generally related to present life, and the last two items are based on emphasizing past life. However, in studies where models are compared or examined together, although Model 3 produces similar goodness-of-fit values to Model 2 (Jovanović, 2019), it is not recommended due to the high correlation between factors (Dahiya & Rangnekar, 2020; McDanold, 1999; Sachs, 2003).

Researchers working on SWLS drew attention to the high factor loadings and item-total correlations of the first three items of the SWLS (Diener et al., 1985, Pavot & Diener, 2009; Vittersø et al., 2009; Kjell & Diener, 2021). Some studies have suggested a one-factor and three-item model by associating the last two items of the 5-item scale weak convergence with the others from a theoretical perspective (Pavot & Diener, 2008). In addition, CFA and Multi-Group CFA (MGCFA) focused studies on the 3-item single factor structure produced findings supporting this suggestion (Espejo et al., 2022; Kjell & Diener, 2021). For this reason, it is considered important to examine the structure of the 3-item single factor model (Model 4) within the scope of the research.

Therefore, it is seen that there is a disagreement in the literature about which is the best model to represent the factor structure of the life satisfaction scale. Based on these cross-cultural findings, many studies emphasize that the in-depth investigation of the factor structure of SWLS, which has a significant potential for use, and the comparison of the relevant factor structure through multi-group applications can contribute to theory and practice (Clench-Aas et al., 2011; Glaesmer et al., 2011; Pavot & Diener, 1993; Pavot & Diener, 2008; Tucker et al., 2006). As a part of the validity studies, questions such as “Do the five items in the scale come together to produce a single implicit feature scoring” or “Are there sub-latent variables in the scale that need to be scored separately?” can arise. The answers to such questions will also affect the decisions to be taken based on the measurements obtained from the SWLS. This study aims to resolve this conflict in the literature based on a meta-analytic structural equation modeling (MASEM) approach. Through this approach, which includes both meta-analysis and structural equation modeling, it is aimed to synthesize the structures obtained regarding life satisfaction by considering the results of MGCFA carried out in different cultures and groups. The main research question covered in the study is: “What is the model that best represents the factor structure of the life satisfaction scale by using the inter-item correlation matrices obtained from SEM-based studies?”

Method

Literature Search

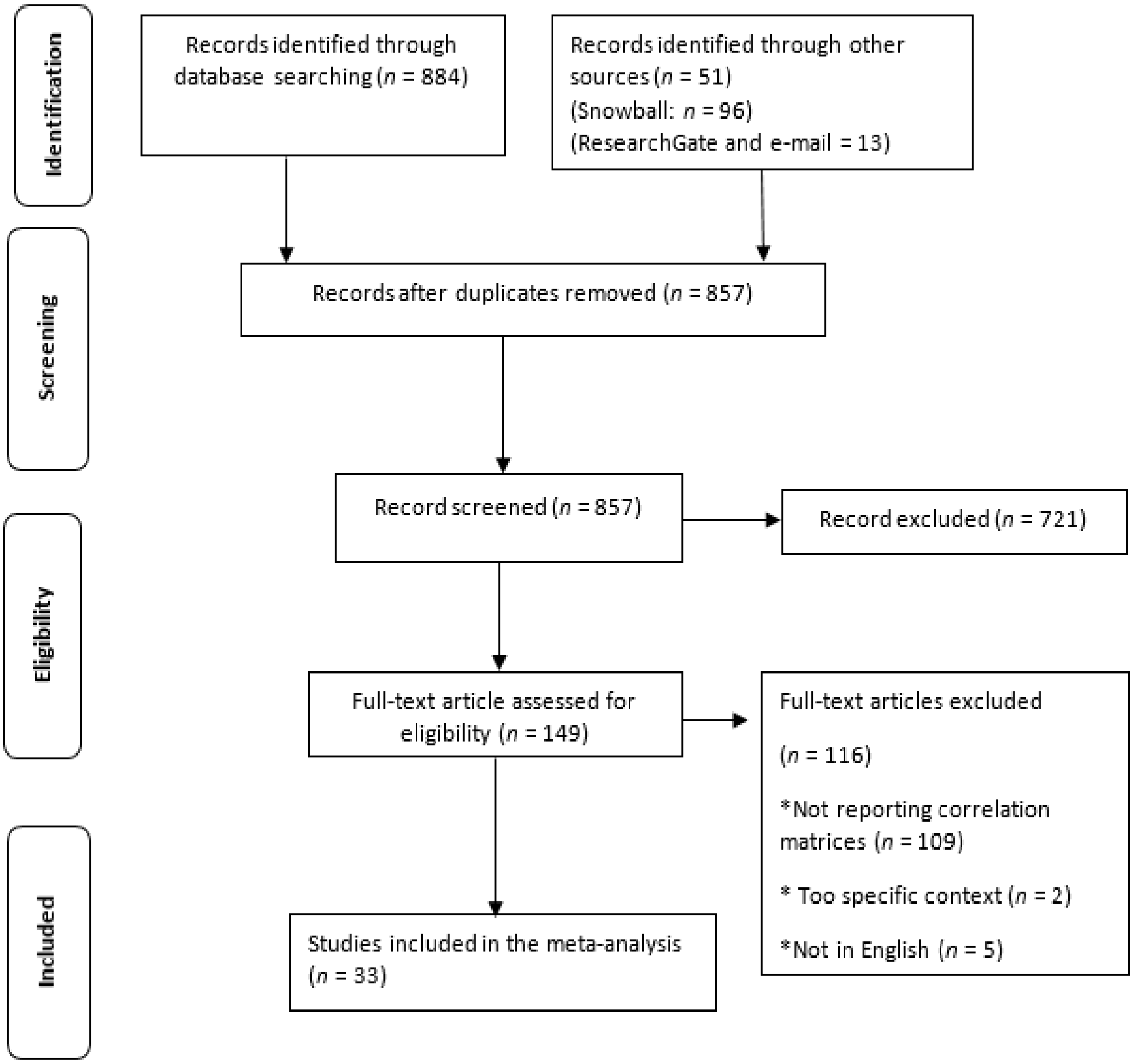
We searched Web of Science, PsycINFO, and Crossref databases to include eligible studies in the meta-analysis. During the scanning, the keywords “Satisfaction with Life Scale” and “SWLS” were scanned together and separately by two different researchers using the Publish and Perish (Harzing, 2007) software and Google Scholar search engine. We tried to determine the appropriate studies by examining (snowball technique) the reference parts of the studies reached. We contacted the study authors via ResearchGate and e-mail for candidate studies whose inter-item correlation was not available but which could potentially be used in the study.

Inclusion Criteria

All studies examining the validity and reliability of the “Satisfaction with Life Scale” are potential candidates for this meta-analysis study. For a potential candidate study to be included in this study, it should meet the following criteria: *(i)* the participants are drawn from a large population (students, adults, society, etc.), *(ii)* the correlation coefficients between the items of the scale or the information necessary for the calculation of these coefficients (covariance coefficients and standard deviations) are reported or obtained by contacting the author, and *(iii)* published in English. We reviewed in detail 35 studies that were planned to be included in the study. As a result of this review, we did not include two studies. The first one was the study by Garcia-Castro et al. (2022) because it conducted with a limited population, such as participants with clinical mental illness and the second one was the study by Silva et al. (2018) because the language of the report was not in English. As a result, we included 33 studies that met the stated inclusion criteria in this meta-analysis. A flowchart of the process of including primary studies in the meta-analysis is given in Figure 1.

Figure 1

PRISMA Flow Diagram


Identification
Screening
Eligibility
Included
Records identified through database searching (n = 884)
Records identified through other sources (n = 51)
(Snowball: n = 96)
(ResearchGate and e-mail = 13)
Records after duplicates removed (n = 857)
Record screened (n = 857)
Record excluded (n = 721)
Full-text article assessed for eligibility (n = 149)
Full-text articles excluded
(n = 116)
*Not reporting correlation matrices (n = 109)
* Too specific context (n = 2)
*Not in English (n = 5)
Studies included in the meta-analysis (n = 33)

Characteristics of Included Studies

We coded the 33 studies included in the meta-analysis according to sample size, age range, characteristics of the participants, and the country where the study was conducted (see Table 1). We divided the participants into three categories (Children/Youth/Adult) according to their mean age. The mean age of the two studies was not reported (Mishra, 2019; Tomás et al., 2015) . When the country where the studies were conducted was examined, 19 studies were conducted in different countries, while some studies were conducted in the same settings (e.g., Balgiu et al., 2021; Macovie et al., 2020; Cazan, 2014; Marcu, 2013). In three studies, the sample consisted of more than one country (Berrios-Riquelme et al., 2021; Esnaola et al., 2021; Tucker et al., 2006).

Two separate coders coded the item correlation matrices that we extracted from the 33 studies included in the meta-analysis. Inter-coder reliability was calculated as 100%.We determined that in four of the studies (Jovanović, 2019; Silva et al., 2015; Wang et al., 2017; Wu & Yao, 2006; Wu et al., 2009) more than one independent group was examined. Therefore, more than one correlation matrix was extracted from these studies. In the study by Wu et al. (2009), we determined that repeated measurements were made with two independent groups at different times. Therefore, we extracted initial item correlation matrices for each independent group from this study. In some studies, (Esnaola et al., 2017; Tucker et al., 2006; Wu & Yao, 2006), covariance matrices have been reported instead of correlation matrices. We calculated the correlation coefficients using covariance matrices and standard deviations of the scores for these studies. As a result, we obtained 41 independent item correlation matrices from a total of 49316 participants. The characteristics of the studies included in the meta-analysis is given in Table 1 and the correlation matrices extracted from the studies are given as Supplemental Material.

As seen in Table 1, the total sample size of 33 primary studies included in the meta-analysis was 49316. The age range of the individuals participating in the studies was quite wide and ranged from seven to 80 years old. In the classification made according to the mean age of the participants, 45.4% (*f*=15) were classified as adults, 36.3% (*f*=12) as youth, 6% (*f*=2) as children, and 6% (*f*=2) as mixed. When the countries in which the studies were carried out were examined, it was seen that four studies were conducted in Romania, three in Greece, two in Portugal, Taiwan, India, Colombia and U.K., and the rest in different countries (e.g. Turkey, Canada, Angola, Spain, Peru).

Table 1

Features of the studies included in the present meta-analysis

Authors	Valid *N*	Age range (*M*/*SD*)	Participant	Country
1. Areepattamannil & Bano (2020)	402	– (15.93/0.7)	Youth	India
2. Bacro et al. (2020)	557	8–16(11.12/1.65)	Children	French
3. Caycho-Rodríguez et al. (2018)	236	– (72.8/6.9)	Adult	Peru
4. Cazan (2014)	342	– (20/–)	Youth	Romania
5. Dirzyte et al. (2021)	2003	– (50.67/17.46)	Adult	Lithuania
6. Espejo et al. (2022a)	1255	18–67 (25.62/8.60)	Adult	Colombia
7. Esnaola et al. (2017)**	701	– (14.93/1.83)	Youth	Mixed
8. Gadermann et al. (2010)	1233	9–14 (11.7/–)	Children	Canada
9. Galanakis et al. (2017)	1797	18–67 (38.06/14.12)	Adult	Greece
10. Jovanović (2019) *	2595	14–55 (23.79/9.71)	Youth	Serbia
11. López-Ortega et al. (2016)	13220	50– (64.7/–)	Adult	Mexica
12. Marcu (2013)	285	16–60 (27.35/13.23)	Adult	Romania
13. Mishra (2019)	426	15–70 (–/–)	–	India
14. Moksnes et al. (2014)	1073	13–18 (15/1.62)	Youth	Norway
15. Sancho et al. (2014)	1003	60–90 (73.1/8.8)	Adult	Portugal
16. Silva et al. (2015)*	885	12–21 (17.7/–)	Youth	Portugal
17. Tomás et al. (2015)	5630	14–65 (–/–)	–	Angola
18. Tucker et al. (2006)**	277	17–62 (29.65/–)	Adult	Mixed
19. Wang et al. (2017)*	2178	15–25 (19.32/–)	Youth	China
20. Wu et al. (2009)*	237	15–23 (19.62/1.29)	Youth	Taiwan
21. Balgiu et al. (2021)	200	– (24.03/0.84)	Youth	Romania
22. Macovei (2020)	124	– (20/–)	Youth	Romania
23. Wu & Yao (2006)**	476	18–30 (20.04/1.67)	Youth	Taiwan
24. Anthimou et al. (2021)	341	17–44 (21.63/3.64)	Youth	Greece
25. García-Castro et al. (2022)	7790	– (66.88/9.58)	Adult	U.K.
26. Theodoropoulou (2021)	360	18–65 (23.54/5.96)	Youth	Greek
27. Berríos-Riquelme et al. (2021)	662	– (34.5/–)	Adult	Mixed
28. Singh et al. (2021)	400	21–60 (35.73/6.28)	Adult	India
29. M^a et al. (2021)	199	– (37.53/12.78)	Adult	Spain
30. Lang & Schmitz (2020)	641	9–18(12.44/1.56)	Children	Germany
31. Sagar & Karim (2014)***	210	19–58 (33.48/8.06)	Adult	Bangladesh
32. Espejo et al. (2022b)***	1222	– (25.66/8.66)	Adult	Colombia
33. Kjell & Diener (2021)***	343	– (34.4/11.9)	Adult	U.K.

* Correlation matrix belonging to more than one independent group was extracted.

** The correlation matrix was calculated by subtracting the covariance coefficients and standard deviations from these studies.

*** Correlation coefficients for the first three items were extracted from these studies.

Evaluating the Quality of Studies

We used the "Quality Assessment Checklist for Survey Studies in Psychology" developed by Protogerou and Hagger (2020) to determine the quality of the studies included in the meta-analysis. This checklist evaluates the study included in the meta-analysis according to 20 criteria in terms of introduction (rationale-variables), participants (sampling), data (Collection-measure-analysis-result-discussion), and ethics (consent-briefing-funding). These criteria vary according to the nature of the study (experimental or cross-sectional). According to this checklist, one point is awarded when these criteria are present in a study and zero when not. The quality score is determined by the ratio of the total score of the study to the total number of criteria. Protogerou and Hagger (2020) reported that studies with a quality score of 70% and above have acceptable quality. Two separate coders scored the 20 studies included in the meta-analysis. As a result of this evaluation, we determined that the quality of the studies varied between 70% and 90%, except for one study (Marcu, 2013). We calculated the quality value of Marcu's (2013) study as 59%.

Data Analysis

We used a two-stage structural equation modeling approach (Cheung, 2015; Cheung & Chan, 2005) to test the magnitude of the inter-item relationship and the goodness-of-fit of the established models using 41 correlation matrices extracted from the 33 studies included in the meta-analysis. We used the codes written by Jak (2015) to prepare the data for analysis. In the first step of this approach, we created a pooled correlation matrix using correlation matrices. In the second stage, we built structural equation models using this correlation matrix and examined the model-data fit of these models. In the first step, we combined the correlation matrices according to the random-effects model, as we collected them from the literature (Borenstein et al., 2009). However, we still examined the heterogeneity test. The fact that the heterogeneity test is statistically significant shows that the studies are heterogeneous and therefore the correlation matrices should be created according to the random effects model (Cheung, 2015). At this stage, we determined the effect size of the relationship between the items and the magnitude of the heterogeneity (I^2). We interpreted the calculated correlation effect sizes as “small” up to 0.1, “medium” up to 0.2, “large” up to 0.3, and “very large” up to 0.4 or greater (Funder & Ozer, 2019). Although not an absolute measure, we evaluated the magnitude of heterogeneity as “small” up to 25%, moderate up to 50%, and “high” up to 75% (Higgins et al., 2003).

We implemented the two-stage structural equation modeling approach using the *R software* Version 4.1.2 (R Core Team, 2021) and *metaSEM R-package* Version 1.2.5.1 (Cheung, 2015). However, we first built the models with *lavaan R-package* Version 0.6-10 (Rosseel, 2012) and then we converted these models into RAM (Responsibility Assignment Matrix) to create asymmetric (A matrix) and symmetric (S matrix) matrices. Thus, by performing the second stage of the analysis, we examined the model-data fit of the models. We examined RMSEA, SRMR, CFI, and TLI to determine if the model-data fit was achieved. RMSEA ≤ .05, SRMR ≤ .05, CFI ≥ .95, and TLI ≥ .95 were considered as “good” model-data fit (Hu & Bentler, 1999). We examined the AIC and BIC values to determine which of the established models fitted to the data, and we determined the model with the smaller values was the best model (Schermelleh-Engel et al., 2003).

In addition to AIC and BIC values in model evaluation, we examined the composite reliability (CR), convergent validity (CV), and divergent validity (DV-only for the third model) evidence suggested by Fornell and Larcker (1981) for each measurement model obtained from large populations via MASEM. Convergent validity is a concept that expresses the relationship between the items and the factor (Yaşlıoğlu, 2017). We obtained CR, Average Variance Extracted (AVE), and Maximum Shared Variance (MSV) values by using estimated standardized loading and error variances of measurement models and correlation between factors in multifactorial structures. CR is calculated by dividing the square of the sum of the standardized loadings by the sum of the squared of the standardized loadings and the error variances (Raykow, 1997). We calculated the AVE by dividing the sum of the square of the factor loading to the number of items. We obtained the MSV by squaring the relationship between the factors for the two-factor structural equation model. In order to assess the relevance of the evidence obtained, we used the criteria CR > 0.70 and AVE < CR with AVE values of 0.5 and above for convergent validity, and MSV< AVE for divergent validity (Esposito Vinzi et al., 2010; Hair et al., 2013). We used Egger's regression test to determine whether the raw correlation coefficients are the product of publication bias. This test is used to examine whether the asymmetry in the funnel plot is statistically significant (Egger et al., 1997). The statistically non-significance of the result of this test is accepted as a finding that there is no publication bias.

Results

Summary of the Effect Sizes of Inter-Item Correlations

In the first step of the two-stage MASEM, we combined the correlation matrices collected from the literature according to the fixed-effect model. As a result of this process, we calculated $Q_{(379)} = 8755.8$, ($p < .001$). This result shows that the studies are heterogeneous. Moreover, both the RMSEA (0.136) and SRMR (0.133) are very large, indicating that homogeneity of correlation matrices did not fit the data well. As the assumption of the homogeneity of the correlation matrices has not been met, we group the studies into clusters based on the study's sample type. If the correlation matrices are homogeneous within the subgroups, the grouping variable may be used to explain the heterogeneity (Cheung, 2015, p.233). In order to determine the source of heterogeneity, we divided the studies included in the meta-analysis into two as "youth" and "adult" according to the mean age of the sample groups. In the context of this study, we combined "children" category with the "youth" category, since the youngest mean age was 11 and there were relatively few studies (n = 2) in this category. We also classified the sample groups whose mean age was over 25 as the "adult" category. Because when the number of studies falling into a category is too small, the test is not powerful enough to reject the null hypothesis of the homogeneity of correlation matrices (Cheung, 2015). The results of the test are given in Table 2.

Table 2

Summary of subgroup analysis

Cluster	Sample	$\chi2$(df)	p	RMSEA [95% CI]	SRMR	TLI	CFI
Youth	25079	3621.68(220)	.000	.119 [.116, .123]	.121	.917	.920
Adult	24237	4353.42(149)	.000	.145 [.141, .149]	.129	.921	.927

According to Table 2, the test statistics and goodness-of-fit indices showed that the hypothesis of the homogeneity of the correlation matrices in these two samples was rejected. In other words, the sample type is not sufficient to explain the heterogeneity of the correlation matrices. For this reason, it is more appropriate to combine the effect sizes of the studies according to the random-effects model rather than the fixed-effect model. We calculated $Q_{(379)} = 7150.81$, ($p < .001$) according to random-effects model. This result also shows that the studies' correlation matrices are heterogeneous. The correlation effect sizes between the items that we obtained as a result of the first stage are given in Table 3.

Table 3

Summary of effect sizes of inter-item correlations (N = 49316)

Associations	***r***	**95% Confidence Interval**		***I*²**
		Lower Limit	**Upper Limit**	
I1&I2	.600	.566	.634	.963
I1&I3	.620	.586	.655	.962
I1&I4	.523	.494	.553	.925
I1&I5	.488	.500	.527	.954
I2&I3	.643	.610	.677	.967
I2&I4	.511	.476	.545	.950
I2&I5	.454	.424	.484	.923
I3&I4	.584	.553	.616	.952
I3&I5	.521	.493	.549	.913
I4&I5	.582	.475	.529	.908

As seen in Table 2, the calculated 10 correlation effect sizes vary between .424 and .629. It can be said that these effect sizes are at a "moderate to high" level according to the Funder and Ozer (2019) classification. These effect sizes indicate a very large and potentially very strong effect size in the short and long term. When the lower and upper limits of the correlation effect sizes are examined, it is seen that all of them are statistically significant (p < .05) at the 95% confidence interval. When the heterogeneity of the correlations between the items is examined (I2), it is seen that it varies between 92.3% and 96.6%. In this case, we can say that the heterogeneity of the correlations between the items is at a “high” level according to the Higgins et al. (2003) classification. Therefore, it can be interpreted that the variance between studies is due to the characteristics of the studies (sample group, measurement tool, etc.) other than sampling error.

We performed Egger's regression test to determine whether the raw correlation coefficients were the product of publication bias. As a result of the test, we determined that the asymmetric distributions of the correlation coefficients between the first item and the second item (I1&I2) in the funnel plot were statistically significant (p<.05), but statistically insignificant (p>.05) in other correlations. According to this results, it can be said that the correlations between the first item and the second item may be the product of publication bias.

Model-Data Fit of Models

We examined the model-data fit of three models in the second stage of the two-stage MASEM. The indices related to the model-data fit of the tested models are given in Table 4.

Table 4

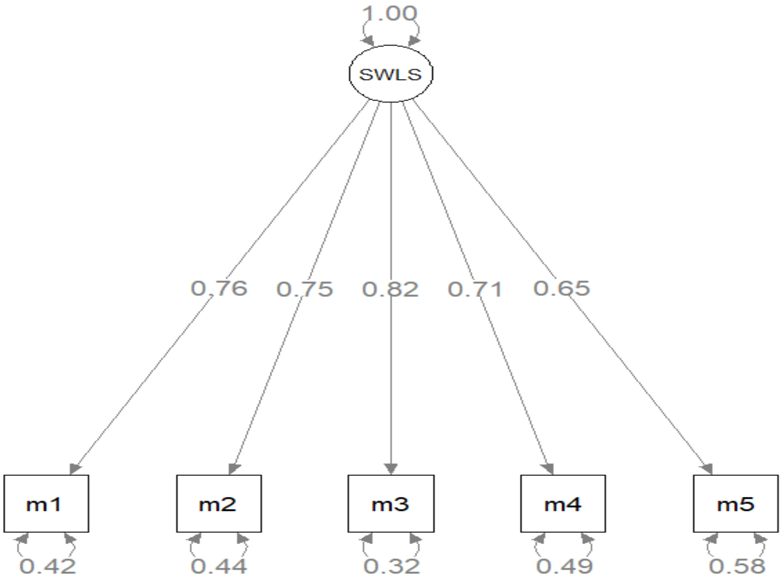
Summary of the indexes of the goodness-of-fit for the tested models (N = 49316)

Models	***χ2*(*df*)**	***p***	**RMSEA [95% CI]**	**SRMR**	**TLI**	**CFI**	**AIC**	**BIC**
Model 1	23.630(5)	.000	.009 [.005, 0.012]	.023	.996	.998	13.63	−30.40
Model 2	5.420(4)	.247	.003 [.000, .008]	.012	.999	.999	−2.58	−37.81
Model 3	5.420(4)	.247	.003 [.000, .008]	.012	.999	.999	−2.58	−37.81
Model 4	.000(0)	.000	.000 [.000, .000]	.000	-Inf	1.00	0.00	0.000

Model 1 (the original model) has a single factor and five-item structure. As seen in Table 4, the chi-square test was statistically significant for the Model 1's 5 degrees of freedom ($\chi 2$ = 23.650, p < .05). The RMSEA (.009, 95% CI [.005, .012]), SRMR (.023) TLI (.996), and CFI (.998) values indicated a good fit. When these indices are evaluated together, we can say that the Model 1 fits to the data. The path diagram for Model 1 is given in Figure 2.

Figure 2

The path diagram for Model 1

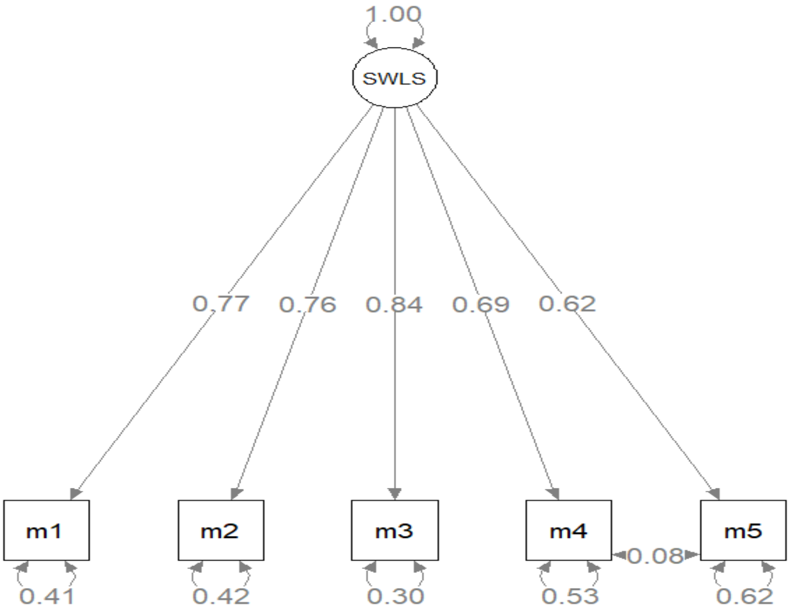


As seen in Figure 2, the standardized factor loadings (effect indicators) of the scale items in a single factor structure vary between .650, 95% CI [.624, .670] and .820, 95% CI [.800, .849]. The error variances of the scale items ranged from .330 to .580. When these variances are taken into account, the variance explanation rate of the latent variable of life satisfaction in the scale items varies between 42% (fifth item) and 68% (third item). When the inter-item residual covariance matrix was examined, we determined that the covariance ranged between -.005 (m1&m3) and .004 (m4&m5). Using standardized loadings and error variances, we calculated CR as .857. This result shows that the model is reliable. In addition, we calculated the AVE as .547. According to these results, we can say that this model satisfies the convergent validity conditions.

The second model is the model (Model 2) in which the correlation between the fourth and fifth items is established. As seen in Table 4, the chi-square test was not statistically significant with the 4 degrees of freedom of the Model 2 (modified original model) ($\chi 2$ = 5.419, *p* > *.05*). According to the results, the RMSEA (.003, 95% CI [.003, .008]), SRMR (.012), TLI (.999), and CFI (.999) values were at a “good” fit level. When these indexes are evaluated together, we can say that the modified original model fits to the data. The path diagram for the Model 2 is given in Figure 3.

Figure 3

The path diagram for Model 2

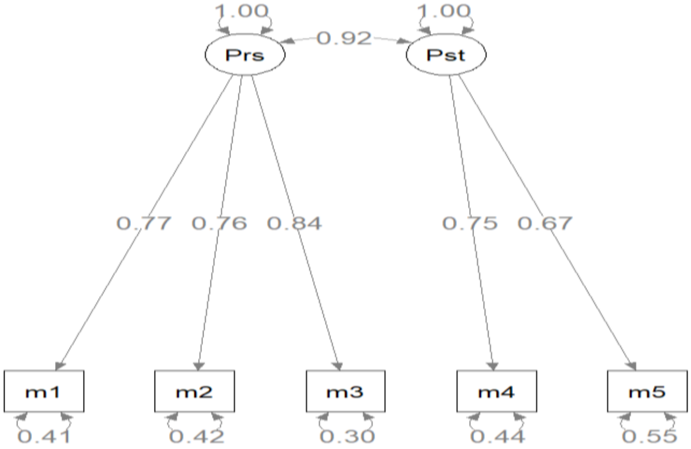


As seen in Figure 3, the standardized factor loadings of the scale items in the modified single factor structure range from .620, 95% CI [.591, .643] to .840, 95% CI [.813, .864]. The error variances of the scale items ranged from .300 to .620. When these variances are taken into account, the variance explanation rate of the latent variable of life satisfaction in the scale items varies between 38% (fifth item) and 70% (third item). The correlation between the fourth item and the fifth item is statistically significant (*p* < *.05*) and small (.08, 95% CI [.043, .115]). When the inter-item residual covariance matrix was examined, we determined that the covariance ranged between -.023 (m1&m3) and .017 (m1&m2). We calculated CR .857 and AVE .547 as additional construct validity evidence for Model 2. These results show that the scale based on this model has a high internal consistency.

In Model 3, the first three items were placed under the present factor (Prs), and the fourth and fifth items were placed under the past factor (Pst). As seen in Table 4, the chi-square test for 4 degrees of freedom of the Model 3 (two-factor model) was not statistically significant ($\chi 2$ = 5.418, *p* > *.05*). The RMSEA (.003 95% CI [.000, .008]), SRMR (.012), TLI (.999), and CFI (.999) values showed good fit. It should be noted that the goodness-of-fit indexes obtained for the two-factor model are the same as for the modified original model. This is because the two-factor model is mathematically equivalent to the modified one-factor model (Jovanović, 2019). The correlation between the first factor (Prs) and the second factor (Pst) was calculated as .920, 95% CI [.882, .954]. This correlation coefficient shows that there is a high level of relationship between the two factors. The path diagram for the two-factor model is given in Figure 4.

Figure 4

The path diagram for Model 3

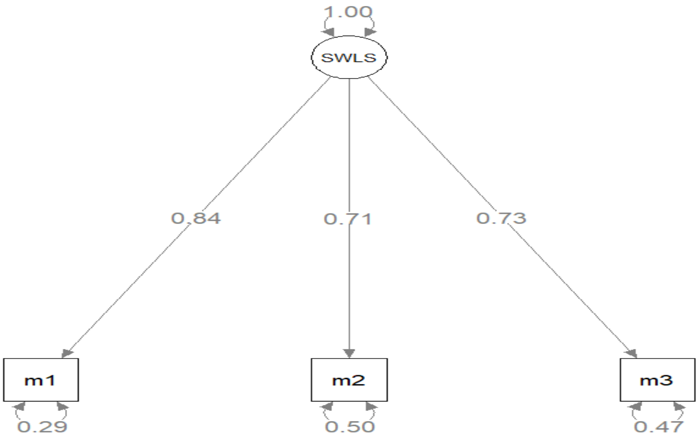


As seen in Figure 4, the standardized factor loadings of the scale items in the two-factor structure range from .670, 95% CI [.647, .698] to .840, 95% CI [.813, .864]. The error variances of the scale items ranged from .300 to .550. When these variances are taken into account, the variance explanation rate of the latent variables of life satisfaction in the scale items varies between 45% (fifth item) and 70% (third item). When the inter-item residual covariance matrix was examined, we determined that the covariance ranged between -.023 (m1&m3) and .017 (m1&m2). The CR and AVE values were .833 and .625 for the first factor but .738 and .585 for the second factor, respectively. According to these results, the conditions for CR and CV were met for both present and past factors. However, another important proof of construct validity calculated based on CFA findings in multifactorial constructs is DV. We calculated the MSV as .846 to obtain evidence for divergent validity. Since this value was greater than the AVE values calculated on the basis of the factor (MSV > AVE), the construct validity became problematic and divergent validity could not be achieved.

In Model 4, we removed the past factor and examined the structure consisting of only the three-item present factor. As seen in Table 4, the chi-square test was statistically significant for the Model 4's 0 degrees of freedom ($\chi 2$ = .000, $p < .05$). As such, the RMSEA (.000, 95% CI [.000, .000]) and SRMR (.000) values indicated good fit. TLI was infinite, and CFI was at a "good" fit with a value of 1.000. This model fits to the data well. The path diagram for Model 5 is given in Figure 5.

Figure 5

The path diagram for Model 4



As seen in Figure 5, the standardized factor loadings of the scale items in a single factor structure vary between .710, 95% CI [.672, .749] and .840, 95% CI [.798, .888]. The error variances of the scale items ranged from .290 to .500. When these variances are taken into account, the variance explanation rate of the latent variable of life satisfaction in the scale items varies between 53% (third item) and 71% (first item). When the inter-item residual covariance matrix was examined, we determined that the covariance ranged between -0.5e-12 (m1&m3) and 1.2e-12 (m2&m3). Using standardized loading and error variances, we calculated the CR value of .868. This result shows that the scale based on this model are consistently measuring the underlying construct. In addition, we calculated the AVE as .604. According to these results, we can say that this model satisfies the CV conditions. However, the goodness-of-fit indexes of the measurement model in Table 3 show that this model is a saturated model. A saturated model has the best fit possible, since it perfectly reproduces all of the variances, covariances, and means (Maruyama, 1997). Since all parameters are calculated in saturated models, these parameters perfectly reflect the covariance matrix of the sample (Sümer, 2000). That's why the saturated model above has a chi-square of zero with zero degrees of freedom. It does matter because the model fit cannot be tested without free *df* and because estimation might fail, but that also depends on the model and data. Thus, this saturated model will produce a GFI of 1, AGFI of 1, CFI of 1 and AIC of 0, BIC of 0, but this is because of the saturated nature of the model and not an indicator of the real perfect fit. Saturated model has no testable implications. Therefore, it is not possible to compare the statistically obtained findings for model 4, considered in this study, with other models. Also, there is no way to know or learn whether Model 4 structure with its directed arrows is correct, partially correct, or complete nonsense.

Which model is the best fit to the data among the three models? AIC (Akaike's Information Criterion) and BIC (Bayesian Information Criterion) indexes are used to determine the fittest model for the data. It is concluded that the model with lower indices than the competitor model is the best fit to the data (Schermelleh-Engel et al., 2003). However, in some cases, while the AIC index of a model is lower than the competitor model, the BIC index may be higher. In such cases, Hyndman and Athanasopoulos (2013) say that using the AIC index is more appropriate than the BIC index. They argue that many statisticians choose the BIC index if there is a real model to test, whereas real models are rare, and even if there is, the model chosen based on the BIC index will not give the best estimate. Accordingly, when the AIC values of Model 1 (original model) and Model 2 (modified-original model), and Model 3 (two-factor model) are compared, the AIC values of Model 2 and Model 3 (−4.980) are lower than that of Model 1 (.680). However, it should be examined whether the decrease in chi-square is statistically significant in choosing the best model (Kline, 1998). The chi-square value of the original model was 10.680 with five degrees of freedom, while the chi-square value of the modified-original model (also in the two-factor model) decreased to 3.020 with four degrees of freedom. As the chi-square value increases, the fit of an overidentified model worsens (Kline, 1998). The difference between the chi-square values of the two models is 7.660 with one degree of freedom. The chi-square table value is 3.840 at the 95% confidence level. Therefore, this decrease in chi-square value is statistically significant (p < .05). Similarly, the RMSA and SRMR values of the modified-single factor model (also the two-factor model) are lower than the original model, and the CFI and TLI values are higher than the original model. Moreover, the residual covariance values of Model 2 are lower than Model 1. As it is known, as the residual covariance of a model decreases, the power of the model to explain the variation in the data increases. When all these values are evaluated together, it can be concluded that the modified model and the two-factor model fit better than the original model. We also calculated the CV for Model 2 as .850. This value is higher than the Cronbach alphas calculated by Vassar (2008) and Busseri (2018) and is consistent with the values reported by Pavot and Diener (1993, 2008). Accordingly, it can be said that the construct reliability of the scale has high internal consistency.

Discussion and Conclusion

The purpose of present study is to determine the most appropriate factor structure for the SWLS. As a result of the analysis we concluded that the modified model and the two-factor model fit better than the original model. When the studies on the factorial structure of the SWLS scale in the literature were examined (Clench-Aas et al., 2011; Hultell & Gustavsson, 2008; Jovanovic, 2016; Jovanovic, 2019; Slocum-Gori et al., 2009; Wu & Yao, 2006; Vautier et al. , 2004), it was observed that the model-data fit of the single-factor and two-factor structures of the scale was generally confirmed. Just like in this study, Jovanovic (2019) emphasized in his study that Model 2 and Model 3 obtained mathematically equivalent statistical values. In this study, the correlation between the two-factor structure was reported to be quite high. Clench-Aas et al.(2011) support a single-factor solution for SWLS with 74% of the variance explained by a single factor. The study also confirmed that the last two items tended to load on the second less important factor reflecting past experiences. They found a high (r = .930) correlation between factors in the two-factor model. In their study, they emphasized the single-factor structure in which error covariances were associated between the two modified items on the grounds that this correlation was evidence that the two factors could not be easily distinguished and that the single-factor structure overlapped with the theory. Similarly, Sachs (2003) compared models on the SWLS Hong Kong Chinese version and again determined the correlation between the factors of Model 3 as "high" (r = .720) and recommended Model 2. He also based the theory of this proposal as "individuals' current experiences cannot be independent of their past experiences in life satisfaction, past experiences also shape their present lives, and therefore life satisfaction will be a general factor formed by the sum of past and present experiences."

Vauiter et al. (2004) argued that the 5 items of the SWLS should be considered on a single main factor in which the sequential effect is taken into account, rather than the scattered positioning of the 5 items on two different factors, and the overall results should be evaluated with a single dimension since the last items perhaps refer to past achievements rather than current conditions. In the literature, the necessity of investigating and understanding the solutions and existing inconsistencies regarding one- and two-factor models has been emphasized not only from a psychometric point of view but also from a cultural point of view (Diener & Diener, 1995; Oishi, 2002; Oishi & Diener, 2001; Oishi & Diener, 2003). For example, Oishi (2006) found in his study that the items 4 and 5 were identified as different across Chinese and American samples. Accordingly, Chinese participants, unlike American participants, did not endorse of items 4 and 5. One of the reasons, Oishi (2006) argues, is that Chinese's concept of life satisfaction is primarily based on external conditions and the current situation rather than on past achievements. These results may indicate that the structure of the scale differs according to the culture groups. In support of this result, Emerson et al. (2017) determined in their review study that SWLS was rarely invariant beyond the structural measurement invariance (MI) level among cultural groups. Emerson et al. (2017) recommends caution when interpreting cross-cultural comparisons, as the meanings of scale items may change during the translation process and most cultural invariance analyses involve comparisons between different language versions of the SWLS. For this reason, although different language versions of the SWLS consistently preserve the meanings of clauses in translation, factor analysis results cannot be generalized to populations from different cultural backgrounds, and people from different cultures may have different definitions, perceptions, and interpretations of SWLS. In this context, it was revealed that measurement invariance studies carried out in different subgroups were primarily investigated on this factorial inconsistency and that the results obtained from the multiple sample analysis based on the single-factor model had strict factorial invariance (Wu & Yao, 2006).

When the models are compared within the framework of the calculated reliability and validity criteria, highest structural reliability belongs to Model 4. When the AVE values for the calculated CR were examined, the structure corresponding to the present structure in Model 3 (which is the same structure as the model tested in Model 4) had the highest explained mean of variance. Also, Model 4 had the second highest AVE value. It is noteworthy that approximately 60% of the structure to be measured in Model 4 is measured with these three items. As expected, these results coincide with the values of the present factor in the two-factor structure. Unlike the Model 3 present structure, the interesting situation in Model 4 is that the standardized load value for item 1 increases and the error variance in this item decreases. In addition, the inter-item residual covariance of this model is close to zero. This finding is in line with the study by Hanzlová (2022) on the 5-item single-factor (Model 1), 4-item single-factor and 3-item single-factor (Model 4) constructs of the SWLS scale. In the study conducted by Hanzlová (2022), the explained total variance value of the 3-item single-factor model was 78%, the Cronbach alpha reliability value was .860 despite the decrease in the number of items, and for all models, the test information functions drawn in line with the item response theory showed that there was enough information in terms of the feature to be measured confirms the findings of the current study. The increase in the total variance explained despite the decrease in the number of items is an indicator of the increase in the representation power of the structure on which the remaining items are focused. As a matter of fact, studies of MI (Espejo et al., 2022; Kjell & Diener, 2021) also confirm this finding and indicate that when Model 1 is used, the meaning attributed to the relevant variable changes and indicates that it causes bias in different subgroups. In these studies, attention was drawn to the fact that the first three models met the measurement of ideal life and perfect conditions, and it was suggested to use model 4 on the grounds that the measurement invariance only met the conditions of partial invariance when other items were added, and the comparisons made lost their meaning.

Implication

We determined that the proposed one-factor modified model (model 2), based on the findings of this study, is largely compatible with the theory and is more understandable for researchers. Suggesting a generally accepted model for SWLS, which has been the subject of many primary structural equation modeling studies in the literature, and the fact that this model is obtained from a very large population can put an end to the debate about the best-fitting model. It is a known fact that SEM-based studies need large sample sizes. Although the primary studies included in the meta-analysis were carried out in relatively large groups, they are rather weak compared to the sample sizes reached by MASEM and can produce different results. Thanks to MASEM, the data obtained from all these studies came together and allowed generalization to be made in the choice of the most suitable model. The researchers' use of this model in studies examining the structural validity of SWLS (e.g., the adaptation of scale, invariance studies, and structural equation modeling studies) may help them obtain more valid and reliable results. Independent of the statistical implications for Model 4, the use of Model 4 can provide researchers with an alternative and useful way to conduct research in which a large number of psychological variables are addressed. Many studies (Kjell & Diener, 2021; Sandy et al., 2017; Ziegler et al., 2014) emphasize that short scales can be a solution in terms of facilitating applicability in different contexts and populations, and making measurements in people with low education level or people with cognitive problems. This can be seen as a fast, advantageous, and common way of obtaining data in research (Sandy et al., 2017).

One of the limitations of this study is that the correlation coefficient calculated between the first and second items of the scale may have been due to publication bias. Therefore, this should be taken into account when discussing the validity of the models. This may limit the validity of the indices obtained from the models. The second limitation is that in the context of this study, moderator analysis was performed on the subgroups formed according to the mean age (youth/adult) in order to determine the source of the variance. However, it was determined that these subgroups could not explain the variance.

Therefore, in future studies based on moderator analysis, the effect of different subgroups can be examined with a larger sample to explain the variance.

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

References marked with an asterisk (*) indicate studies included in the meta-analysis

Appleton, S., & Song, L. (2008). Life satisfaction in urban China: Components and determinants. *World Development, 36*(11), 2325-2340. https://doi.org/10.1016/j.worlddev.2008.04.009

*Anthimou, A., Koutsogiorgi, C., & Michaelides, M. (2021). Psychometric properties of the Satisfaction with Life Scale in a Cypriot student sample. *Psychologia*, *26*, 273–282. https://doi.org/10.12681/psy_hps.29152

Areepattamannil, S., & Bano, S. (2020). Psychometric Properties of the Satisfaction with Life Scale (SWLS) Among Middle Adolescents in a Collectivist Cultural Setting. *Psychological Studies*, *65*(4), 497–503. https://doi.org/10.1007/s12646-020-00578-4

Atienza, F. L., Pons, D., Balaguer, I., & García-Merita, M. (2000). Psychometric properties of the satisfaction with life scale in adolescents. *Psicothema*, *12*, 314–319. https://psycnet.apa.org/record/2000-03241-023

Bacro, F., Coudronnière, C., Gaudonville, T., Galharret, J.-M., Ferrière, S., Florin, A., & Guimard, P. (2020). The French adaptation of the Satisfaction with Life Scale (SWLS): Factorial structure, age, gender and time-related invariance in children and adolescents. *European Journal of Developmental Psychology*, *17*(2), 307–316. https://doi.org/10.1080/17405629.2019.1680359

Balgiu, B., Sfeatcu, R., Ilinca, R., & Bucur, V. (2021). Investigating Construct Validity of the Satisfaction with Life Scale in a Sample of Romanian Dental Students. *Romanian Journal of Oral Rehabilitation*, *13*, 7–12. http://rjor.ro/wp-content/uploads/2021/10

Berríos-Riquelme, J., Pascual-Soler, M., Frías-navarro, D., & Maluenda Albornoz, J. (2021). Psychometric properties and factorial invariance of the satisfaction with life scale in Latino immigrants in Chile, Spain, and United States. *Terapia Psicológica, 39*, 199–218. https://doi.org/10.4067/s0718-48082021000200199

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

Busseri, M. A. (2018). Examining the structure of subjective well-being through meta-analysis of the associations among positive affect, negative affect, and life satisfaction. *Personality and Individual Differences*, *122*, 68–71. https://doi.org/10.1016/j.paid.2017.10.003

Caycho-Rodríguez, T., Ventura-León, J., García Cadena, C. H., Barboza-Palomino, M., Arias Gallegos, W. L., Dominguez-Vergara, J., Azabache-Alvarado, K., Cabrera-Orosco, I., & Samaniego Pinho, A. (2018). Psychometric Evidence of the Diener's Satisfaction with Life Scale in Peruvian Elderly. *Revista Ciencias de La Salud*, *16*(3), 488. https://doi.org/10.12804/revistas.urosario.edu.co/revsalud/a.7267

Cazan, A.-M. (2014). The Romanian Version of the Satisfaction with Life Scale. *Romanian Journal of Experimental Applied Psychology*, *5*(1), 42–47. https://www.researchgate.net/profile/Ana-Maria-Cazan-2/publication/263768270

Cheung, M. W.-L. (2015). *Meta-Analysis: A Structural Equation Modeling Approach*. John Wiley&Sons.

Cheung, M. W.-L., & Chan, W. (2005). Meta-analytic structural equation modeling: A two-stage approach. *Psychological Methods*, *10*(1), 40–64. https://doi.org/10.1037/1082-989X.10.1.40

Clench-Aas, J., Nes, R., Dalgard, O., & Aarø, L. (2011). Dimensionality and measurement invariance in the Satisfaction with Life Scale in Norway. *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *20*, 1307–1317. https://doi.org/10.1007/s11136-011-9859-x

Dağli, A., & Baysal, N. (2016). Yaşam Doyumu Ölçeğinin Türkçe'ye Uyarlanması: Geçerlik Ve Güvenirlik Çalışması. *Elektronik Sosyal Bilimler Dergisi*, *15*(59), 59. https://doi.org/10.17755/esosder.263229

Dirzyte, A., Perminas, A., & Biliuniene, E. (2021). Psychometric Properties of Satisfaction with Life Scale (SWLS) and Psychological Capital Questionnaire (PCQ-24) in the Lithuanian Population. *International Journal of Environmental Research and Public Health*, *18*(5), 2608. https://doi.org/10.3390/ijerph18052608

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Egger M, Smith GD, Schneider M, Minder CBias in meta-analysis detected by a simple, graphical test. *BMJ (Clinical Research Ed.)*, *315*, 629–634. https://doi.org/10.1136/bmj.315.7109.629

Emerson, S. D., Guhn, M., & Gadermann, A. M. (2017). Measurement invariance of the Satisfaction with Life Scale: Reviewing three decades of research. *Quality of Life Research*, *26*(9), 2251–2264. https://doi.org/10.1007/s11136-017-1552-2

Esnaola, I., Benito, M., & Antonio-Agirre, I. (2017). Measurement invariance of the Satisfaction With Life Scale (SWLS) by country, gender and age. *Psicothema*, *29*(4), 596–601. https://doi.org/10.7334/psicothema2016.394

Espejo, B., Martín-Carbonell, M., & Checa, I. (2022). Psychometric Properties and Measurement Invariance by Gender of the Abbreviated Three-Item Version of the Satisfaction with Life Scale in a Colombian Sample. *International Journal of Environmental Research and Public Health*, *19*(5), 2595. https://doi.org/10.3390/ijerph19052595

Espejo, B., Martín-Carbonell, M., Checa, I., Paternina, Y., Fernández-Daza, M., Higuita, J. D., Albarracín, A., & Cerquera, A. (2022). Psychometric Properties of the Diener Satisfaction With Life Scale With Five Response Options Applied to the Colombian Population. *Frontiers in Public Health*, *9*, 767–534. https://doi.org/10.3389/fpubh.2021.767534

Esposito Vinzi, V., Trinchera, L., & Amato, S. (2010). *PLS Path Modeling: From Foundations to Recent Developments and Open Issues for Model Assessment and Improvement*. In Handbook of Partial Least Squares (pp. 47–82). https://doi.org/10.1007/978-3-540-32827-8_3

Fabio, A., & Gori, A. (2015). Measuring Adolescent Life Satisfaction: Psychometric Properties of the Satisfaction With Life Scale in a Sample of Italian Adolescents and Young Adults. *Journal of Psychoeducational Assessment*, *34*. https://doi.org/10.1177/0734282915621223

Fornell, C., & Larcker, D. F. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research*, *18*(1), 39–50. https://doi.org/10.2307/3151312.

Frisch, M. B. (1994). *Quality of life inventory manual and treatment guide*. NCS Pearson and Pearson Assessments.

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Gadermann, A. M., Schonert-Reichl, K. A., & Zumbo, B. D. (2010). Investigating Validity Evidence of the Satisfaction with Life Scale Adapted for Children. *Social Indicators Research*, *96*(2), 229–247. https://doi.org/10.1007/s11205-009-9474-1

Galanakis, M., Lakioti, A., Pezirkianidis, C., & Karakasidou, E. (2017). Reliability and Validity of the Satisfaction with Life Scale SWLS in a Greek Sample. *International Journal of Humanities & Social Studies*, *5*(2), 8. https://internationaljournalcorner.com/index.php/theijhss/article/view/125249

García-Castro, F. J., Bendayan, R., & Blanca, M. J. (2022). Measurement Invariance and Validity of the Satisfaction With Life Scale in Informal Caregivers. *Psicothema*, *34*(2), 299–307. https://doi.org/10.7334/psicothema2021.313

Glaesmer, H., Grande, G., Braehler, E., & Roth, M. (2011). 'The German version of the Satisfaction with Life Scale (SWLS): Psychometric properties, validity, and population-based norms': Correction to Glaesmer et al. (2011). *European Journal of Psychological Assessment*, *27*, 299–299. https://doi.org/10.1027/1015-5759/a000081

Gouveia, V. V., Milfont, T. L., da Fonseca, P. N., & Coelho, J. A. P. de M. (2009). Life Satisfaction in Brazil: Testing the Psychometric Properties of the Satisfaction With Life Scale (SWLS) in Five Brazilian Samples. *Social Indicators Research*, *90*(2), 267–277. https://doi.org/10.1007/s11205-008-9257-0

Hair, J., Ringle, C., & Sarstedt, M. (2013). Partial Least Squares Structural Equation Modeling: Rigorous Applications, Better Results and Higher Acceptance. *Long Range Planning*, *46*, 1–12. https://doi.org/10.1016/j.lrp.2013.08.016

Hanzlová, R. (2022). Measuring wellbeing in European Social Survey (ESS). *Conference: Well-Being 2022: Knowledge for informed decisions*. https://doi.org/10.13140/RG.2.2.25204.01921

Harzing, A.-W. (2007). Publish or Perish (8.2.3883.8074). https://harzing.com/resources/publish-or-perish

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ : British Medical Journal*, *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

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

Jak, S. (2015). *Meta-Analytic Structural Equation Modelling*. Springer International Publishing. https://doi.org/10.1007/978-3-319-27174-3

Jovanović, V. (2019). Measurement Invariance of the Serbian Version of the Satisfaction With Life Scale Across Age, Gender, and Time. *European Journal of Psychological Assessment, 35*(4), 555–563. https://doi.org/10.1027/1015-5759/a000410

Kjell, O. N. E., & Diener, E. (2021). Abbreviated Three-Item Versions of the Satisfaction with Life Scale and the Harmony in Life Scale Yield as Strong Psychometric Properties as the Original Scales. *Journal of Personality Assessment*, *103*(2), 183–194. https://doi.org/10.1080/00223891.2020.1737093

Lang, J., & Schmitz, B. (2020). German Translation of the Satisfaction With Life Scale for Children and Adolescents. *Journal of Psychoeducational Assessment*, *38*(3), 291–304. https://doi.org/10.1177/0734282919849361

López-Ortega, M., Torres-Castro, S., & Rosas-Carrasco, O. (2016). Psychometric properties of the Satisfaction with Life Scale (SWLS): Secondary analysis of the Mexican Health and Aging Study. *Health and Quality of Life Outcomes*, *14*(1), 170. https://doi.org/10.1186/s12955-016-0573-9

Ma, D., Merino, M., Privado, J., & Durán, R. (2021). Satisfaction with Life Scale (SWLS) adapted to work: Psychometric Properties of the Satisfaction with Work Scale (SWWS) Psychological and subjective well-being. *Anales de Psicología, 37*, 557–566. https://doi.org/10.6018/analesps.430801

Macovei, C. M. (2020). Psychometric Properties and Factor Structure of the Satisfaction with Life Scale in a Military Students Sample. *International Conference knowledge-based organization, 26*(2), 300–304. https://doi.org/10.2478/kbo-2020-0094

Marcu, R. (2013). New Psychometrical Data on the Efficiency of Satisfaction with Life Scale in Romania. *Journal of Psychological and Educational Research (JPER), 1*, 79–90. https://www.ceeol.com/search/article-detail?id=148635

Margolis, S., Schwitzgebel, E., Ozer, D.J., & Lyubomirsky, S.(2018). A New Measure of Life Satisfaction: The Riverside Life Satisfaction Scale. *Journal of Personality Assessment*, 1532-7752. https://doi.org/10.1080/00223891.2018.1464457

Mishra, K. K. (2019). Psychometric Evaluation of Hindi version of Satisfaction with Life Scale. *Mind and Society*, *8*(1&2), 30–34. https://mindandsociety.in/index.php/MAS/article/view/96

Moksnes, U. K., Løhre, A., Byrne, D. G., & Haugan, G. (2014). Satisfaction with Life Scale in Adolescents: Evaluation of Factor Structure and Gender Invariance in a Norwegian Sample. *Social Indicators Research*, *118*(2), 657–671. https://doi.org/10.1007/s11205-013-0451-3

Oishi, S. (2006). The concept of life satisfaction across cultures: An IRT analysis. *Journal of Research in Personality*, *40*(4), 411–423. https://doi.org/10.1016/j.jrp.2005.02.002

Protogerou, C., & Hagger, M. S. (2020). A checklist to assess the quality of survey studies in psychology | Elsevier Enhanced Reader. *Methods in Psychology*, *3*(2020), 1–14. https://doi.org/10.1016/j.metip.2020.100031

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing (4.1.2). https://www.R-project.org/

Raykov T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*, 173-184. . https://doi.org/10.1177/01466216970212006

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*, 1–36. https://doi.org/10.18637/jss.v048.i02

Sagar, M. M., & Karim, A. (2014). The psychometric properties of Satisfaction With Life Scale for police population in Bangladeshi Culture. *The International Journal of Social Sciences*, *28*, 24–31. https://www.academia.edu/9594002/The_psychometric_properties_of_Satisfaction_With_Life_Scale_for_police_population_in_Bangladeshi_Culture

Sancho, P., Galiana, L., Gutierrez, M., Francisco, E.-H., & Tomás, J. M. (2014). Validating the Portuguese Version of the Satisfaction With Life Scale in an Elderly Sample. *Social Indicators Research*, *115*(1), 457–466. https://doi.org/10.1007/s11205-012-9994-y

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research*, *8*(2), 23–74. https://www.stats.ox.ac.uk/~snijders/mpr_Schermelleh.pdf

Shin, D. C., & Johnson, D. M. (1978). Avowed happiness as an overall assessment of the quality of life. *Social Indicators Research*, *5*(1), 475–492. https://doi.org/10.1007/BF00352944

Silva, A. D., Taveira, M. do C., Marques, C., & Gouveia, V. V. (2015). Satisfaction with Life Scale Among Adolescents and Young Adults in Portugal: Extending Evidence of Construct Validity. *Social Indicators Research*, *120*(1), 309–318. https://doi.org/10.1007/s11205-014-0587-9

Singh, R., Gull, M., & Husain, A. (2021). Standardization of Diener's Satisfaction with Life Scale in Hindi. *International Journal of Indian Psychology*, *9*(4). https://doi.org/10.25215/0904.138

Slocum-Gori, S., Zumbo, B., Michalos, A., & Diener, E. (2009). A Note on the Dimensionality of Quality of Life Scales: An Illustration with the Satisfaction with Life Scale (SWLS). *Social Indicators Research*, *92*, 489–496. https://doi.org/10.1007/s11205-008-9303-y

Swami, V., & Chamorro-Premuzic, T. (2009). Psychometric evaluation of the Malay Satisfaction With Life Scale. *Social Indicators Research*, *92*, 25–33. https://doi.org/10.1007/s11205-008-9295-7

Theodoropoulou, E. (2021). Validity and reliability of the Greek version of the Satisfaction with Life Scale. *Global Journal For Research Analysis*. https://doi.org/10.36106/2715007

Tomás, J. M., Gutiérrez, M., Sancho, P., & Romero, I. (2015). Measurement invariance of the Satisfaction With Life Scale (SWLS) by gender and age in Angola. *Personality and Individual Differences, 85*, 182–186. https://doi.org/10.1016/j.paid.2015.05.008

Tucker, K. L., Ozer, D. J., Lyubomirsky, S., & Boehm, J. K. (2006). Testing for Measurement Invariance in the Satisfaction with Life Scale: A Comparison of Russians and North Americans. *Social Indicators Research*, *78*(2), 341–360. https://doi.org/10.1007/s11205-005-1037-5

Vassar, M. (2008). A note on the score reliability for the Satisfaction With Life Scale: An RG study. *Social Indicators Research*, *86*, 47–57. https://doi.org/10.1007/s11205-007-9113-7

Vittersø, J., Oelmann, H., & Wang, A. (2009). Life Satisfaction is not a Balanced Estimator of the Good Life: Evidence from Reaction Time Measures and Self-Reported Emotions. *Journal of Happiness Studies*, *10*, 1–17. https://doi.org/10.1007/s10902-007-9058-1

Wang, D., Hu, M., & Xu, Q. (2017). Testing the factorial invariance of the Satisfaction with Life Scale across Chinese adolescents . *Social Behavior and Personality: An International Journal*, *45*(3), 505–516. https://doi.org/10.2224/sbp.6222

Wu, C., & Yao, G. (2006). Analysis of factorial invariance across gender in the Taiwan version of Satisfaction With Life Scale. *Personality and Individual Differences*, *40*, 1259–1268. https://doi.org/10.1016/j.paid.2005.11.012

Wu, C.-H., Chen, L. H., & Tsai, Y.-M. (2009). Longitudinal invariance analysis of the satisfaction with life scale. *Personality and Individual Differences*, *46*(4), 396–401. https://doi.org/10.1016/j.paid.2008.11.002

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(4); 472-491

Appendix

Raw Correlation considered for meta-analysis

Research ID	Sample Size	I1&I2	I1&I3	I1&I4	I1&I5	I2&I3	I2&I4	I2&I5	I3&I4	I3&I5	I4&I5	Mean of the Items				
												I1	I2	I3	I4	I5
1.Areepattamannil & Bano (2020)	417	.570	.590	.470	.510	.690	.610	.640	.570	.550	.500	4.28	4.58	4.69	4.71	4.20
2.Bacro et al.(2020)	557	.540	.630	.500	.540	.650	.450	.510	.520	.630	.540	-	-	-	-	-
3.Caycho-Rodríguez et al.(2018)	236	.760	.756	.592	.775	.791	.619	.810	.616	.806	.631	3.58	3.58	3.58	3.55	3.69
4.Cazan (2014)	342	.507	.659	.487	.484	.560	.384	.342	.535	.520	.409	-	-	-	-	-
5.Dirzyte et al.(2021)	2003	.650	.643	.579	.557	.732	.609	.519	.705	.550	.611	3.50	3.70	4.08	4.00	3.64
6.Esnaola et al.(2017)	701	.612	.680	.553	.567	.586	.434	.447	.608	.576	.526	4.88	5.42	5.57	5.23	4.81
7.Espejo et al.(2022)	1255	.539	.59	.587	.435	.571	.514	.458	.658	.513	.437	3.67	4.05	4.00	3.94	3.47
8.Gadermann et al.(2010)	1233	.690	.690	.560	.610	.750	.570	.610	.60	.650	.560	-	-	-	-	-
9.Galanakis et al.(2017)	1797	.620	.630	.460	.450	.640	.470	.440	.570	.490	.450	4.45	4.15	4.80	4.80	4.07
10.Jovanović (2019)(1)	1097	.560	.640	.450	.480	.570	.310	.380	.430	.560	.440	4.42	5.45	5.48	4.01	4.44
11.Jovanović (2019)(2)	998	.570	.710	.530	.550	.550	.360	.370	.510	.570	.550	4.42	4.84	5.28	4.39	4.41
12.Jovanović (2019)(3)	500	.660	.700	.620	.560	.640	.590	.480	.630	.600	.520	3.86	4.06	4.57	4.06	3.78
13.López-Ortega et al.(2016)	13220	.430	.350	.320	.270	.450	.400	.320	.480	.310	.340	1.4	1.5	1.2	1.3	1.5
14.Marcu (2013)	285	.540	.590	.400	.430	.510	.370	.420	.520	.500	.410	4.71	4.66	5.07	5.33	4.38
15.Mishra (2019)	426	.440	.420	.292	.286	.577	.481	.247	.435	.327	.221	4.78	4.77	5.16	4.74	3.88
16.Moksnes et al.(2014)	1073	.610	.600	.500	.440	.720	.580	.490	.700	.570	.590	4.24	4.71	5.04	4.82	4.29
17.Sancho et al.(2014)	1003	.745	.825	.654	.630	.819	.818	.585	.828	.617	.657	3.34	3.06	3.37	3.32	3.60
18.Silva et al.(2015)(1)	461	.480	.440	.510	.520	.440	.440	.360	.380	.380	.530	4.70	5.20	5.30	4.90	4.20
19.Silva et al.(2015)(2)	317	.540	.630	.570	.530	.670	.460	.430	.650	.600	.590	4.50	4.60	5.10	5.0	4.30
20.Silva et al.(2015)(3)	107	.440	.410	.580	.460	.530	.400	.470	.580	.640	.530	4.40	4.70	4.80	4.60	3.60
21.Tomás et al.(2015)	5630	.287	.345	.288	.213	.509	.413	.292	.455	.342	.356	3.36	2.99	3.42	3.12	2.69
22.Tucker et al.(2006)(1)	148	.495	.374	.405	.394	.437	.422	.418	.353	.404	.414	4.65	4.80	5.18	5.07	4.35
23.Tucker et al.(2006)(2)	129	.299	.643	.600	.530	.328	.294	.168	.484	.345	.291	3.63	3.65	4.50	3.79	4.08

* Assoc. Prof. Dr., Mersin University, Faculty of Education, Mersin-Türkiye, skanadli@mersin.edu.tr, ORCID ID: 0000-0002-0905-8677
** Assoc. Prof. Dr., Mersin University, Faculty of Education, Mersin-Türkiye, buzun@mersin.edu.tr, ORCID ID: 0000-0003-2293-4536

To cite this article:
Kanadlı, S. & Uzun N. B. (2023). The factor structure of the satisfaction with life scale (SWLS): A meta-analytic structural equation modeling. *Journal of Measurement and Evaluation in Education and Psychology, 14*(4), 472-491. https://doi.org/10.21031/epod.1221365

Received: 7.03.2023
Accepted: 20.11.2023

Research ID	Sample Size	I1&I2	I1&I3	I1&I4	I1&I5	I2&I3	I2&I4	I2&I5	I3&I4	I3&I5	I4&I5	Mean of the Items				
												I1	I2	I3	I4	I5
24.Wang et al.(2017)(1)	552	.780	.730	.560	.420	.800	.640	.460	.700	.480	.470	3.95	4.06	4.24	4.18	3.34
25.Wang et al.(2017)(2)	566	.770	.730	.640	.560	.790	.700	.570	.710	.570	.570	4.23	4.26	4.46	4.43	3.63
26.Wang et al.(2017)(3)	1060	.740	.710	.620	.520	.810	.650	.530	.700	.560	.590	3.97	3.95	4.13	4.04	3.28
27.Wu et al.(2009)	237	.778	.639	.572	.453	.763	.604	0.466	.628	.480	.577	4.32	4.32	4.52	4.47	3.70
28. Balgiu et al. (2021)	200	.56	.56	.41	.31	.58	.27	.36	.6	.5	.53	4.77	4.88	5.62	5.84	4.62
29. Macovei (2020)	124	.662	.556	.394	.43	.44	.334	.502	.327	.28	.288	-	-	-	-	-
30. Wu & Yao (2006)(1)	207	.708	.692	.631	.683	.739	.611	.545	.677	.664	.661	3.97	4.04	4.42	4.12	3.85
31.Wu & Yao (2006)(2)	269	.733	.759	.612	.86	.708	.567	.436	.641	.537	.521	4.08	4.24	4.49	4.32	3.97
32. Anthimou et al. (2021)	341	.603	.702	.606	.425	.742	.557	.4	.622	.471	.471	-	-	-	-	-
33. García-Castro et al. (2022)	7790	.82	.795	.663	.561	.798	.662	.54	.733	.566	.561	5.09	5.06	5.41	5.57	4.68
34. Theodoropoulou (2021)	360	.767	.774	.67	.585	.809	.642	.6	.717	.647	.664	-	-	-	-	-
35. Berríos-Riquelme et al. (2021)	662	.592	.587	.508	.312	.625	.573	.391	.616	.472	.473	-	-	-	-	-
36. Singh et al. (2021)	400	.5	.454	.496	.429	.622	.524	.389	.539	.429	.547	5.19	5.03	5.12	5.04	4.44
37. Mª et al. (2021)	199	.55	.675	.6	.666	.598	.527	.522	.665	.556	.544	-	-	-	-	-
38. Lang & Schmitz (2020)	641	.58	.59	.38	.47	.77	.39	.49	.39	.54	.46	4.03	4.08	4.38	4.01	3.58
39. Sagar & Karim (2014)	210	.53	.37	.32	.13	.57	.58	.21	NA	NA	NA	-	-	-	-	-
40. Espejo et al. (2022b)***	1222	.59	.539	NA	NA	.571	NA	NA	NA	NA	NA	3.67	4.00	4.05	-	-
41. Kjell & Diener (2021)***	343	.666	.74	NA	NA	.71	NA	NA	NA	NA	NA	4.75	4.93	5.19	-	-

Note. The items of SWLS: (I1)In most ways my life is close to ideal; (I2)The conditions of my life are excellent; (I3)I am satisfied with my life; (I4)So far, I have gotten the important things I want in life; (I5)If I could live my life over, I would change almost nothing

An Illustration of a Latent Class Analysis for Interrater Agreement: Identifying Subpopulations with Different Agreement Levels

Ömer Emre Can ALAGÖZ* Yılmaz Orhun GÜRLÜK** Mediha KORKMAZ***
Gizem CÖMERT****

Abstract

This study illustrates a latent class analysis (LCA) approach to investigate interrater agreement based on rating patterns. LCA identifies which subjects are rated similarly or differently by raters, providing a new perspective for investigating agreement. Using an empirical dataset of parents and teachers evaluating pupils, the study found two latent classes of respondents, one belonging to a moderate agreement pattern and the other belonging to low agreement pattern. We calculated raw agreement coefficient (RAC) per behaviour in the whole sample and each latent class. When RAC was calculated in the whole sample, many behaviour had low/moderate RAC values. However, LCA showed that these items had higher RAC values in the high agreement and lower RAC values in the low agreement class.

Keywords: Interrater Agreement, Latent Class Analysis, Raw Agreement Coefficient, Agreement Methods, Mixture Modelling

Introduction

Using self-report methods for measuring unobservable psychological constructs (i.e., latent variables, traits, factors) is sometimes not possible due to several reasons. For example, researchers need external raters (i.e., observers, evaluators) to gather information about the subjects when the study focuses on disadvantaged groups, students or infants. More specifically, asking children to describe their aggression level by means of filling a questionnaire might be an unrealistic goal. Rather, researchers may employ teachers as raters to assess students' aggression level by observation. Generally, researchers prepare an evaluation list or a manual to inform raters about the indicators of the latent constructs and how to rate them. These ratings can be based on a behaviour's presence/absence, traits, or frequency of behaviour (Bıkmaz Bilgen & Doğan, 2017; Tanner & Young, 1985; Uebersax, 1990; Von Eye & Mun, 2005).

These rating scores are used for different purposes such as research (Leising et al., 2013) or diagnosis of mental illness (Shaffer et al., 1993). Therefore, it is very important to give objective and reliable information to the raters, otherwise, any result from their ratings is likely to be inaccurate. To improve the rating accuracy, researchers usually employ multiple raters. These raters are expected to rate a subject in a similar way since they all follow the same objective instructions. Therefore, using multiple raters and evaluating the consistency between them can also help researchers to understand the quality of instructions (e.g., clarity, objectivity). This consensus among the raters is referred to as interrater agreement (Hallgren, 2012; Landis & Koch, 1977; Uebersax, 1990). Researchers have suggested several methods for testing interrater agreement. These methods can be collected under two headings: 1) classical methods 2) latent variable methods. Since this study focuses on discrete variables, we are only

* Research Assistant, University of Mannheim, alagoez@uni-mannheim.de, ORCID ID: 0000-0003-3305-6564
** PhD student, Ege University, Faculty of Literature, İzmir-Türkiye, yilmazorhungurluk@gmail.com, ORCID ID: 0000-0002-1134-3776
*** Assoc. Prof., Ege University, Faculty of Literature, İzmir-Türkiye, medihakrkmz@gmail.com, ORCID ID: 0000-0001-6504-5822
****PhD student, Ege University, Faculty of Literature, İzmir-Türkiye, cmrtgizem@gmail.com, ORCID ID:0000-0001-7555-6378

To cite this article:
Alagöz, Ö.,E.,C., Gürlük, Y.,O., Korkmaz, M. & Cömert, G. (2023). An illustration of a latent class analysis for interrater agreement: identifying subpopulations with different agreement levels. *Journal of Measurement and Evaluation in Education and Psychology*, *14*(4), 492-507. https://doi.org/10.21031/epod.1308732

Received: 1.06.2023
Accepted: 23.11.2023

concerned with the methods suitable for them. Belonging to classical methods, Cohen's Kappa (Cohen, 1960), Fleiss' Kappa (Fleiss, 1971), and Krippendorff's Alpha coefficients are three of the most popular interrater agreement tests for discrete variables (see footnote 1). However, there is a disadvantage to using them, namely these coefficients are biased when most of the raters only use a specific rating category (Gisev et al., 2013; Göktaş & İşci, 2011, Hayes & Krippendorff, 2007; Yarnold, 2016). In such cases, it is suggested that bias caused by the sparse contingency matrix can be prevented by using raw agreement coefficient (RAC; i.e., percentage of agreement; Feinstein & Cicchetti, 1990; Viera & Garrett, 2005).

In recent years, interrater agreement is studied with latent variable-based probability statistics. Although approaches based on probability distributions were widespread during the 60s, usage of these methods has become more common thanks to software developments (Jiang, 2019; Raykov et al., 2013). There are several advantages to using latent variable models for measuring the interrater agreement. First, we can test the rating consistency between different raters (Agresti, 1992; Yilmaz & Saracbasi, 2017; Yilmaz & Saracbasi, 2019). Second, we can extract the agreement patterns. Third, we can detect anomalies in these patterns. Fourth, we can calculate the sensitivity of raters who give the same rating to participants. Finally, by comparing the agreement patterns obtained from different measurements of the same construct (e.g., two independent studies using the same test with multiple raters), latent variable approaches inform researchers about the reliability and validity of the measurement tool (Jiang, 2019; Kottner et al., 2011; Tanner & Young, 1985; Schuster & Smith, 2002).

Here, we briefly explain two previous latent variable approach to interrater agreement, one treating agreement continuous and one treating it discrete. The first approach is a confirmatory factor analysis that is proposed by Raykov and colleagues (2013). In their approach, rating pattern of a rater can be examined with category thresholds. It is important to note that these category thresholds are rater specific, which means that we do not obtain a summary rating pattern for the whole sample, but we obtain rating pattern for each rater separately. One can test the identity of thresholds across raters to investigate the invariance of cut-offs that raters use to evaluate subjects. By examining the rater-specific thresholds, one can identify those raters who have aberrant rating pattern.

The second latent variable approach is a Latent Class Analysis (LCA) model, which is proposed by Schuster and Smith (2002). Their LCA model consists of two categorical latent variables. The first latent variable represents the true class of a subject, and the second latent variable represents whether a subject is obvious (i.e., all raters agree on them) or ambiguous (i.e., at least one rater disagrees with the rest). Raters can easily identify the true class membership of an obvious subject, whereas they randomly guess the class membership of an ambiguous subject. The agreement on the obvious and ambiguous subjects are referred to as *systematic agreement* and *chance agreement*, respectively. The former can be interpreted as the true agreement between raters. Note that LCA can be used when the rating is done with categorical variables. If rating is done with continuous variables, one can use another approach that is based on Latent Profile Analysis (LPA; Major et al., 2018). For some other LCA approaches in interrater agreement studies, we refer readers to Basten and colleagues (2015), De Los Reyes and colleagues (2009), and Major and colleagues (2018).

To our knowledge, classical approach or latent variable modelling, interrater agreement methods and studies using them focus merely on the consistency between raters (e.g., Forster et al., 2007; Miller, 2011; Thomson, 2003). Indeed, the consistency between raters are vital parts of studies using multiple raters, but another important question arises whenever there is no perfect agreement between raters: Do raters disagree on every subject for every item? In other words, we are interested in whether there are different sets of subjects, each of which is rated in a different way by the raters. There could be, for instance, one set of subjects that are similarly rated by raters for most of the items, and another set of subjects that are rated differently by raters for most of the items. If such a heterogeneity is ignored and an interrater agreement method is used, the result is likely to show a disagreement. In such a case, an important piece of information is disregarded: the subset of subjects on whom raters agreed. Therefore, we propose a LCA approach that detects latent classes of subjects which differ in how they are rated. In the method section, we describe the details of the proposed LCA approach. Then, we analyse an empirical data set and interpret the results.

Sample

In the study, some data within the framework of the "I'm Learning to Protect Myself with Mika" sexual abuse prevention program developed by Kızıltepe and colleagues (2022) were used with the permission of the researchers (see footnote 2). In order to determine whether the aforementioned intervention program had side effects, the researchers created 10 dichotomously scored items (for item contents, see Table 3). The parents evaluated only their children and the teachers evaluated all of the students who have taken the MIKA prevention program. For the examined agreement between teachers and parents two rater blocks were structured as parent and teacher. The blocks were handled as two raters, and it was tested whether there was concordance between the evaluations of the teachers and the parents.

The sample of the study consists of 290 children in the 5-year-old age group from the lower, middle, and upper socio-economic status attending kindergarten and their parents. In the study, considering the districts where the kindergartens are located, two schools from the lower socio-economic status, three from the middle socio-economic status and one school from the upper socio-economic status were selected. Two of the schools are private and four of them were kindergartens within the part of a public institution. This form was answered by parents and teachers for 290 children. The proportion of boys and girls in the sample were 52.76% (*N*=153) and 47.24% (*N*=137), respectively. After data screening, 13 observations were omitted (Both parents and teachers of 4 and only parents of 9 did not evaluate the children). In order to create intervention and control groups that are equivalent in terms of age, socio-economic status, gender and so on, one classroom from each school was included in the training group and one classroom in the control group. Only the intervention group was used to examine the agreement between raters. The ages of the children ranged from 50-72 months (Mean = 61.80, SD = 6.1). The ages of the mothers of the children participating in the study ranged from 22 to 47 (Mean = 24.32, SD = 5.25), while the age of their fathers ranged between 25 and 53 (Mean = 37.67, SD= 5.59). In the study, 84% of the parents participating were mothers and 16% were fathers.

Methods

LCA Approach to Interrater Agreement

We utilize LCA to detect subpopulations of subjects that differ by how they are rated on several categorical items by different raters. Different from other LCA rater agreement approaches, our approach does not classify raters into "agreement" and "disagreement" classes and does not investigate the similarity of ratings at item-level. Rather, we classify subjects into classes depending on the similarity of scores given to them by different raters. Therefore, this approach does not necessarily find "agreement" or "disagreement" classes, but it captures whether raters evaluate all respondents similarly on a set of items.

First, we transform the data set into a new form by subtracting the scores given by one group of raters from the scores given by the other group of raters. That is, assume that X^A and X^B are $N \times J$ matrices of scores given to N number of subjects on J number of items by rater group A and rater group B, respectively. Then, the matrix of rating distances between rater groups, X^{A-B}, is an $N \times J$ matrix that is calculated by $X^A - X^B$. If raters evaluate subjects on $M \in \{1, \ldots, m\}$ response categories, $X_{nj}^{A-B} \in \{1 - m, \ldots, 0, \ldots, m - 1\}$ is the signed distance between two raters for subject n and variable *j.* A negative (positive) X_{nj}^{A-B} means that the subject is assigned a higher (lower) score by the rater in group B than the rater in group A. If X_{nj}^{A-B} is zero, then subject n is assigned the same score by raters in both groups, therefore raters agree with each other.

Second, we conduct LCA with the transformed data set X^{A-B}. Each column of X^{A-B} is specified as an indicator variable. Therefore, classes are defined with how divergent do raters score subjects. We fit models with increasing number of latent classes and investigate several fit indices to decide on the number of latent classes. By investigating conditional response probabilities, we can find which subjects on which items are evaluated similarly (or differently) by the raters.

Third, we calculate the posterior class probabilities for each subject and assign them to the class for which their posterior class probability is the highest (i.e., modal assignment, Dias & Vermunt, 2008).

Optionally, if we believe classes represent subjects that are rated similarly and differently by raters, then we can conduct a classical rater agreement analysis (e.g., Kappa, RAC) in each class to verify whether it was the case.

Latent classes can differ due to presence/absence of agreement, type of disagreement, and items where these (dis-)agreements occur. If latent classes differ due to presence/absence of agreement, then the response probability in one class will be the highest for the category "0", whereas in other classes it will be the lowest for the category "0". If classes differ regarding the type of disagreement (see footnote 3), response probabilities in one class(es) can be higher for positive categories, whereas in the other class(es), it can higher for negative categories. Finally, the reason for class differences can differ item by item. That is for one item, classes can differ due to presence/absence of agreement, and for another item, they can only differ due to type of disagreement. In conclusion, it is very important to carefully examine the conditional response probabilities to make sense of classes. Indeed, it is a rule that applies to all latent class analyses.

Procedure

In the first step, we conducted LCA in the abovementioned way. Then, the number of classes that differ in the similarity of ratings between two rater groups were determined, and respondents were assigned into the class for which their posterior class probability was the highest. Next, we investigated the conditional response probabilities for rating distances to understand the characteristics of each class. Finally, we calculated the RAC in each class to confirm our interpretations about class characteristics. We used Latent GOLD (Vermunt & Magidson, 2008) for conducting LCA and IBM SPSS Statistics 25 for calculating RAC. In our empirical illustration, we calculated RAC due to skewed ratings provided by one rater group. Hereby we explain how RAC can be calculated. Given that two raters evaluate *N* number of subjects on M number of categories, the below $M \times M$ table can be constructed. There, n_{jk} denotes the total number of subjects that are rated with category *j* by rater A and with category *k* by rater B. On diagonal, we have numbers of cases where raters evaluate subjects with the same categories.

Table 1:
Summary of ratings by two raters. $M \times M$ table of ratings provided by two raters, where each cell n_{jk} represents the number of subjects who are rated with category j by rater A and with category k by rater B.

	Rater A				
Rater B	1	2	…	M	Column Sums
1	n_{11}	n_{12}	…	n_{1M}	n_{1+}
2	n_{21}	n_{22}	…	n_{2M}	n_{2+}
…	…	…	…	…	…
M	n_{M1}	n_{M2}	…	n_{MM}	n_{M+}
Row Sums	n_{+1}	n_{+2}	…	n_{+M}	N

Then RAC is calculated as the proportion of respondents rated with the same category by both rater A and rater B:

$$RAC = \frac{1}{N}\sum_{m=1}^{M} n_{mm}$$

Results

In line with the model described in the method section, we fit latent class models with 1 to 5 classes. To decide on the number of latent classes, we compared Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

AIC and BIC are information criteria that takes model fit and model complexity into account and inform us about the balance between fit and complexity. The lower the AIC and BIC get, there is a better balance between fit and complexity. Since including more parameters to the model will increase the model fit as it approaches to saturation, these criteria penalize the fit index with the number of observed units (i.e., sample size) in AIC and with both the number of observed units and number of parameters in BIC. For this reason, BIC is mostly preferred over AIC. Also for LCA, the model with the lowest AIC or BIC should be preferred, but in case of very close values, researchers can choose among best fitting models according to theory and interpretability of results (Nylund and colleagues, 2007).

In Table 2, we provide the AIC and BIC values for LCA model with different number of classes. Accordingly, as is seen in Table 2, AIC favours the model with 5 classes, whereas BIC suggests the model with 3 classes fit the data best. We base our decision on BIC results since it penalizes the model with more parameters. However, since the difference in BIC between the model with 2 classes and 3 classes is very small, we chose the model with 2 classes for parsimony and interpretability reasons. Also, a further investigation of the model with 3 classes showed that the size of the added class is very close to zero, which then made sense to choose the model with 2 classes. We care for brevity and parsimony because the reason for using LCA is to capture different rating patterns but not to make substantial inferences about classes. In this 2-class model, the size of Class 1 was found 0.62 (*N*=183) and of Class 2 was found 0.38 (*N*=107).

Table 2:

Model comparison. From the leftmost column to the rightmost column, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and degrees of freedom (df)

Number of Classes	AIC	BIC	Df
1	7171.56	7432.12	219
2	7021.89	7322.82	208
3	6977.21	7318.51	197
4	6972.28	7353.95	186
5	6957.54	7379.58	175

We investigated conditional probabilities (see Table 3) of the differences between parents' $(X^{parents})$ and teachers'$\left(X^{teachers}\right)$ ratings to understand in what sense these classes differ from one another. The most visible difference between classes is the probability of category zero, in other words, agreement. We see that subjects in Class 1 have higher probabilities of being identically rated by their parents and teachers on all items. Actually, the probability of agreement is generally very high and larger than .50 in all items except items 4, 5, and 6. However, in Class 2, the probability of category zero is roughly below .50 in all items except item 3.

In Table 3, we also see for Class 1 that the second largest category probability after "0" is "+1", and "+1" is followed by "-1" (item 6 can be an exception). Meaning of this pattern is, in Class 1, the most probable outcomes are either a perfect agreement or a difference of one unit between raters. This finding further supports that Class 1 is associated with high agreement. If we look at Class 2, we see that it is either the extreme disagreement category "+4" or intermediate disagreement categories "+1, +2, +3" have the highest probabilities (item 3 can be an exception).

The implication of this pattern is that Class 2 is associated with higher disagreement, where teachers provide low ratings and parents provide high ratings for the subjects.

The reason for disagreement between parents and teachers can be that teachers systematically evaluate their students lower than parents or parents systematically evaluate their children higher than teachers.

To investigate if either one is the case, we investigated the category frequencies and variances for parents and teachers. The typical finding was parents showed a higher variance in their ratings compared to teachers and teachers used smaller categories more often than parents. For example, the variances of parents' and teachers' ratings about "Anger Problems" item were 1.29 and 0.45, respectively. Moreover, parents rated 105 children with the lowest category, whereas teachers rated 254 students with the lowest category. Therefore, the disagreement occurs because teachers systematically give a low score to the children. The reason for such tendency could be that teachers have only limited time and a fixed context to observe children, whereas parents spend more time and observe their children in different contexts (also they have biases from the times before the study).

Table 3:

Conditional probabilities of rating differences between different raters $Pr\left(X_{nj}^{A-B} = m \middle| Class = c\right)$

Items	Classes	$m\ (X^{parents} - X^{teachers})$									RAC
		-4	-3	-2	-1	**0**	1	2	3	4	
Item 1	Class 1	-	-	.01	.02	**.71**	.23	.03	.01	.01	**.72**
	Class 2	-	-	.01	.01	**.37**	.38	.17	.04	.04	**.32**
Item 2	Class 1	-	.01	-	.01	**.71**	.20	.04	.01	.01	**.72**
	Class 2	-	.01	-	.01	**.46**	.26	.10	.06	.11	**.44**
Item 3	Class 1	-	-	.01	.01	**.82**	.12	.03	.01	-	**.84**
	Class 2	-	-	.01	.01	**.75**	.17	.06	.01	-	**.75**
Item 4	Class 1	.02	.07	.10	.15	**.28**	.20	.09	.03	.03	**.31**
	Class 2	.01	.03	.05	.10	**.26**	.24	.15	.07	.07	**.21**
Item 5	Class 1	.01	.03	.04	.10	**.47**	.25	.06	.02	.01	**.52**
	Class 2	.01	.01	.01	.03	**.28**	.31	.16	.10	.10	**.23**
Item 6	Class 1	-	.03	.03	.06	**.24**	.32	.18	.08	.06	**.25**
	Class 2	-	.01	.01	.01	**.07**	.18	.21	.20	.32	**.06**
Item 7	Class 1	-	.01	.01	.08	**.72**	.15	.03	.01	.01	**.69**
	Class 2	-	.01	.01	.02	**.52**	.27	.12	.04	.02	**.54**
Item 8	Class 1	.01	.02	.01	.05	**.62**	.21	.05	.01	.01	**.67**
	Class 2	.01	.01	.01	.01	**.31**	.28	.18	.09	.12	**.27**
Item 9	Class 1	-	.01	.01	.02	**.83**	.12	.01	.01	.01	**.86**
	Class 2	-	.01	.01	.01	**.51**	.27	.09	.07	.04	**.46**
Item 10	Class 1	.01	.01	.01	.09	**.69**	.18	.01	-	.01	**.71**
	Class 2	.01	.01	.01	.02	**.48**	.39	.05	-	.06	**.45**

Note. The cells with "-" means that the difference between parents' and teachers' was never yielded the category *m* for any respondent. Conditional probabilities of the agreement category "0" and class-specific raw agreement coefficient (RAC) are given in bold face.

Since there is a visible difference in the agreement probabilities between classes, we further investigated whether Class 1 that has higher agreement probabilities yielded a higher agreement coefficient with a classical interrater agreement analysis. Therefore, we calculated RAC for each item once for respondents in Class 1 and once for respondents in Class 2. When RAC in Class 1 was examined, we see values larger than .70 for five items and larger than .50 for eight items (see Table 4). These eight items are the ones that conditional response probabilities suggested similarity in ratings between parents and teachers. Furthermore, both conditional probabilities of "0" from LCA and small RAC values pointed out that there is a difference between parents' and teachers' rating patterns for items 5 and 6 in Class 1. For the results of Class 2, it was visible that all items have RAC smaller than around .50 except for item 3 and 7, for which conditional response probabilities also suggested dissimilar rating patterns between parents and teachers.

What is interesting is that conditional response probability of category "0" is almost identical to the class-specific RAC for all items. However, this is not much surprising since RAC is quantified by the total proportion of subjects that are evaluated with the same category by two raters, subtraction of which is equal to "0". Yet, RAC is merely calculated with the observed variables and deterministic, whereas the RAC-like values obtained via LCA is probabilistic. This finding implies that conducting LCA can be adequate to also quantify the interrater agreement without further separate analysis.

Table 4:

Raw Agreement Coefficients per item for Class 1 (left), for Class 2 (middle), for the whole sample (right).

Items	Class 1	Class 2	Overall
Afraid of Animals	0.72	0.32	0.57
Separation Anxiety	0.72	0.44	0.61
Questions about Sexuality	0.84	0.75	0.81
Difficulty of Expressing Emotions	0.31	0.21	0.27
Disobedience	0.52	0.23	0.41
Whining	0.25	0.06	0.17
School Avoidance	0.69	0.54	0.63
Anger Problems	0.67	0.27	0.52
Stranger Anxiety	0.86	0.46	0.71
Afraid of Adults	0.71	0.45	0.61
Mean	0.63	0.37	0.53

Discussion and Conclusion

In this study, we proposed a LCA to investigate rater agreement for categorical rating data. We first explained the need for and benefits of such an approach then described how LCA parameters help investigating rater agreement with a fictitious example. Lastly, we analysed an empirical dataset with the proposed approach.

Previous classical or latent variable approaches for rater agreement research focus on quantifying the agreement between two or more raters for all respondents at once. However, it is also possible that raters give similar scores to one set of respondents and different scores to another set of respondents. To detect such respondent groups, one can first create new variables by subtracting the ratings of one rater group from the ratings of another rater group. Hence, this new variable indicates the distance and its direction between the ratings of two rater groups. Then, one can conduct LCA on these new variables to capture respondent subpopulations who were rated similarly (zero distance) and differently (non-zero distances) by rater groups.

The current practice is to ignore the existence of such subpopulations associated with different levels of agreement and to calculate a single agreement coefficient (e.g., RAC) for the whole sample. However, in the presence of subpopulations, the sample RAC is roughly the weighted average of RAC values calculated for subpopulations. Hence, the sample RAC is likely to be biased. In our empirical example, we identified two latent classes, one related with smaller rating distances and the other related with larger rating distances between raters. Indeed, we showed that the RAC calculated for the whole sample was around the weighted average of class-specific RAC values. Furthermore, we showed that RAC calculated in smaller distance class was higher than the RAC calculated in larger distance class.

Another advantage of using LCA is that, in case of disagreement between raters, conditional response probabilities inform us about the direction of disagreement. If researchers conduct a classical agreement analysis, they obtain a single coefficient value quantifying the agreement. To understand more about why disagreement occurs, they need to examine descriptive statistics. However, the conditional response probabilities in LCA already tells researchers about which rater group tends to give higher or smaller scores than the other group. Moreover, researchers can also easily evaluate how severe is the disagreement. For example, disagreement is less severe when the conditional response probabilities are higher for ± 1 distance categories compared to when they are higher for ± 5 distance categories.

The reason for using RAC in our analysis was the sparse contingency tables of ratings. That is, some raters did not use some of the categories. In case of a sparse contingency table, other classical agreement analyses than RAC were found to yield biased estimates. As the second step after finding latent classes, one can always use other analyses within each class to see if they indeed represent different levels of agreement. However, calculating RAC for a variable after LCA was redundant in our analysis since the conditional response probability of distance category “0” has always corresponded to the RAC value of that specific class (see Table 2).

In the proposed LCA approach, we include all items to the analysis at once, whereas one calculates the agreement item by item in the classical rater agreement approaches. By doing so, all items contribute to the classification of respondents into classes. One can of course include only one set of items to the analysis, but it rather conflicts with the main idea of using LCA for rater agreement analysis, that is to make use of rating patterns across many items rather than doing item by item analyses.

As in all studies, there are also limitations to our approach. First limitation is the sample size requirement of LCA. Usually, it is required to have at least 500 respondents in the data set for using LCA. Although this requirement is not a limitation to the LCA approach to the rater agreement, it is for our empirical example. However, we do not see our sample size (i.e., 290 children) as a problem for two reasons: 1) the main aim for using LCA is to capture differences between rating patterns, and 2) the empirical example is only used to demonstrate how LCA parameters are interpreted in rater agreement domain but not to answer substantive research questions about the MIKA measurement tool.

Another limitation that we are aware of is the ambiguity of interpretations of classes. Actually, it is not a limitation but a general feature of LCA. That is, researchers need to examine class-specific parameters to understand or speculate about the characteristics or the labels of classes. We see it as a limitation in the sense that our approach does not yield two clear cut classes related to rater agreement and disagreement. Indeed, in the empirical illustration, there were some items that contradict with our class explanations. However, we believe that the majority of items were consistent with our class interpretation. Yet, we acknowledged that such interpretations can be sometimes subjective, that is why we included Table 3 with detailed conditional probabilities to be transparent and to allow readers to better evaluate our interpretation (as ours is maybe only one out of many alternatives). Moreover, in Appendix B, we provide the results from the model with three classes, which was the favourite of BIC, to see if it leads to a different class explanation (see footnote 4). However, the newly added class was almost practically empty (*N*=9), so none of our interpretations have differed. Compared to our empirical example, some data sets might require more time and effort to make sense of the meaning of latent classes.

Future research can include respondent level covariates that might explain why they were assigned into different classes. If classes separate who are rated similar or different (as in our empirical analysis), then such covariates would tell why raters agreed or disagreed on the rating of an item. Also, future research can focus on a confirmatory mixture model that classifies raters into agreement and disagreement classes. Further, it would be interesting to see how this approach works with other data sets, both from similar and different domains, and whether similar types of classes arise with other data sets. Finally, future research should examine using LCA with non-sparse data to see whether other classical agreement analyses are also suitable for quantifying class-specific rater agreements in the second step.

Despite the limitations, we believe that latent variable models can help us learn more about the rater agreement. We also believe that our approach focuses on an important aspect of the rater agreement, which is the respondents that are being rated. With the proposed LCA approach, one does not have to disregard the substantive research question at hand because of the rater disagreement, but they can focus on the respondents for whom there is an agreement between raters.

Footnotes:

1) For interested readers, among many other, Konstantinidis and colleagues (2022), Zapf and colleagues (2016), Sertdemir and colleagues (2013), and Ato and colleagues (2011) provide extensive simulation studies and theoretical discussions regarding the classical agreement methods and the comparison of their performances. These methods and their comparisons are beyond the scope of this study, therefore they are not discussed in this paper.
2) Requests for accessing the data that support the findings of this study should be made to the the corresponding author of Kızıltepe and colleagues (2022).
3) It can co-exist with the presence/absence classes. For example, there can be one class with respondents who are scored similarly, another class with respondents for whom raters in group A provide higher scores than raters in group B, and another class with respondents for whom raters in group B provide higher scores than raters in group A.
4) We thank the anonymous reviewer for suggesting to add this alternative model results.

Acknowledgements

We would like to thank Rukiye KIZILTEPE, Duygu ESLEK, Türkan YILMAZ IRMAK, Duygu GÜNGÖR CULHA for sharing the research data with us.

Data Availability

Requests for accessing the data that support the findings of this study should be made to the corresponding author of Kızıltepe and colleagues (2022).

Declarations

Author Contribution: Ömer Emre Can ALAGÖZ – conceptualization, methodology, software, analyses, writing, editing. Yılmaz Orhun GÜRLÜK – conceptualization, software, analyses, writing, editing. Mediha KORKMAZ – conceptualization, supervising. Gizem CÖMERT – conceptualization, writing, editing, visualization.

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

Ethical Approval: Secondary data were used in this study. Therefor ethical approval is not applicable.

References

Agresti, A. (1992). Modelling patterns of agreement and disagreement. Statistical Methods in Medical Research, 1, 201-218. https://doi.org/10.1177/096228029200100205

Ato, M., López, J. J., & Benavente, A. (2011). A simulation study of rater agreement measures with 2x2 contingency tables. Psicológica, 32(2), 385–402.

Basten, M., Tienmeier H., Althoff, R., van de Schoot, R., Jaddoe, V. W. V., Hofman, A., Hudziak, J. J., Verhulst, F. C. & Van der Ende, J. (2015). The stability of problem behavior across the preschool years: an empirical approach in general population, Journal of Abnormal Child Psychology, 44(2), 393-404. https://doi.org/10.1007/s10802-015-9993-y

Bıkmaz Bilgen, Ö. ve Doğan, N. (2017). Puanlayıcılar arası güvenirlik belirleme tekniklerinin karşılaştırılması, Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 8(1), 63-78. https://doi.org/10.21031/epod.294847

Cohen (1960). A coefficient of rater agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46. https://doi.org/10.1177/001316446002000104

De Los Reyes, A., Henry, D. B., Tolan, P. H. T. & Wakschlag, L. S. (2009). Linking informant discrepancies to observed variations in young children’s disruptive behavior, Journal of Abnormal Psychology, 37(5), 637-652. https://doi.org/10.1007/s10802-009-9307-3

Feinstein, A. R. & Cicchetti, D. V. (1990). High agreement but low kappa: I. the problems of two paradoxes, Journal of Clinical Epidemiology, 43(6), 543-549. https://doi.org/10.1016/0895-4356(90)90158-L

Fleiss, J. L. (1971). Measuring agreement for multinomial data. Psychological Bulletin, 76(5), 378-382. https://doi.org/10.1037/h0031619

Forster, A. J., O'Rourke, K., Shojania, K. G., & van Walraven, C. (2007). Combining ratings from multiple physician reviewers helped to overcome the uncertainty associated with adverse event classification. Journal of clinical epidemiology, 60(9), 892-901.

Gisev, N., Simon Bell, J. & Chen, T. F. (2013). Interrater agreement and interrater reliability: key concepts, approaches, and applications. Research in Social and Administrative Pharmacy, 9, 330-338. https://doi.org/10.1016/j.sapharm.2012.04.004

Göktaş, A. & İşçi, Ö. (2011). A comparison of the most commonly used measures of association for doubly ordered square contingency tables via simulation. Metodoloski Zvezki, 8(1), 17-37. https://doi.org/10.51936/milh5641

Hallgren, K. (2012). Computimg inter-rater reliability for observational data: an overview and tutorial, Tutorials in Quantitative Methods for Psychology, 8(1), 23-34. https://doi.org/10.20982/tqmp.08.1.p023

Hayes, A. F. & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding, Communication Methods and Measures, 1(1), 77-89. https://doi.org/10.1080/19312450709336664

Jiang, Z. (2019). Using the iterative latent-class analysis approach to improve attribute accuracy in diagnostis classification models. Behavior Research Method, 51, 1075-1084. https://doi.org/10.3758/s13428-018-01191-0

Kızıltepe R., Eslek, D., Yılmaz Irmak, T. & Güngör, D. (2022). I am learning to protect myself with Mika:” a teacher-based child sexual abuse prevention program in Turkey. Journal of Interpersonal Violence, 37(11-12), 1-25. https://doi.org/10.1177/0886260520986272

Konstantinidis, M., Le, L. W., & Gao, X. (2022). An empirical comparative assessment of inter-rater agreement of binary outcomes and multiple raters. Symmetry, 14(2), 262. https://doi.org/10.3390/sym14020262

Kottner, J.,Audige, L., Brorson, S., Donner, A., Gajewski, B., Hrobjartsson, A., Roberts, C., Shoukri, M. & Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. Journal of Clinical Epidemiology, 64, 96-106. https://doi.org/10.1016/j.ijnurstu.2011.01.016

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174. https://doi.org/10.2307/2529310

Leising, D., Ostrovski, O., & Zimmermann, J. (2013). "Are we talking about the same person here?" Interrater agreement in judgments of personality varies dramatically with how much the perceivers like the targets. Social Psychological and Personality Science, 4(4), 468-474. https://doi.org/10.1177/1948550612462414

Major, S., Seabra-Santos, M. J. & Martin, R. P. (2018). Latent profile analysis: another approach to look at parent-teacher agreement on preschoolers' behavior problems. European Early Childhood Education Research Journal, 26(5), 701-717. https://doi.org/10.1080/1350293X.2018.1522743

Miller, W. E. (2011). A latent class method for the selection of prototypes using expert ratings. Statistics in Medicine, 31(1), 80-92.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Structural Equation Modeling, 14(4), 535–569. https://doi.org/10.1080/10705510701575396

Raykov, T., Dimitrov, D. M., von Eye, A. & Marcoulides, G. A. (2013). Interrater Agreement Evaluation: a latent variable modeling approach. Educational and Psychological Measurement, 20(10). 1-20. https://doi.org/10.1177/0013164412449016

Schuster, C. & Smith, D. A. (2002). Indexing systematic rater agreement with a latent-class model. Psychological Methods, 7(3), 384-395. https://doi.org/10.1037/1082-989X.7.3.384

Sertdemir, Y., Burgut, H. R., Alparslan, Z. N., Unal, I., & Gunasti, S. (2013). Comparing the methods of measuring multi-rater agreement on an ordinal rating scale: a simulation study with an application to real data. Journal of Applied Statistics, 40(7), 1506-1519. https://doi.org/10.1080/02664763.2013.788617

Shaffer, D., Schwab-Stone, M., Fisher, P., Cohen, P., Placentini, J., Davies, M. & Regier, D. (1993). The diagnostic interview schedule for children-revised version (DISC-R): I. Preparation, field testing, interrater reliability, and acceptability. Journal of the American Academy of Child & Adolescent Psychiatry, 32(3), 643-650. https://doi.org/10.1097/00004583-199305000-00023

Tanner, M. A. & Young, M. A. (1985). Modelling agreemet among raters. Journal of the American Statistical Association, 80(389), 175-180. https://doi.org/10.1080/01621459.1985.10477157

Thompson, D. M. (2003). Comparing SAS-based applications of latent class analysis using simulated patient classification data. The University of Oklahoma Health Sciences Center.

Uebersax, J. S. & Grove, W. M. (1990). Latent class analysis of diagnostic agreement. Statistisc in Medicine, 9(5), 559-572. https://doi.org/10.1002/sim.4780090509

Viera, A. J. & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistics, Family Medicine, 37(5), 360-363. PMID: 15883903

Von Eye, A. & Mun, E. Y. (2005). Analyzing rater agreement manifest variable methods (1st ed.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410611024

Yarnold, P. R. (2016). ODA vs. π and κ: paradoxes of kappa, Optimal Data Analysis, 5, 160-161. Accessed at: https://www.researchgate.net/publication/309681250_ODA_vs_p_and_k_Paradoxes_of_Kappa, 23.03.2023

Yilmaz, A. E. & Saracbasi, T. (2017). Assessing agreement between raters from the point of coefficients and log-linear models. Journal of Data Science, 15, 1-24. https://doi.org/10.6339/JDS.201701_15(1).0001

Yilmaz, A. E. & Saracbasi, T. (2019). Agreement and adjusted degree of distinguishability for square contingency tables. Hacettepe Journal of Mathematics and Statistics, 48(2), 592-604. https://doi.org/10.15672/hjms.2018.620

Zapf,A.,Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data–which coefficients and confidence intervals are appropriate?. BMC medical research methodology, 16, 1-10. https://doi.org/10.1186/s12874-016-0200-9

APPENDICES

Appendix A: Latent GOLD Syntax

```
options
  algorithm
    tolerance=1e-008 emtolerance=0,01 emiterations=250 nriterations=50;
  startvalues
    seed=0 sets=10 tolerance=1e-005 iterations=50;
  bayes
    categorical=1 variances=1 latent=1 poisson=1;
  montecarlo
    seed=0 replicates=500 tolerance=1e-008;
  quadrature  nodes=10;
  missing  includeall;
  output
    parameters=first standarderrors probmeans=posterior profile bivariateresiduals;
variables
  dependent hay, ayr, cin, duy, it, miz, okul, of, yab, yet;
  latent
    Cluster nominal 2;
equations
  Cluster <- 1;
  hay <- 1 + Cluster;
  ayr <- 1 + Cluster;
  cin <- 1 + Cluster;
  duy <- 1 + Cluster;
  it <- 1 + Cluster;
  miz <- 1 + Cluster;
  okul <- 1 + Cluster;
  of <- 1 + Cluster;
  yab <- 1 + Cluster;
  yet <- 1 + Cluster;
```

Appendix B: Results for the model with three classes

We also provide the results for the model with three classes as it had slightly lower BIC value than the model with two classes. First, we investigated the class sizes. We found that the size of the first class was 0.61, the size of the second class was 0.35, and the third class was 0.04. With modal assignment, 184 subjects were assigned to class 1, 97 subjects were assigned to class 2, and only 9 subjects were assigned to class 3. These class sizes provide further reasons for sticking to the model with two classes, because the newly added class is very small; therefore, its estimates would be less accurate. Regardless of that, we provide the conditional probabilities of rating differences between raters given classes in Table B1.

Table B1:

Conditional probabilities of rating differences between different raters $Pr\left(X_{nj}^{A-B} = m \middle| Class = c\right)$

Items	Classes	$m\ (X^{parents} - X^{teachers})$									RAC
		-4	-3	-2	-1	**0**	1	2	3	4	
Item 1	Class 1	-	-	.01	.02	**.71**	.23	.03	.01	.01	**.73**
	Class 2	-	-	.01	.01	**.33**	.38	.19	.04	.04	**.28**
	Class 3	-	-	.01	.01	**.55**	.33	.09	.01	.01	**.44**
Item 2	Class 1	-	.01	-	.01	**.72**	.20	.04	.01	.01	**.73**
	Class 2	-	.01	-	.01	**.43**	.24	.12	.05	.15	**.41**
	Class 3	-	.01	-	.01	**.64**	.23	.07	.02	.03	**.44**
Item 3	Class 1	-	-	.01	.01	**.85**	.11	.02	.01	-	**.86**
	Class 2	-	-	.01	.01	**.75**	.18	.05	.01	-	**.72**
	Class 3	-	-	.01	.01	**.67**	.22	.09	.01	-	**.67**
Item 4	Class 1	.02	.06	.10	.14	**.28**	.20	.10	.04	.04	**.31**
	Class 2	.01	.03	.05	.10	**.26**	.23	.15	.09	.10	**.20**
	Class 3	.03	.09	.13	.16	**.28**	.17	.08	.03	.02	**.13**
Item 5	Class 1	.01	.01	.04	.09	**.48**	.25	.07	.03	.01	**.53**
	Class 2	.01	.01	.01	.03	**.30**	.29	.15	.12	.10	**.21**
	Class 3	.12	.15	.15	.15	**.07**	.07	.01	.01	.01	**.11**
Item 6	Class 1	-	.01	.02	.06	**.23**	.34	.18	.08	.08	**.22**
	Class 2	-	.01	.01	.01	**.07**	.20	.20	.18	.34	**.07**
	Class 3	-	.24	.18	.18	**.25**	.13	.02	.03	.01	**.44**
Item 7	Class 1	-	.01	.01	.08	**.70**	.16	.02	.01	.01	**.71**
	Class 2	-	.01	.01	.02	**.50**	.29	.11	.04	.03	**.52**
	Class 3	-	.01	.01	.06	**.68**	.20	.04	.01	.01	**.33**
Item 8	Class 1	.01	.01	.01	.04	**.62**	.22	.06	.01	.01	**.68**
	Class 2	.01	.01	.01	.01	**.33**	.28	.19	.09	.11	**.26**

	Class 3	.20	.40	.14	.12	**.13**	.01	.01	.01	.01	**.00**
Item 9	Class 1	-	.01	.01	.02	**.85**	.13	.01	.01	.01	**.88**
	Class 2	-	.01	.01	.01	**.50**	.28	.09	.07	.05	**.43**
	Class 3	-	.10	.20	.29	**.41**	.01	.01	.01	.01	**.33**
Item 10	Class 1	.01	.01	.02	.10	**.70**	.17	.01	.01	.01	**.71**
	Class 2	.01	.01	.01	.02	**.46**	.38	.05	.01	.09	**.42**
	Class 3	.01	.01	.01	.04	**.63**	.29	.02	.01	.01	**.67**

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(4); 508-523

Modelling the Differences in Social and Emotional Skills with Polytomous Explanatory IRT: The Example of Assertiveness Skill*

Fatma Nur AYDIN ** Kübra ATALAY KABASAKAL ***

Abstract

Explanatory item response theory models can simultaneously decompose the covariance between persons and items, as well as analyze items by adding item-related predictors for differences between item difficulties and/or person-related predictors for differences between individuals. In the current study, we calculated the parameter estimations regarding the skill of assertiveness according to the rating scale model and partial credit model, which are descriptive (traditional) item response theory models as well as latent regression partial credit model including only person-level predictors, and then examined the results comparatively. We used the raw score belonging to the skill of assertiveness of Türkiye belonging to the OECD Social and Emotional Skills Study, and we included gender, socioeconomic level, perceived relationships with teachers, bullying at school, sense of belonging at school, global mindedness, and test anxiety as person-level predictors. Current study findings suggest that; (1) the latent regression partial credit model produces a better data fit when compared to the rating scale model and partial credit model, and (2) sense of belonging at school, global mindedness, and socioeconomic level are significant predictors to explain the differences between persons. We discussed the current study findings in terms of the rich body of knowledge provided by explanatory item response theory and presented some suggestions.

Keywords: polytomous IRT models, polytomous explanatory IRT models, social and emotional skills, assertiveness

Introduction

Item response theory (IRT) is a mathematical theory of measurement that indicates that it is possible to establish a relationship between one’s performance in a test, and latent traits or abilities assumed to underlie this performance (Hambleton & Swaminathan, 1985). IRT aims to make inferences about the features measured by a test (Baker, 2016). There are various IRT models such as those including items in two categories (Rasch model; one, two, three parameter models) (Embretson & Reise, 2000) and polytomous models (partial credit model (Masters, 1982), rating scale model (Andrich, 1978a; 1978b), graded response model (Samejima, 1969). When IRT models are considered within the framework of generalized linear mixed models and non-linear mixed models, it is possible to get descriptive and explanatory models (De Boeck & Wilson, 2004). If a model explains item qualities with parameters such as item difficulty and item discrimination, while it explains individuals’ performances in terms of ability scores, it is called a “descriptive” measurement model (De Boeck & Wilson, 2004). The aforementioned traditional IRT models are considered to be descriptive models. On the other hand, generalized linear mixed models allow IRT models to be addressed with a multi-level approach. According to that, responses to items are dependent on the emerging hierarchical structure. In other words, responses to items are addressed as repetitive measurements nested in individuals (De Boeck & Wilson, 2004). It is possible to estimate the impact of predictor variables with the help of such a

* A part of this study was presented as an oral presentation at the 8. Congress of Measurement and Evaluation in Education and Psychology (CMEEP)

**Teacher., Republic of Türkiye Ministry of National Education, Ankara-Türkiye, fatmanuraydin.2012@gmail.com, ORCID ID: 0000-0003-0887-395X

*** Assoc. Prof., Hacettepe University, Faculty of Education, Ankara-Türkiye, kkatalay@gmail.com, ORCID ID: 0000-0002-3580-5568

To cite this article:
Aydın, F.N., Kabasakal Atalay, K. (2023). Modelling the differences in social and emotional skills with polytomous explanatory IRT: The example of assertiveness skill. *Journal of Measurement and Evaluation in Education and Psychology, 14*(4), 508-523. https://doi.org/10.21031/epod.1299468

Received: 19.05.2023
Accepted: 21.11.2023

modelling approach (Stanke & Bulut, 2019). Explanatory item response theory models (De Boeck & Wilson, 2004) refer to models that analyse items by focusing on differences between item difficulties through adding item-related predictors and/or focusing on differences between individuals through adding person-related predictors as well as dissociating the common variance between person and items at the same time (Briggs, 2008; De Boeck & Wilson, 2004). Although studies in the literature often use descriptive measurement models that make it possible to find an answer to many problems, these models do not provide information on systematic effects that can explain observations, and so they cannot explain the common variability among persons and items (De Boeck & Wilson, 2004; Stanke & Bulut, 2019). However, explanatory IRT models can meet this demand, and these models can be divided into four according to the predictors they include. According to that, if the model does not include a predictor at the level of person or item, it is called "doubly descriptive (i.e. traditional IRT models)"; if it includes a predictor at the person level, it is called "person explanatory (i.e. the latent regression Rasch model)"; if it includes a predictor only at the level of item, it is called "item explanatory (i.e. linear logistic test model)"; if it includes a predictor both at the level of person and item, it is called "doubly explanatory (i.e. the latent regression linear logistic test model)"(De Boeck & Wilson, 2004). Various studies are using these models in the literature. Atar and Çobanoğlu Aktan (2013) added gender, positive attitude towards science, the importance given to science, self-confidence towards science and parents' education level to the model as individual-level predictors to explain the differences between student achievement. Demirkol and Ayvallı Karagöz (2023) compared various explanatory IRT models in which item format and cognitive domain level of items were added as predictors to explain the differences in item difficulty parameters. Demirkol and Kelecioğlu (2022) examined the item position effect and its interaction with various student characteristics (gender, SES, test anxiety, achievement motivation) in a test in reading. Stanke and Bulut (2019) examined individuals' reactions to items by adding various predictor variables at the item level for the verbal aggression (Vansteelandt, 2000) dataset. In the study, type of behaviour (curse, scold, or shout), type of blame (others or self), and blame mode (want or do) were used as item-level explanatory variables. Atar (2011) established explanatory and descriptive IRT models in her study. Accordingly, the variables of gender, positive attitude towards mathematics, giving importance to mathematics and self-confidence in learning mathematics were used as individual-level characteristics in the study. In the same study, two different predictors were used as item-level predictors: cognitive domain (knowledge, application, reasoning) and subject area (numbers, algebra, data analysis and probability, geometry).

Studies on IRT mostly focus on data with items scored in dichotomous form. However, it is more common to use polytomous measurement tools when it comes to ability tests or scales. Polytomous data give more information on response patterns and more detailed insight into the construct to be measured as well as measurement tools (De Boeck & Wilson, 2004; Stanke & Bulut, 2019). It is necessary to use appropriate models to analyze the data that can reveal this extra information. For example, data loss will be inevitable when polytomous data is turned into one having two categories and then analysed (De Boeck & Wilson, 2004; Stanke & Bulut, 2019). The analysis given in the book written by De Boeck and Wilson (2004) about explanatory IRT can be an example of this. In related examples, they compared the fit of various descriptive and explanatory IRT models with the verbal aggression (Vansteelandt, 2000) dataset converted into a dichotomous response format (Wilson & De Boeck, 2004) and its polytomous version (Tuerlinckx & Wang, 2004). Firstly, results obtained from the partial credit model were compared to the latent regression Rasch model, and the coefficients were found to be almost equal, while standard errors regarding the coefficients were found to be 60% more in the dichotomous data set. According to another finding, in the dichotomous data set, gender was not a significant variable, whereas it was found to be significant in the polytomous data set (Tuerlinckx & Wang, 2004). On the other hand, Kim and Wilson (2019) conducted a study to extend the linear logistic test model approach, and they developed two different item explanatory models to find out that polytomous item explanatory IRT models could contribute to test development processes more than descriptive models due to the information they provided on the content of the items.

Literature review shows that there is a limited number of studies on polytomous explanatory IRT (i.e. Kahraman, 2014; Kim & Wilson, 2019, Stanke & Bulut, 2019; Tuerlinckx & Wang, 2004). Tuerlinckx and Wang (2004) conducted the study that founded the basis to formalize and interpret explanatory IRT models via a polytomous data set. In that study, the researchers analysed the data set on verbal aggression via partial credit model, person explanatory partial credit model, rating scale model, person explanatory rating scale model, and explanatory partial credit model including person and item-level predictors. Study findings show that person-explanatory models display a better fit when compared to their traditional counterparts. On the other hand, Stanke and Bulut (2019) conducted a study to make a new parameterization that made it possible to explain the distances between the thresholds by flexing the formulas regarding polytomous item response theory models. In that study, they analysed data sets on verbal aggression via rating scale model, partial credit model, explanatory partial credit model and cross-classified explanatory partial credit model. The study findings show that the partial credit model resulted in a better fit according to the AIC value, while the cross-classified explanatory partial credit model resulted in a better fit according to the BIC value. Another study on polytomous explanatory IRT models was carried out by Kim and Wilson (2019), who developed two different item explanatory IRT models. In that study, the researchers analyzed two different data sets (carbon cycle and verbal aggression). Results of the analysis conducted with the data set on the carbon cycle show that the explanatory many-facet Rasch model resulted in a better fit, while the researchers could not reach a definite result at the end of the analysis conducted with verbal aggression according to AIC and BIC values. Another study which made use of the explanatory IRT approach was conducted by Kahraman (2014). That study which used the data obtained from a performance test in the field of medicine took the advantage of partial credit model to compare explanatory IRT models to which various predictor variables (gender, time to respond, number of the items, test score) were added individually. The study results show that the explanatory partial credit model to which the test score was added as a predictor displayed a better fit to the data.

According to Stanke and Bulut (2019), polytomous explanatory IRT models have mostly focused on the first threshold between item response categories (e.g., Tuerlinckx & Wang, 2004). However, possible variances between the thresholds can be ignored in such a case (Stanke & Bulut, 2019). Therefore, Stanke and Bulut (2019) added a new parameter and flexed the explanatory IRT models. In this context, the log-odd of response in category j instead of $j-1$ given by the individual n for the item i is written as below according to the explanatory partial credit model employed in the current study:

$$\log(\frac{P_n(j)}{P_{n(j-1)}})= Z_n\theta_n - X_n'\delta_i + W_n'\tau_{ii}$$

Here, Z_n represents a matrix used to estimate fixed and random effects related to personal traits θ_n refers to the level of latent qualities of a person, and it has a normal distribution $(\mathrm{N}(\mu_n, \sigma_n^2))$. X_n' is a matrix of item-related information that describes the characteristics of individual items. δ_i is the position of the first threshold between the first and second response categories for the item i. W_n' is a matrix that is used to estimate the fixed and random effects regarding the distances between the thresholds. τ_{ii} refers to the distance between the threshold of $(j-2)/(j-1)$ and $(j-1)/j$ for the item i (Stanke & Bulut, 2019). In the current study, in addition to the explanatory partial credit model, predictions were also made according to the rating scale model (RSM) and the partial credit model (PCM). Accordingly, the model equation for RSM and PCM is given below (Stanke and Bulut, 2019):

$$\log(\frac{P_n(j)}{P_{n(j-1)}}) = \theta_n - (\delta_i + \tau_{ii})$$

In this equation, θ_n and δ_i have the same meaning as the explanatory partial credit model. The only difference between RSM and PCM is that τ_{ii} is the same for all items in RSM (Stanke and Bulut, 2019).

Social and Emotional Skills

Today, continuously changing social, economic and environmental conditions lead to changes in individuals' lives and the flow of social activities. As globalization and digitalization connect people, the world has become a more complicated place full of uncertainties. The content of these skills that are necessary to be successful in such a world and adapt to these changes are also changing in time (Kankaraš & Suarez-Alvarez, 2019; OECD, 2021a). Cognitive skills that are commonly associated with academic achievement are thought to be of prime importance. These skills are very significant as they increase the likelihood of people getting positive results in later life by making them competent in many social and emotional skills such as perseverance, sociability, and self-respect. In today's world, for individuals to become competent in transforming skills, it is important to measure social and emotional skills that interact with cognitive skills in education systems and to take initiatives to support the development of skills accordingly (OECD, 2015). Social and emotional skills are known to be effective in many fields such as academic achievement, productivity in work life or subjective well-being. A high level of social and emotional skills increases trust and tolerance in society, and they lead to a decrease in criminal and anti-social behaviours (Kankaraš & Suarez-Alvarez, 2019; OECD, 2021a). Accepted to be one of the most comprehensive evaluations of these skills in the international arena, the OECD Social and Emotional Skills Study (2021a) was conducted to identify the factors that support or hinder students' social and emotional skills. The study findings were intended to provide the shareholders of education with reliable information (OECD, 2021b). The social and emotional skills specified in that study are described as "individual capacities that can be (a) manifested in consistent patterns of thoughts, feelings and behaviours, (b) developed through formal and informal learning experiences, and (c) important drivers of socioeconomic outcomes throughout the individual's life" (OECD, 2015, p.35). In that study, which was conducted at an international level, the theoretical framework of social and emotional skills relied on the basic components of "big five" personality traits to develop "big five social and emotional skills model." In this context, the basic skills were listed as engaging with others (sub-domains; assertiveness/dominance, sociability, energy/enthusiasm), task performance (sub-domains; persistence, self-control/self-discipline, responsibility/trustworthiness), emotional regulation (sub-domains; optimism/positive emotion, stress resistance/resilience vs. anxiety, emotional control), collaboration (sub-domains; empathy/compassion, trust, co-operation/relationship harmony), open-mindedness (sub-domains; creativity/imagination, tolerance/cultural flexibility, intellectual curiosity) (Kankaraš & Suarez-Alvarez, 2019). The assertiveness skill examined in the current study is associated with expressing one's ideas, feelings and needs responsibly and liking leadership. Individuals having this skill can express their thoughts directly when they disagree with others, and they do not need the guidance of others (Kankaraš & Suarez-Alvarez, 2019). Because of that, being assertive plays an important role in increasing one's level of well-being, while emphasizing individual rights at the same time (Eskin, 2003). As individuals with this skill can more clearly reveal their will (Kankaraš & Suarez-Alvarez, 2019), they can find a solution for their problems at work, school or home more rationally and appropriately. Those who have a high level of assertiveness will have a higher level of self-confidence as well as better decision-making skills, and they will be able to deal with negative feelings such as anger more healthily. This will also have a positive impact on school performance (Sitota, 2018).

It is important to try to explain individual differences as to the social and emotional skills that have an important role in people's family, work and school life and that are also related to cognitive skills. Literature review shows that studies that employed an explanatory IRT mostly focused on cognitive skills (see. Atar, 2011, Atar & Çobanoğlu Aktan, 2013; Briggs, 2008; Büyükkıdık & Bulut, 2022; Demirkol & Ayvallı Karagöz, 2023; Demirkol & Kelecioğlu, 2022; Kahraman, 2014; Kim & Wilson, 2020). However, we think that identifying the variability of social and emotional skills among individuals will make it possible to understand the construct better and accordingly to give more effective feedback to individuals at schools. On the other hand, when the studies that employed an explanatory IRT were evaluated in terms of the qualities of the data set, it was clear that studies mostly employed dichotomous data (see. Atar, 2011; Atar & Çobanoğlu Aktan, 2013; Briggs, 2008; Büyükkıdık & Bulut, 2022). However, most measurement tools that measure latent features have a polytomous data format. In that sense, it is important to make use of explanatory IRT models appropriate for polytomous data sets. Also, the studies on polytomous explanatory IRT generally ignored the distances between the thresholds and focused only on the estimation of the first threshold parameter. This can cause data loss due to ignoring the possible variability between the thresholds (Stanke & Bulut, 2019). In this context, the current study aims at identifying the individual differences in an affective skill via a polytomous explanatory IRT model that makes it possible to estimate the distances between the thresholds. In addition, the explanatory IRT model used is analyzed in comparison with the predictions obtained from RSM and PCM. In line with the study purpose, the research questions are as below:

How are the parameter estimations obtained from the rating scale model regarding the skill of assertiveness?

How are the parameter estimations obtained from the partial credit model regarding the skill of assertiveness?

How are the parameter estimations obtained from the latent regression partial credit model regarding the skill of assertiveness?

What are the significant predictors at the personal level regarding the skill of assertiveness?

How are the results of model-data fit regarding the rating scale model, partial credit model and latent regression partial credit model?

Method

In the current study, we comparatively investigated individual differences regarding the skill of assertiveness within the scope of polytomous descriptive and explanatory IRT models. Therefore, we employed a descriptive study model (Büyüköztürk et al., 2014).

Participants

This study uses data collected in Türkiye as part of the OECD Social and Emotional Skills Study conducted in 2019. Only Istanbul was included in this survey conducted by the OECD. A total of 5869 individuals in the age group of 10 and 15 years participated in the study. We conducted the analysis using the responses of individuals included in the age group of 15 (n=3168), who responded to the items in the Assertiveness Scale. We used the responses given by 2968 individuals after examining the data set in terms of missing data and outliers. The study group included 1764 (59.43%) female and 1204 (40.57%) male students.

Measurement Tools

In the current study, we used the raw data obtained from the sub-scale of assertiveness measured within the scope of the OECD Social and Emotional Skills Study as well as the indices calculated using the raw data. Social and emotional skills were measured using questionnaires administered to the student, the teacher, and the parents. With the raw data from the questionnaires, indices representing different characteristics were calculated (Kankaraš & Suarez-Alvarez, 2019, OECD, 2021b).

In the current study, we used a 5-point (strongly disagree, disagree, neutral, agree, strongly agree) Likert-type scale including 8 items (Item 5 is reverse coded) for the skills of assertiveness, while we did not include item number 7 as it was found to be statistically insignificant in estimations conducted via IRT models. This item was also excluded in the study conducted by OECD as it produced a high tau (slope) value, and it displayed a duplication with item number 6 (OECD, 2021b). The indices we used as a predictor at the person level were perceived relationships with teachers, bullying at school, sense of belonging at school, global mindedness, and test anxiety. Furthermore, the other two predictor variables of the study were gender and socio-economic level. Table 1 below gives the items of the Assertiveness Scale used in the study.

Table 1.

Scale Items

Item Number	Item Content
1	A leader
2	Want to be in charge
3	Know how to convince others to do what I want
4	Enjoy leading others
5	Dislike leading a team
6	Like to be a leader in my class
7	-
8	Dominant, and act as a leader

Table 1 above presents information about the content of scale items as explained in the technical report published by OECD. As item number 7 was excluded from the study at the end of the analysis, the report did not include content about this item. Table 2 below shows the content of the variables that were used as predictors in the study (see. OECD, 2021b).

Table 2.

Content about the Person-Level Predictors

Name of the Index	Question	Alternatives
Perceived relationships with teachers	During the past 12 months, how often did you have the following experiences at school?	Most of my teachers treated me fairly. I got along well with most of my teachers. Most of my teachers were interested in my well-being.
Bullying at school	During the past 12 months, how often have you had the following experiences in school?	Other students made fun of me. I was threatened by other students. Other students took away or destroyed things that belonged to me. I got hit or pushed around by other students.
Sense of belonging at school	Thinking about your school: To what extent do you agree with the following statements?	I feel like an outsider (or left out of things) at school. I make friends easily at school. I feel like I belong at school. I feel awkward and out of place in my school. Other students seem to like me. I feel lonely at school.
Global mindedness	How informed are you about the following topics?	Climate change and global warming Global health (e.g. epidemics) International conflicts Causes of poverty Equality between men and women in different parts of the world
Test anxiety	To what extent do you agree or disagree with the following statements about yourself?	I often worry that it will be difficult for me to take a test. Even if I am well prepared for a test I feel very anxious. I get very tense when I study for a test.

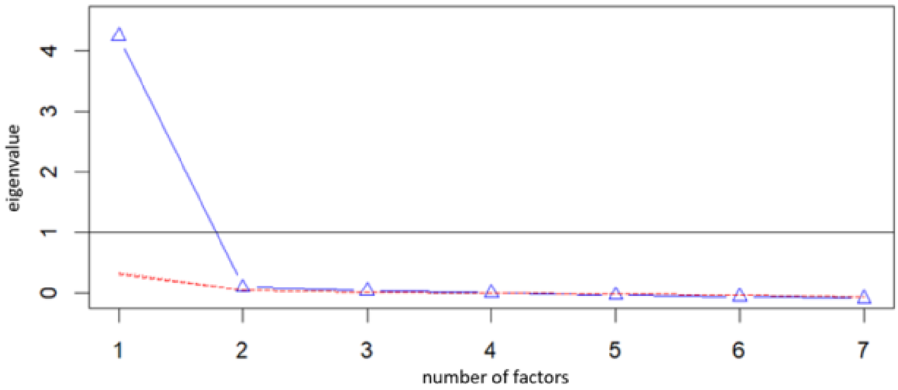
Data Analysis

In the current study, we conducted data cleaning and checked the assumptions on R software (R Core Team, 2022) via the packages of haven (Wickham et al., 2022), stringr (Wickham, 2022), olsrr (Hebbali, 2020), dplyr (Wickham et al., 2022), ltm (Rizopoulos, 2006), psych (Revelle, 2022), MVTests (Bulut, 2019), ggplot2 (Wickham, 2016), mirt (Chalmers, 2012). For estimations relating to the partial credit model, rating scale model, and latent regression partial credit model, the eirm package (Bulut, 2021) that is also included in R software was used. We examined the data and checked the assumptions as stated below:

We excluded 37 individuals based on missing value analysis and 163 individuals based on univariate and multivariate outlier analysis. We then conducted a multicollinearity analysis. Simple pairwise correlation values between independent variables (ranging from -0.261 to 0.291), tolerance values (ranging from 0.826 to 0.961) and variance influence factor values (ranging from 1.039 to 1.209) showed that there was no multicollinearity problem. In the final case, we continued the analyses with data from 2968 individuals. On the other hand, Cronbach's alpha value calculated for reliability was found to be 0.603[1]. For Cronbach alpha, a value equal to or higher than 0.70 is acceptable, while a value higher than 0.80 implies a high level of reliability (Nunnally & Bernstein, 1994). Hence, it is possible to state that the reliability value calculated in the current study was low. This might result from the length of the test which is a factor that affects reliability. It is known that a short test affects reliability in a negative way (Crocker & Algina, 1986). In the current study, the reliability value for the data set belonging to the 7 items was low, which we think might be due to the shortness of the test[2].

Finally, we tested the assumptions of unidimensionality and local independence, which are necessary assumptions for item response theory analyses. First, we conducted an exploratory factor analysis and parallel analysis to test the unidimensionality assumption. To simplify the narrative, only parallel analysis results are given[3]. For this analysis, we used the function of fa.parallel in the package psych of R software (Revelle, 2022). Figure 1 below shows the related results.

Figure 1.
Results of parallel analysis



In Figure 1, the blue line refers to values regarding the real data, and the red line refers to values regarding the data produced randomly. One of the factors obtained from the real data set has an eigenvalue noticeably higher than the eigenvalues of data produced randomly. In this case, it is possible to state that the scale has only one factor. When all the results are taken into consideration, the factor loads of items of the scale which was decided to be unidimensional are 0.811; 0.549; 0.428; 0.904; -0.877; 0.877; 0.862 respectively. Secondly, to test the local independence assumption, we performed parameter estimations for the rating scale model and the partial credit model. We examined correlation

[1] The Cronbach's alpha value calculated without removing the 7th item from the data set was 0.725.

[2] In the technical report published by the OECD, both alpha and omega coefficients for the assertiveness subscale for the 15-year-old age group are reported as 0.88 (OECD, 2021b).

[3] Exploratory factor analysis (EFA) was also conducted to examine the unidimensionality assumption. According to the EFA results, the assertiveness subscale was found to be unidimensional (eigenvalue of the first factor=4.550, variance explained=61%).

values between the residuals. According to the results obtained from the rating scale model, these values range between -0.789 and 0.421, and the results obtained from the partial credit model range between -0.670 and 0.408. As stated by Christensen et al. (2016), the studies in the literature mostly use the critical value of 0.2 suggested by Chen and Thissen (1997) for local independence. The values higher than this value are said to violate local independence. However, other critical values are also used (Christensen et al., 2016). In this context, Christensen et al. (2016) mentioned some studies in which critical values of 0.1; 0.3; 0.5 and even 0.7 were used. On the other hand, it is stated that local independence can be violated as personality assessments include very similar items (Steinberg & Thissen, 1996, cited by Embretson & Reise, 2000, p.232). According to that, considering all the studies conducted previously, we concluded that the assumption of local independence in the scale was not violated.

Preparing the Data Set for Analysis

In this part, the procedures for making the data set suitable for analysis are explained. Estimations regarding dichotomous and polytomous explanatory IRT models can be done via package eirm (Bulut, 2021) which conducts transactions through function glmer in package lme4 (Bates et al., 2015). In explanatory IRT models, as items are nested in persons, it is necessary to turn data into a long format in which there are answers about one item in each line and each person has more than one line. On the other hand, to use the functions in the package, the responses should display binominal distribution, and so it should be dichotomous. Therefore, the responses should first be transformed into multiple dichotomous formats to analyse polytomous data. These two processes can be done with the function polyreformat (Bulut et al., 2021). Table 3 gives information on multiple dichotomous coding conducted to analyse the 5-category data set.

Table 3.
Transforming Polytomous Responses into Multiple Dichotomous Responses

Original response	"I disagree"	"Neutral"	"I agree"	"I strongly agree"
Strongly disagree	0	NA	NA	NA
Disagree	1	0	NA	NA
Neutral	NA	1	0	NA
Agree	NA	NA	1	0
Strongly agree	NA	NA	NA	1

According to the recording given in Table 3, for example, for an individual who responded to an item in the strongly disagree category, the responses transformed into multiple dichotomous responses were recoded as 0 in the "disagree" category and NA in the other categories. Table 4 below shows the R codes used in the study.

Table 4.
R codes used in the Study

Model	R codes
Rating Scale Model	rsm <- eirm(formula = "polyresponse ~ -1 + item + polycategory + (1|person)", data = long_data)
Partial Credit Model	pcm <- eirm(formula = "polyresponse ~ -1 + item + item:polycategory + (1|person)", data = long_data)
Latent Regression Partial Credit Model	lrm_poly <- eirm(formula = "polyresponse ~ -1 + item + item:polycategory + gender+relteach+ bully+belong+ global+ anxtest+ SES+ (1|person)", data = long_data)

Results

In this part, we explained the results obtained from the rating scale, partial credit, and latent regression item response theory model according to the research questions. The tables show the estimations regarding the easiness parameters.

Table 5.

Estimations Regarding Rating Scale Model

Item Number	**Location for Strongly Disagree/Disagree**			**Distance for Disagree/Neutral****			**Distance for Neutral/Agree**			**Distance for I agree/I strongly agree**		
	Estimation	SE	Z value	Estimation	SE	Z value	Estimation	SE	Z value	Estimation	SE	Z value
1	0.580	0.036	16.072*	-0.561	0.033	-17.196*	-0.689	0.033	-21.149*	-1.195	0.034	-34.729*
2	0.804	0.037	21.512*									
3	1.130	0.040	28.392*									
4	0.732	0.037	19.781*									
5	0.325	0.035	9.315*									
6	0.681	0.037	18.592*									
7	0.614	0.036	16.953*									

*p<0.05

**In empty cells, the values written in the relevant columns are repeated. For clarity, we have not repeated the same values.

Table 5 gives the results obtained from the rating scale model for the sub-scale of assertiveness. According to that, item number 3 (Know how to convince others to do what I want) with logit 1.130 is the item which has the easiest likelihood to respond in a higher category according to location parameter estimation for the threshold of “I strongly disagree/I disagree”. For this item, it seems exp(1.130)= 3.095 times easier to respond in the category of “I disagree” instead of “I strongly disagree”. In the rating scale model, distance value is estimated the same in all items for each category. According to the distance value obtained for the threshold of “I disagree/I am neutral”, after checking the level of assertiveness, it is exp(-0.561)=0.570 times easier for items to respond in the category of “I am neutral” than the categories of “I strongly disagree” and “I disagree”. According to the distance value obtained for the threshold of “I am neutral/I agree”, after checking the level of assertiveness, it is exp(-0.689)=0.502 times easier for items to respond in the category of “I agree” than the categories of “I strongly disagree” and “I disagree”. Lastly, according to the distance value obtained for the threshold of “I agree/I strongly agree”, after checking the level of assertiveness, it is exp(-1.195)=0.302 times easier for items to respond in the category of “I strongly agree” than the categories of “I strongly disagree” and “I disagree”.

Table 6.

Estimations Regarding Partial Credit Model

Item Number	Location for Strongly Disagree/Disagree			Distance for Disagree/Neutral			Distance for Neutral/Agree			Distance for I agree/I strongly agree		
	Estimation	SE	Z value	Estimation	SE	Z value	Estimation	SE	Z value	Estimation	SE	Z value
1	0.563	0.061	9.289*	-0.562	0.080	-7.056*	-0.703	0.081	-8.694*	-1.086	0.088	-12.306*
2	1.525	0.098	15.600*	-1.092	0.111	-9.827*	-1.512	0.108	-13.946*	-2.288	0.114	-20.070*
3	1.485	0.145	10.215*	-0.434	0.162	-2.672*	-1.082	0.153	-7.073*	-1.765	0.152	-11.578*
4	0.490	0.064	7.644*	-0.636	0.087	-7.355*	-0.106	0.085	-1.254	-0.842	0.084	-10.001*
5	0.204	0.050	4.054*	-0.521	0.072	-7.193*	-0.577	0.080	-7.231*	-0.597	0.090	-6.605*
6	0.661	0.065	10.179*	-0.606	0.084	-7.242*	-0.735	0.084	-8.770*	-0.989	0.088	-11.259*
7	0.620	0.062	9.963*	-0.605	0.081	-7.471*	-0.734	0.082	-8.972*	-1.094	0.088	-12.400*

*p<0.05

Table 6 gives the results obtained according to the partial credit model for the sub-scale of assertiveness. There is one thing to be careful about while interpreting the distance values. According to that, while calculating threshold values except for the location parameter of the related item, it is necessary to add the distance value estimated in each category and the value estimated for the location parameter (1st threshold). For instance, the location value for the first item is 0.563. For the same item, the estimated distance value for the category of “I disagree/I am neutral” is -0.562. This value points to the distance between the first and second thresholds for the first item. When these two values are added (0.563+(-0.562)), the second threshold is obtained. In other words, the first threshold value is taken as a reference to make calculations to find each threshold. According to that, item number 2 (Want to be in charge) with logit 1.525 is the item which has the easiest likelihood of responding in a higher category. For this item, it seems exp(1.525)= 4.595 times easier to respond in the category of “I disagree” instead of “I strongly disagree”. Unlike the rating scale model, the distance value is estimated differently in all items for each category when it comes to the partial credit model. According to the distance value obtained for the threshold of “I disagree/I am neutral”, after checking the level of assertiveness, it is exp(0.563+(-0.562)) =exp(1.051)=2.860 times easier for the third item to respond in the category of “I am neutral” than the categories of “I strongly disagree” and “I disagree”. According to the distance value obtained for the threshold of “I am neutral/I agree”, after checking the level of assertiveness, it is exp (0.403) =1.496 times easier for the third item to respond in the category of “I agree” than the categories of “I strongly disagree” and “I disagree”. Lastly, according to the distance value obtained for the threshold of “I agree/I strongly agree”, after checking the level of assertiveness, it is exp (-0.281)=0.755 times easier for the third item to respond in the category of “I strongly agree” than the categories of “I strongly disagree” and “I disagree”.

Table 7.
Estimations Regarding Latent Regression Partial Credit Model

Item Number	**Location for Strongly Disagree/Disagree**			**Distance for Disagree/Neutral**			**Distance for Neutral/Agree**			**Distance for I agree/I strongly agree**		
	Estimation	SE	Z value	Estimation	SE	Z value	Estimation	SE	Z value	Estimation	SE	Z value
1	-0.354	0.126	-2.810*	-0.580	0.080	-7.264*	-0.751	0.081	-9.245*	-1.180	0.089	-13.289*
2	0.617	0.147	4.199*	-1.109	0.111	-9.971*	-1.561	0.108	-14.390*	-2.379	0.114	-20.827*
3	0.589	0.182	3.231*	-0.453	0.163	-2.784*	-1.126	0.153	-7.343*	-1.849	0.153	12.081*
4	-0.430	0.128	-3.368*	-0.650	0.087	-7.494*	-0.139	0.085	-1.633	-0.914	0.085	-10.792*
5	-0.781	0.124	-6.321*	-0.482	0.073	-6.634*	-0.520	0.080	-6.485*	-0.530	0.091	-5.838*
6	-0.253	0.128	-1.980*	-0.622	0.084	-7.415*	-0.770	0.084	-9.267*	-1.079	0.088	-12.214*
7	-0.298	0.127	-2.349*	-0.619	0.081	-7.615*	-0.778	0.082	-9.465*	-1.182	0.089	-13.326*

*p<0.05

Table 7 gives the results obtained from the latent regression partial credit model for the sub-scale of assertiveness. Distance values are calculated as in the partial credit model. According to that, considering the location parameters values, item number 2 (Want to be in charge) with logit 0.617 is the item which has the easiest likelihood to respond in a higher category. For this item, it seems exp (0.617)=1.53 times easier to respond in the category of “I disagree” instead of “I strongly disagree”. According to the distance value obtained for the threshold of “I disagree/I am neutral”, after checking the level of assertiveness, it is exp(0.135)=1.144 times easier for the third item (Know how to convince others to do what I want) to respond in the category of “I am neutral” instead of the categories of “I strongly disagree” and “I disagree”. According to the distance value obtained for the threshold of “I am neutral/I agree”, after checking the level of assertiveness, it is exp(-0.538)=0.583 times easier for the third item to respond in the category of “I agree” than the categories of “I strongly disagree” and “I disagree”. Lastly, according to the distance value obtained for the threshold of “I agree/I strongly agree”, after checking the level of assertiveness, it is exp(-1.260)=0.283 times easier for the third item to respond in the category of “I strongly agree” than the categories of “I strongly disagree” and “I disagree”.

Table 8 below shows the estimation regarding the predictors in the latent regression partial credit model in which analysis is conducted by including predictors at the person level.

Table 8.
Estimations Regarding Predictors

Predictors	**b**	**SE**	**exp (b)**	**Z value**
Gender	0.045	0.023	1.046	1.929
Perceived relationships with teachers	0.000	0.001	1.000	0.056
Bullying at school	0.002	0.001	1.002	1.660
Sense of belonging at school	0.009	0.001	1.009	8.194*
Global mindedness	0.009	0.001	1.009	8.109*
Test anxiety	-0.001	0.000	0.999	-1.544
Socio-economic level	0.028	0.012	1.028	2.306*

*p<0.05

According to the results, the predictors of a sense of belonging at school, global mindedness and socio-economic level were found to be significant, and the predictors of gender, perceived relationships with teachers, bullying at school and test anxiety were found to be insignificant. According to that, for a 1-unit change in sense of belonging at school, the likelihood of getting a higher score from assertiveness is 1.009 times more. For a 1-unit change in the level of global mindedness, the likelihood of getting a higher score from assertiveness is 1.009 times more. Lastly, for a 1-unit change in the level of socio-economic level, the likelihood of getting a higher score from assertiveness is 1.028 times more. In other words, it is possible to state that individuals with a more advantageous socio-economic level are more likely to have a higher level of assertiveness.

Table 9.
Results of Model-Data Fit

Model	AIC	BIC	Deviation
Rating Scale Model	47996.6	48089.9	47974.6
Partial Credit Model	47680.0	47926.0	47622.0
Latent Regression Partial Credit Model	47502.3	47807.7	47430.3

Table 9 gives the results of model-data fit regarding the three models developed in the current study. As the models were not nested, they were compared to relative fit indices. According to that, the results were examined according to the Akaike Information Criterion (AIC) (Akaike, 1974), Bayesian Information Criterion (BIC) (Schwarz, 1978) and deviation values. Having a small index refers to a better model-data fit. According to that, the latent regression explanatory partial credit model is the one with the lowest AIC, BIC, and deviation values. In other words, the model that best fits the data is the latent regression partial credit model.

Discussion and Conclusion

In the current study, individual differences in assertiveness skills of individuals were examined through the rating scale model, partial credit model and latent regression partial credit model. We interpreted the study findings within the framework of parameter estimations and model-data fit. According to the first study finding which we obtained from the rating scale model and partial credit model, neither of which included a predictor, the partial credit model showed a better fit to the data. According to the distance values obtained from the partial credit model, it was easier for the third item to respond in a category in the 2nd, 3rd and 4th thresholds, while it was easier for the second item to respond in a category in the 1st threshold. Therefore, there was a variance in the distance values. In this context, according to the rating scale model, which makes the same estimations as to all items for each category distance, it is possible to state that the partial credit model gives deeper information and reveals variances between items better. This study finding is supported by a previous study conducted by Tuerlinckx and Wang (2004). In that study, the researchers concluded that the partial credit model provides a better fit than the rating scale model. This study finding is partially in line with the results of the study conducted by Stanke and Bulut (2019). The researchers in that study reported that when the items in the data set of that study were estimated via a partial credit model, the resulting values varied between -0.10 and 1.13. According to that, the researchers compared the estimations of distance values obtained from the partial credit model and rating scale model and concluded that the partial credit model explained the variance between the items more. Also, in the same study, the partial credit model produced a better fit according to the AIC value, whereas the rating scale model produced a better fit according to the BIC value.

Secondly, the current study reveals that the items which are easier to respond to in a higher category are the same according to the partial credit model and latent regression partial credit model. Thus, we concluded in the current study that the results of these two models were overlapping in terms of location and distance value estimation.

According to the third finding of the current study, the best-fitting model for all indices is the latent regression partial credit model. The fact that explanatory IRT models are models that can simultaneously decompose the covariance between persons and items in addition to analyzing items (Briggs, 2008) by adding predictors related to items for differences in item difficulties and/or adding predictors related to persons for differences between persons (De Boeck & Wilson, 2004) was thought to be effective in the emergence of a tendency to better fit the data. Furthermore, it is stated in the literature that one of the reasons why explanatory IRT models have a better fit is the increase in the number of estimated parameters (De Boeck & Wilson, 2004). Stanke and Bulut (2019) came up with a similar finding in their study. In that study, they compared the explanatory partial credit model and cross-classified explanatory partial credit model and found out that the second model produced a better fit. This was thought to result from the latter model including more parameters. When all models developed in the study were compared, they concluded the model that produced the best fit according to AIC value was the partial credit model, while the model that produced the best fit according to BIC value was the cross-classified explanatory partial credit model. Tuerlinckx and Wang (2004) developed the rating scale model, partial credit model and person-explanatory partial credit model, and they compared the values obtained from these models to find out that person explanatory partial credit model produced the best fit of all. On the other hand, Kahraman (2014) found that the explanatory partial score model including the test score variable provided a better fit. Another important finding belongs to a study carried out by Briggs (2008) to explain the differences in achievement in science. In that study, after conducting ability estimation through one multidimensional Rasch model, the researcher conducted linear regression by taking ability scores obtained from these estimations as the dependent variable and race/ethnicity as the independent variable. Then the researcher compared the results of this analysis which was stated to be a two-step approach and the estimations were done via the latent regression Rasch model. It was stated that if the reliability of the test scores is high, the results of the two-step approach and explanatory IRT approach will overlap. On the other hand, when reliability is low-medium, the ability estimates in the two-stage approach will narrow towards the population mean and therefore the regression coefficients obtained in the second stage will be weakened due to measurement error. The researcher suggested using an explanatory IRT approach when the aim is to identify group differences. In addition, it was emphasized that the explanatory approach allows for more detailed interpretations of achievement differences at the group level depending on race/ethnicity. Büyükkıdık and Bulut (2022) conducted a study to model individuals' achievement in science through test, student, and school-related predictors, and they developed explanatory IRT models including the Rasch model and various predictors. The study results showed that the best fit was obtained from the explanatory model including the variables of gender and school type. Atar and Çobanoğlu Aktan (2013) also carried out a study in which they comparatively examined the differences in achievement in science using the latent regression two-parameter logistic model and traditional two-parameter logistic model, and they found that the latent regression two-parameter model produced a better fit.

Fourthly, according to the results of the latent regression partial credit model including person-level predictors, the predictors of sense of belonging at school, global mindedness and socio-economic level were found to be significant, whereas the values regarding the predictors of gender, bullying at school, perceived relationships with teachers and test anxiety were found to be insignificant. According to the result related to school belonging, it was observed that the higher the sense of belonging to the school, the higher the probability of having a higher level of assertiveness. According to the Social and Emotional Skills Turkey results published by OECD (2021a), a statistically significant positive relationship was found between a sense of belonging to school and assertiveness in the 15-year-old age group. On the other hand, for the variable of the level of awareness of global events, which was found to be significant in the study, it was found that the higher the level of awareness, the higher the probability of receiving higher assertiveness scores. However, the report published by OECD (2021a) did not include an explanation of this variable being significant. In the study, it was also found that there was a positive relationship between socioeconomic level and assertiveness scores, which means that those having a more advantageous socioeconomic level are more likely to have a higher score of assertiveness. A similar result was obtained from OECD (2021a) Türkiye results. There are various studies in the literature that focus on the relationship between assertiveness and socioeconomic level. Kılıç (2009), one of these studies, concluded that the level of assertiveness varied statistically significantly according to the perceived economic level. It was found in that study that individuals included in the middle income and over-middle/high-income families had a higher level of assertiveness than those included in under-middle/low-income families at a statistically significant level. These findings can be attributed to the fact that families from a more advantageous socioeconomic background could invest more in the development of their kids' social and emotional skills (OECD, 2021a).

Gender is one of the variables that was not found to be significant in the current study. In the study, the mean score of girls was found to be 21.535, whereas it was 22.379 for boys. Studies on the relationships between assertiveness and gender have various results. Some studies reported that male participants had a higher score of assertiveness than female participants (Karaaslan Arkan, 2016; OECD, 2021a; Sitota, 2018). On the other hand, some other studies concluded that there was no significant difference between assertiveness and gender (Eskin, 2003; Kılıç, 2009). In some studies that concluded that there was no significant difference, it was found that the assertiveness scores of males were slightly higher than females (Castedo et al., 2015; Kaya & Karaca, 2018). It is striking to see that males have a higher level of assertiveness when studies that examine the relationships between assertiveness and gender are in question. Furthermore, the level of significance regarding this relationship varies from study to study. In the current study, the fact that males had slightly higher assertiveness scores than females, but that there was no significant difference between them, overlaps with the findings in the literature to a certain extent. However, we think that the diversity of findings among the studies in the literature may be because each study has a different study group, uses different analysis techniques, etc. Another variable of the current study, perceived relationships with teachers wasn't found to be a significant predictor. However, the OECD (2021a) report indicated that there was a statistically significant positive relationship between the said variable and the skill of assertiveness. On the other hand, Stake et al. (1983) found a significant increase in the self-esteem levels of individuals who thought that they received more positive reactions from their teachers due to their assertive behaviours. In the current study, we found out that there was a negative but insignificant relationship between assertiveness and exposure to bullying, whereas it was stated in the OECD report (2021a) that there was a significant negative relationship between the two variables. Similarly, Keliat et al. (2015) found a low-level statistically significant negative relationship between stories of abuse and assertiveness. Lastly, the variable of test anxiety was not found to be a significant predictor in the current study. A similar result was reported in the OECD (2021a) study.

When the results of the current study are evaluated together, it is seen that the explanatory IRT model used in the study, in line with the literature, shows a better fit and explains the differences between individuals and the variability in the data better. In addition to that, although some of the estimation results about the predictors are in parallel with the literature, there are also some differences. There are some disagreements, especially with the results reported by OECD (2021a). We think that this might result from using different techniques of analysis.

In conclusion, in light of the findings of the current study, we believe that explanatory IRT models can contribute to improving the scope of studies on the latent traits targeted to be measured. For instance, for a polytomous data set, as is the case in the current study, revealing the potential variance between categories can contribute to developing measurement tools more felicitously. Also, identifying the variables that are related to the skill of assertiveness can help enrich educational content and preventive guidance activities that can be presented for this skill at schools. A clearer understanding of the construct and the contextual factors associated with it can support formal and non-formal education activities involving students, teachers, and parents.

Limitations and Suggestions

The study has several limitations. Only personal characteristics were included as predictors in the study. In other studies, both items and personal characteristics can be added as predictors. In addition, the rating scale model, partial credit model and latent regression partial credit models were compared in the study. Different models can be established in other studies. On the other hand, only the responses of 15-year-olds to the assertiveness scale of the OECD Social and Emotional Skills survey were used in the study. Similar studies can be conducted with other sub-skills in the OECD study and/or with the 10-year-old age group. At the same time, studies comparing the two age groups can be conducted. Finally, the eirm

(Bulut, 2021) package was used for the analysis in this study. This package only analyzes explanatory IRT models based on the Rasch model family. Therefore, different explanatory IRT models including discrimination parameters cannot be estimated with this package (Bulut et al., 2021).

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723. https://doi.org/10.1109/TAC.1974.1100705.

Andrich, D. (1978a). Application of a psychometric model to ordered categories which are scored with successive integers. *Applied Psychological Measurement,* 2(4), 581-594. https://doi.org/10.1177/014662167800200413.

Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika, 43* (4), 561-573. https://doi.org/10.1007/BF02293814.

Atar, B. (2011). Tanımlayıcı ve açıklayıcı madde tepki modellerinin TIMSS 2007 Türkiye matematik verisine uyarlanması. *Eğitim ve Bilim, 36* (159), 255-269.

Atar, B., & Çobanoğlu Aktan, D. (2013). Birey açıklayıcı madde tepki kuramı analizi: Örtük regresyon iki parametreli lojistik modeli. *Eğitim ve Bilim, 38* (168), 59-68.

Baker, F. B. (2016). *Madde tepki kuramının temelleri*. (N., Güler, Trans Ed.) & (M., İlhan, Trans.). Ankara. https://doi.org/10.14527/9786053185673.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01.

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education,* *21*(2), 89 - 118. https://doi.org/10.1080/08957340801926086.

Bulut, H. (2019). An R Package for Multivariate Hypothesis Tests: MVTests. *Technological Applied Sciences, 14* (4), 132-138. https://doi.org/10.12739/NWSA.2019.14.4.2A0175.

Bulut, O. (2021). *eirm: Explanatory item response modeling for dichotomous and polytomous item responses*. https://CRAN.R-project.org/package=eirm.

Bulut, O., Görgün, G., & Yıldırım Erbaşlı, S. N. (2021). Estimating explanatory extensions of dichotomous and polytomous Rasch models: The eirm package in R. *Psych,* *3*(3), 308-321. https://doi.org/10.3390/psych3030023.

Büyükkıdık, S., & Bulut, O. (2022). Analyzing the effects of test, student, and school predictors on science achievement: An explanatory IRT modelling approach. *Journal of Measurement and Evaluation in Education and Psychology, 13*(1), 40-53. https://doi.org/10.21031/epod.1013784.

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö., Karadeniz, Ş., & Demirel, F. (2014). *Bilimsel araştırma yöntemleri*. Pegem Akademi.

Castedo, A. L., Juste, M. P., & Alonso, J. D. (2015). Social competence: Evaluation of assertiveness in Spanish adolescents. *Psychological reports, 116* (1), 219-229. https://doi.org/10.2466/21.PR0.116k12w5.

Chalmers, R.P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48* (6), 1-29. https://doi.org/10.18637/jss.v048.i06.

Christensen, K.B., Makransky, G., & Horton, M.C. (2017) Critical values for yen's q3: identification of local dependence in the rasch model using residual correlations. *Applied Psychological Measurement*, *41* (3), 178-194. https://doi.org/10.1177/0146621616677520.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer-Verlag.

Demirkol, S., & Ayvallı Karagöz, M. (2023). PISA 2015 okuma becerisi maddelerinin güçlük indeksini etkileyen madde özelliklerinin incelenmesi. *Ana Dili Eğitimi Dergisi,* *11*(3), 567-579. https://doi.org/10.16916/aded.1212049

Demirkol, S., & Kelecioğlu, H. (2022). Analyzing the interaction of item position effect and student characteristics within explanatory IRT models. *Journal of Measurement and Evaluation in Education and Psychology, 13*(4), 282-304. https://doi.org/10.21031/epod.1126368

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

Eskin, M. (2003). Self-reported assertiveness in Swedish and Turkish adolescents: A cross-cultural comparison. *Scandinavian Journal of Psychology, 44*(1), 7-12. https://doi.org/10.1111/1467-9450.t01-1-00315.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer Science & Business Media.

Hebbali A. (2020). *olsrr: Tools for building OLS regression models*. https://CRAN.R-project.org/package=olsrr.

Kahraman, N. (2014). An explanatory item response theory approach for a computer-based case simulation test. *Eurasian Journal of Educational Research, 14* (54), 117-134. https://doi.org/10.14689/ejer.2014.54.7.

Kankaraš, M., & Suarez-Alvarez, J. (2019). *Assessment framework of the OECD study on social and emotional skills*. OECD Publishing. https://doi.org/10.1787/5007adef-en.

Karaaslan Arkan, R. (2016). *Ergenlerin atılganlık, utangaçlık ve yalnızlık düzeyleri*. [Unpublished master thesis]. Beykent Üniversitesi.

Kaya, Z., & Karaca, R. (2018). Ergenlerin atılganlık ve sürekli kaygı düzeylerinin bazı değişkenlere göre incelenmesi. *Van Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, 15*(1), 1490-1516. https://doi.org/10.23891/efdyyu.2018.113

Keliat, B. A., Tololiu, T. A., Daulima, N. H. C., & Erawati, E. (2015). Effectiveness assertive training of bullying prevention among adolescents in West Java Indonesia. *International Journal of Nursing, 2*(1), 128-134. https://doi.org/10.15640/ijn.v2n1a14.

Kılıç, G. (2009). *Lise öğrenimi görmekte olan ergenlerin atılganlık düzeylerinin ebeveynlerine bağlanma örüntülerine ve bazı demografik değişkenlere göre incelenmesi: Darıca ilçesi örneği*. [Unpublished master thesis]. Maltepe Üniversitesi.

Kim, J., & Wilson, M. (2020). Polytomous item explanatory item response theory models. *Educational and Psychological Measurement, 80*(4), 726-755. https://doi.org/10.1177/0013164419892667.

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

Nunnally, J., & Bernstein, I. (1994), *Psychometric Theory*, McGraw-Hill.

OECD. (2015). Skills for social progress: The power of social and emotional skills. OECD Publishing. https://doi.org/10.1787/9789264226159-en.

OECD. (2021a). Beyond academic learning: First results from the survey of social and emotional skills. OECD Publishing. https://doi.org/10.1787/92a11084-en.

OECD. (2021b). OECD Survey on social and emotional skills technical report. OECD Publishing.

R Core Team (2022). *R: A language and environment for statistical computing*. https://www.R-project.org/.

Revelle, W. (2022) *Psych: procedures for personality and psychological research*. https://CRAN.R-project.org/package=psych.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17* (5), 1-25. https://doi.org/10.18637/jss.v017.i05.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34* (1), 1-97. https://doi.org/10.1007/BF03372160.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136.

Sitota, G. (2018). Assertiveness and academic achievement motivation of adolescent students in selected secondary schools of Harari peoples regional state, Ethiopia. *International Journal of Education and Literacy Studies, 6*(4), 40-46. https://doi.org/10.7575/aiac.ijels.v.6n.4p.40.

Stake, J. E., DeVille, C. J., & Pennell, C. L. (1983). The effects of assertive training on the performance self-esteem of adolescent girls. *Journal of Youth and Adolescence, 12*(5), 435-442. https://doi.org/10.1007/BF02088725.

Stanke, L., & Bulut, O. (2019). Explanatory item response models for polytomous item responses. *International Journal of Assessment Tools in Education, 6*(2), 259–278. https://doi.org/10.21449/ijate.515085.

Tuerlinckx, F., & Wang, W.C. (2004). Models for polytomous data. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 75–109). Springer-Verlag.

Vansteelandt, K. (2000). *Formal models for contextualized personality psychology* [Unpublished doctoral dissertation]. K.U.Leuven.

Wickham H. (2022). *stringr: Simple, consistent wrappers for common string operations*. https://CRAN.R-project.org/package=stringr.

Wickham H., François R., Henry L., & Müller K. (2022). *dplyr: A grammar of data manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham H., Miller E., & Smith D. (2022). *haven: Import and export 'SPSS', 'Stata' and 'SAS' files*. https://CRAN.R-project.org/package=haven.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer.

Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 43–74). Springer-Verlag