# Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

*ISSN: 1309-6575* *Yaz* 2024 *Summer* 2024 *Cilt: 15-Sayı: 2* *Volume: 15-Issue: 2*

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309 – 6575

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

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

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

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

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

**Hakem Kurulu / Referee Board**

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

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

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Haziran 2024, Sayı: 15-2)

Journal of Measurement and Evaluation in Education and Psychology (June 2024, Issue: 15-2)

# İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
Review; 2024; 15(2); 79-93

# A Systematic Review of Factor Mixture Model Applications

Sedat ŞEN* Allan S. COHEN**

**Abstract**

In this study, a systematic review was conducted on peer-reviewed articles of factor mixture model (FMM) applications. A total of 304 studies were included with 334 applications published from 2003–2022. FMM was mostly used in these studies to detect latent classes and model heterogeneity. Most of the studies were conducted in the U.S. with samples including students, adults, and the general population. The average sample size was 3,562, and the average number of items was 17.34. Measurement tools containing mostly Likert type items and measuring structures in the field of psychology were used in these FMM analyses. Most FMM studies that were reviewed were applied with maximum likelihood estimation methods as implemented in Mplus software. Multiple fit indices were used, the most common of which were AIC, BIC, and entropy. The mean numbers of classes and factors across the 334 applications were 2.96 and 2.17, respectively. Psychological and behavioral disorders, gender, and age variables were mostly the focus of these studies and included use of covariates in these analyses. As a result of this systematic review, the trends in FMM analyses were better understood.

*Keywords: mixture models, factor mixture model, systematic review*

## Introduction

The factor mixture model (FMM; Muthén & Shedden, 1999) is a combination of common factor and latent class models (Lubke & Muthén, 2005). The FMM is a type of mixture model and comprises a family of statistical models useful for evaluating data in which there may be multiple latent variables that underlie the observed variables. FMMs can provide a powerful tool for analyzing data where multiple unobserved variables may be influencing the observed variables. As with other latent variable mixture models, FMM is also a flexible analytical method that enables researchers to explore research problems about data patterns and assess the degree to which observed patterns are related to important variables (Berlin et al., 2013). Typically, FMMs attempt to estimate latent classes within a sample based on the response patterns (i.e., the observed variables) respondents have made to a given set of items. Thus, they are considered "person-centered" statistical methods as the detection of latent classes is based on person characteristics. FMMs are often used to explain unobserved population heterogeneity as well as to detect latent classes by relaxing the assumption that all respondents in the sample are drawn from the same population. The latent classes may differ either qualitatively or quantitatively or both.

Unlike latent class analysis (LCA; Lazarsfeld & Henry, 1968) and exploratory factor analysis (EFA; Spearman, 1904), FMMs have the flexibility to model hybrids of continuous latent variables (factors) and categorical latent variables (latent classes). Thus, FMMs are also sometimes known as hybrid latent variable models. In FMM, the factor analysis part seeks to uncover shared latent content (i.e., factors) among the observable variables, the latent class analysis part is intended to identify latent subgroups or latent classes of a study population. In addition to FMM, several models with different names have been developed in the literature based on combining categorical and continuous latent variables. These fall

* Assoc. Prof. Dr., Harran University, Faculty of Education, Şanlıurfa-Turkey, sedatsen@harran.edu.tr, ORCID ID: 0000-0001-6962-4960
** Professor Emeritus, The University of Georgia, Department of Educational Psychology, Georgia-USA, acohen@uga.edu, ORCID ID: 0000-0002-8776-9378

To cite this article:
Şen, S., & Cohen, A. S. (2024). A systematic review of factor mixture model applications. *Journal of Measurement and Evaluation in Education and Psychology, 15*(2), 79-93. https://doi.org/10.21031/epod.1423427

*Received: 21.01.2024*
*Accepted: 6.06.2024*

under the heading of mixture item response theory (mixture IRT; Mislevy & Verhelst, 1990; Rost, 1990), and latent class factor analysis (LCFA; Magidson & Vermunt, 2001).

The combination of both categorical and continuous latent variables enables the structure to be simultaneously categorical and dimensional, making the FMM useful for researchers (see Clark et al., 2013). This is because the FMM allows for the simultaneous modeling of latent class membership and the distribution both within and between latent classes. One of the main advantages of FMMs is their capability to handle multiple types of data within the same model, and to simultaneously model the relationships between the latent variables and the observed variables. This makes FMMs particularly useful for data where the relationships between the variables are complex and multifaceted.

The general FMM equation can be written as follows:

$$\mathbf{Y}_{ik} = \boldsymbol{\tau}_k + \boldsymbol{\Lambda}_k \boldsymbol{\eta}_{ik} + \boldsymbol{\varepsilon}_{ik}, \qquad (1)$$

where the *i* indicates persons and *k* indicates latent classes to which individuals are assigned. The $\mathbf{Y}_{ik}$ matrix in Equation (1) includes the response sets of the *i*-th individual in latent class *k*. Parameters $\boldsymbol{\tau}_k$, $\boldsymbol{\Lambda}_k$, $\boldsymbol{\eta}_{ik}$, and $\boldsymbol{\varepsilon}_{ik}$ represent the intercept vector, the factor loading vector, the vector of an individual's factor scores, and the residual, respectively. In addition, residuals are assumed to be normally distributed with a mean of zero and a variance of $\boldsymbol{\Theta}_k$.

The FMM assigns each individual to a latent class based on posterior probabilities (a.k.a. class probabilities). Once individuals have been classified, the FMM allows for individual intra-class differences by estimating a factorial model for each class (Clark et al., 2013). Class-specific item thresholds and slopes can be estimated in addition to factor variance(s), covariances, and mean(s). The parameters of FMMs can be estimated using frequentist (e.g., maximum likelihood) or Bayesian (e.g., Markov chain Monte Carlo) methods.

There are different model labels, however, according to the restrictions applied within FMM itself (Clark et al., 2013). Clark et al. identified four different types of FMMs, labeled as FMM-1, FMM-2, FMM-3, and FMM-4. Each of these has different parameter restrictions and measurement invariance assumptions. The most restricted FMM is FMM-1 and is equivalent to the LCFA. The least restrictive is FMM-4. With respect to FMM-1, the factor mean is the only parameter that changes across classes. The item thresholds and factor loadings are constrained to be equal across classes. In addition, the factor covariance matrix is fixed at zero in FMM-1 in order to assign the same factor scores to all individuals within a single latent class. In FMM-2, factor means, factor variances, and covariances are freely estimated. This model is also known as a mixture factor analysis (McDonald, 2003; Yung, 1997). FMM-1 and FMM-2 incorporate strict factorial invariance (Masyn et al., 2010). FMM-3 allows the factor covariance matrix and item thresholds to change across classes, but holds the factor loadings to equality, and fixes factor means to zero for identification purposes. Finally, in FMM-4, the factor means are fixed to zero and all other elements (intercepts, loadings, and factor covariances) can vary across classes.

Although there are examples of FMM used for confirmatory purposes, FMM is an exploratory model. That is, the model is used to estimate different solutions and the numbers of factors and latent classes are determined based on the model that best explains the data. Researchers typically analyze different models by changing the number of factors and increasing the number of classes one by one. The best fitting model is the one among all the candidates that is best fitting, both theoretically and statistically. Reporting all models that fit the data, comparing them, and outlining the decision-making process is an important part of the model selection process. Model selection can be challenging as the statistical results and the content-based theory do not always agree on the same number of latent classes.

There are several fit indices that can be used for model selection, including information criteria (IC) indices and likelihood ratio (LR) tests (McLachlan & Peel, 2000). The traditional LR test, which is appropriate for nested models, cannot be used for FMMs because regularity conditions are not met (see Nylund et al., 2007 and McLachlan & Peel, 2000). Thus, several adjusted versions of LR tests have been developed for use with FMMs. These are the bootstrapped LR test (BLRT; McLachlan & Peel, 2000), Lo-Mendell-Rubin LR test (LMR; Lo et al., 2001), adjusted LR test (aLMR; Lo et al., 2001), and the

Vuong-Lo-Mendell-Rubin LR test (VLMR; Lo et al., 2001). IC indices including Akaike's information criterion (AIC; Akaike, 1974), Bayesian information criterion (BIC; Schwarz, 1978), and extensions of these two indices, consistent AIC (CAIC), corrected AIC (AICc), and sample size adjusted BIC (SABIC), are also used for model selection. Nylund et al. (2007) describe simulation studies on how some of the fit indices perform in FMM selection. Based on the result in Nylund et al., BIC and BLRT are among the best choices for selecting the model with the correct number of latent classes.

An important advantage of FMMs is that covariates, such as gender, age, and education level, can also be added to the model in order to account for the uncertainty in class membership and to validate the FMM (Brown, 2013). Covariate inclusion can be done as either one-step or multiple-step (e.g., two- or three-step) approaches. In the first approach, covariates can be added directly to the FMM model. In the second approach, an unconditional FMM is analyzed first, then latent classes are estimated and the relationship between the latent class memberships and covariates is examined with a regression model. Both approaches have advantages and disadvantages (see Wang, et al., 2022 for a comparison). FMMs have been employed in several disciplines, including psychology, education, sociology, and the health sciences (e.g., Lin & Masse, 2021; Moors et al., 2014; Morin & Marsh, 2015).

Previous reviews on latent variable mixture models (e.g., Killian et al., 2019) have largely focused on other models, including mixture IRT (Sen & Cohen, 2019), latent profile analysis (LPA; Spurk et al., 2020), latent class analysis (LCA; Ulbricht et al., 2018) and growth mixture modeling (GMM; Baron et al., 2017). A review of the literature reveals that there are some review studies on FMMs. A few investigations have reported small-scale studies of the FMM. In this regard, Hofmans et al. (2020) presented an overview of four studies of careers, career counseling, and vocational behavior that used FMM analysis. Krawietz and Pett (2023) have conducted a systematic review of 95 studies including latent variable mixture modeling in communication scholarship. Kim et al. (2023) have recently conducted a systematic review of 76 FMM applications. However, this search was based only on studies in the PsycInfo database and did not select keywords that would include all possible factor mixture models such as LCFA. As a result, there is a lack of review studies that include all FMM applications. This present study fills this gap by systematically reviewing 334 applications of FMMs published in peer-reviewed journals across a variety of databases. This study (i) reviews existing applications of FMMs, (ii) to improve understanding of FMMs and how they are applied. The findings in this review can improve our understanding of how FMMs are applied. Inconsistent and incomplete reporting practices can set a bad example for any new study that plans to use FMMs. Thus, it is necessary to identify and summarize best practices to ensure the quality of FMM applications. Knowledge gained from this review will provide useful information for future researchers who may use FMMs in their studies.

## Method

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; Moher, Liberati, Tetzlaff, Altman, & The PRISMA Group, 2009) standards were followed for conducting this study. Methods used in the review are presented below.

### Inclusion/Exclusion Criteria

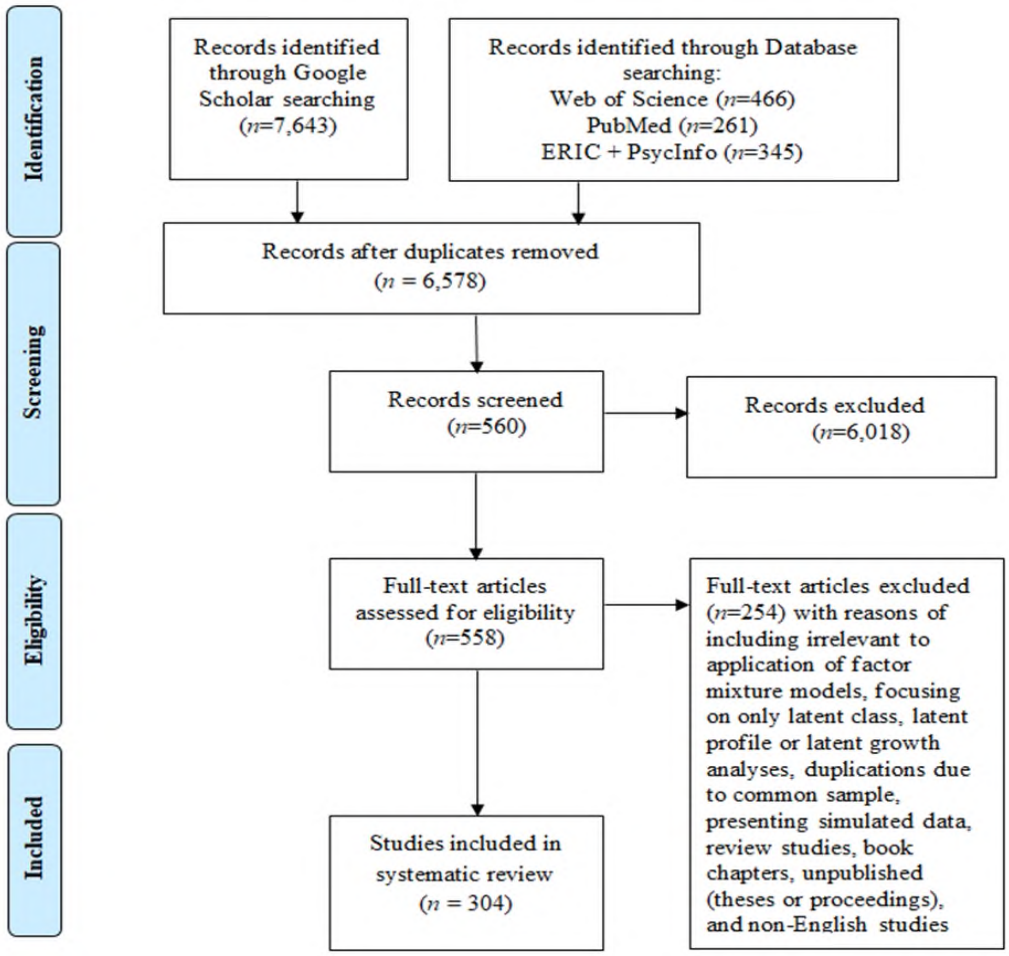
The studies included in this systematic review were screened to include only those published in peer-reviewed journals in English before 2023. Studies needed to apply at least one FMM to be included in this review. Studies which included only simulations, used simulated data, contained use of only latent class, latent profile, or latent growth analysis, unpublished studies (such as theses or proceedings), book chapters, review studies, and studies not written in English also were excluded. However, the application part of the methodology studies that include application is also included in the review. Mixture IRT based studies were not included in this review as this review was written within the framework of factor analysis, and a systematic review has already been reported on mixture IRT studies (see Sen & Cohen, 2019).

## Search Strategy

The review strategy used in this study is reported below. The search was conducted on January 1, 2023 using the Google Scholar search engine and included several databases including the Web of Science, PubMed, ERIC, and PscyInfo. Different labels were used for factor mixture model, including factor mixture analysis, mixture factor model, mixture factor analysis, latent class factor model, and latent class factor analysis. To cover all of these terms in the search process, the following search strings were used: "factor mixture*" OR "mixture factor*" OR "latent class factor*". The first search yielded 7,643 studies on Google Scholar, 466 studies on the Web of Science, 266 studies on PubMed, and 345 studies on the ERIC and PsycInfo databases. Duplicate articles ($n$=1,065) were deleted. From the remaining 6,578 unique studies, all articles that did not use any form of factor mixture models ($n$=6,018) were excluded. Among the remaining 558 full studies, 254 were excluded as being irrelevant to the application of factor mixture models, focusing on only latent class, latent profile, latent growth analyses, duplications due to a common sample, presenting only simulated data, review studies, book chapters, unpublished (theses or proceedings), or studies written in a language other than English. The final sample consisted of 304 peer-reviewed articles. Some studies included more than one factor mixture model analysis applied to different samples. This resulted in 334 applications from the 304 studies. A PRISMA (Moher et al., 2009) diagram showing each step of this search process is presented in Figure 1. The references of the studies included in the review can be requested from the first author.

**Figure 1**

*PRISMA flow diagram of study selection*



### Data coding and analysis

Each study included in the review was coded for study and model characteristics using the following coding scheme: (a) characteristics of the study (author(s), year of publication, and journal title); (b) country and region of application; (c) construct measured; (d) use of FMM; (e) FMM type; (f) other model types applied before FMM analysis; (g) sample size and number of items; (h) population type; (i) model fit statistics; (j) number of classes checked and decided; (k) software package; (l) estimation type; (m) covariate used in FMMs; (n) missing data handling method. While creating this coding scheme, some previous review studies (Killian et al., 2019; Sen & Cohen, 2019; Spurk et al., 2020; Ulbricht et al., 2018) were taken as references. Some of these variables are coded continuously, and some are coded categorically. The percentage and frequencies of the categorical variables are reported, and also the arithmetic mean and standard deviations of the continuous variables.

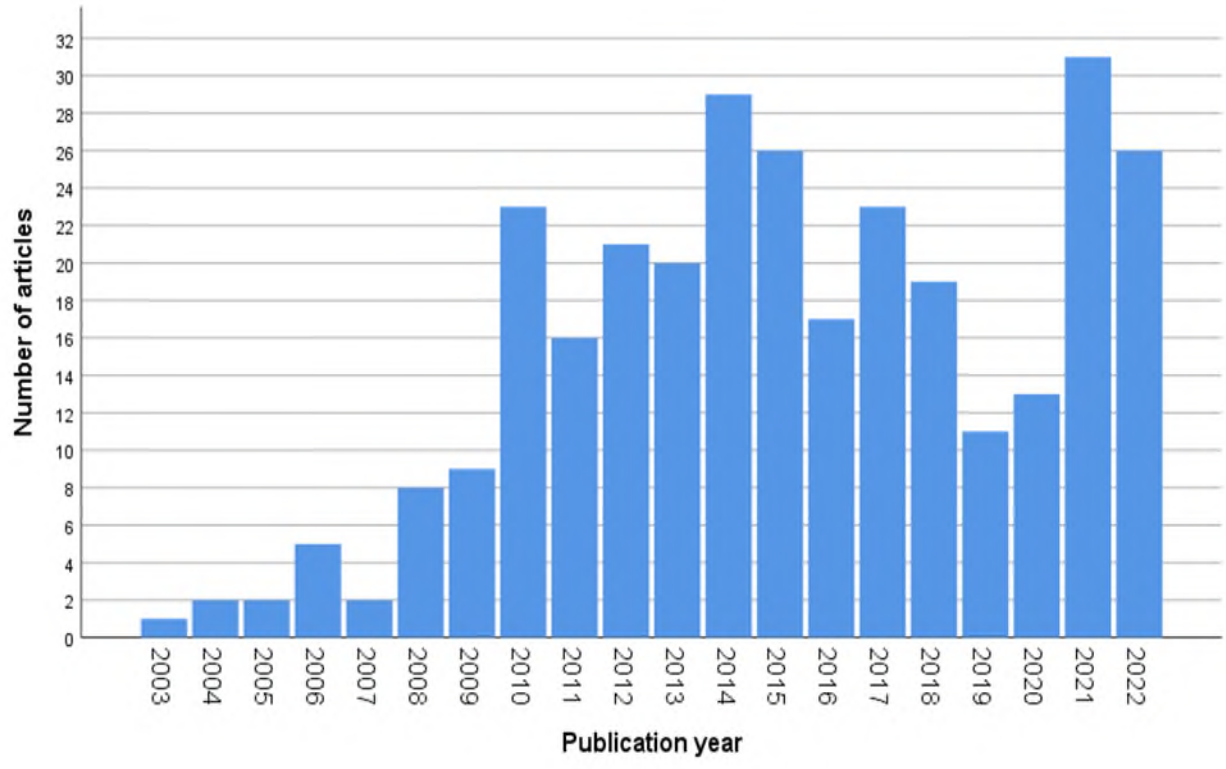
## Results

### Characteristics of the study

In this section we review information about the publication year of the 304 studies, the country/region in which they were published, and the journals in which they were published. The papers in our review were published in 194 different journals between 2003 and 2022 (see Figure 2). The distribution of published studies by year is shown in Figure 2. The year in which the fewest studies were published was 2003 ($n$=1), and the year in which the most studies were published was 2021 ($n$=31). The average number of studies per year was 15.2. The fact that the number of studies is higher in the last 10 years than in the previous decade is likely evidence that studies using FMM are on the rise.

**Figure 2**

*FMM applications published between 2003 and 2022*



Among journals with the most FMM studies published are *Structural Equation Modeling: A Multidisciplinary Journal* (*n*=11), *Psychological Assessment* (*n*=8), *Frontiers in Psychology* (*n*=7), *Psychological Medicine* (*n*=6), *Journal of Personality Assessment* (*n*=5), *Journal of Abnormal Psychology* (*n*=5), *Journal of Anxiety Disorders* (*n*=5), *Psychiatry Research* (*n*=5), *Plos One* (*n*=5), and *Social Indicators Research* (*n*=4). The studies considered in this review were conducted in 36 different countries. The country in which most FMM studies were conducted was the U.S. (*n*=145) followed by Australia (*n*=20) and the Netherlands (*n*=19). Fourteen FMM studies were conducted in Germany, 13 FMM studies were conducted in Canada, and 12 FMM studies were conducted in Finland and Italy. Seventeen studies were conducted with multinational samples. Six percent of the studies did not mention the region of the study. Overall, 48.8% FMM applications were conducted in North America and 32% in Europe. The percentages of studies conducted on other continents are as follows: Asia (7.2%), Oceania (6.3%), South America (1.8%), and Africa (0.6%).

**Table 1**

*Overview of study and FMM characteristics of the reviewed studies (k=304 with 334 applications)*

| **Study Characteristics** | ***N* (%), or *M* (SD), Median** |
|---|---|
| Average number of studies per year | *M*=15.2 |
| Region | |
| North America | 163 (48.8%) |
| Europe | 107 (32.0%) |
| Asia | 24 (7.2%) |
| Oceania | 21 (6.3%) |
| South America | 6 (1.8%) |
| Not reported or unclear | 6 (1.8%) |
| Africa | 2 (0.6%) |
| Multi-continent | 5 (1.5%) |
| Sample description | |
| Students | 63 (18.9%) |
| Adults | 53 (15.9%) |
| Patients | 38 (11.4%) |
| General population | 33 (9.9%) |
| Children | 24 (7.2%) |
| Adolescents | 22 (6.6%) |
| Employees | 19 (5.7%) |
| Twins | 13 (3.9%) |
| Young adults & Teens | 8 (2.4%) |
| Others | 72 (21.6%) |
| Sample size, median | *M*=3562.42, *SD*=7662.53, Med=888 |
| Number of items/indicators | *M*=17.34, *SD*=17.71, Med=12 |
| Item type | |
| Likert | 209 (62.6%) |
| Dichotomous/binary | 78 (23.4%) |
| Continuous | 24 (7.2%) |
| Mixed | 11 (3.3%) |
| Not reported/unclear | 10 (3.0%) |
| 0-10 Rating scale | 2 (0.6%) |
| Modeling approach | |
| Exploratory | 275 (82.3%) |
| Confirmatory | 59 (17.7%) |
| FMM Type | |
| FMM-1/LCFA | 109 (32.6%) |
| FMM-2 | 46 (13.8%) |
| FMM-3 | 33 (9.9%) |
| FMM-4 | 19 (5.7%) |
| ML-FMM | 10 (3.0%) |
| Not clear | 155 (46.4%) |

**Table 1 Continued**

| Study Characteristics | *N* (%), or *M* (SD), Median |
|---|---|
| Models estimated before FMM analyses | |
| CFA | 141 (42.2%) |
| LCA | 109 (32.6%) |
| EFA | 79 (23.7%) |
| LPA | 23 (6.9%) |
| ESEM | 15 (4.5%) |
| SEM | 13 (3.9%) |
| IRT | 13 (3.9%) |
| LTA | 11 (3.3%) |
| Others | 14 (4.2%) |
| Not reported | 87 (26.0%) |
| Software package | |
| Mplus | 259 (77.5%) |
| Latent GOLD | 37 (11.1%) |
| R packages | 12 (3.6%) |
| Others (mdltm, Mx, OpenMx, WINBUGS) | 8 (2.4%) |
| Not reported | 19 (5.7%) |
| Type of estimator | |
| Frequentist | 245 (73.4%) |
| Bayesian | 1 (0.3%) |
| Not reported/unclear | 88 (26.3%) |
| Missing data methods | |
| FIML | 68 (20.4%) |
| Pairwise/Listwise/Excluded | 58 (17.4%) |
| Imputation (single, multiple, nonparametric, mean, recorded as zero) | 12 (3.6%) |
| Complete data | 16 (4.8%) |
| Not reported | 180 (53.9%) |
| Number of classes in final models | *M*=2.96, *SD*=1.28, Med=3 |
| Number of factors in final models | *M*=2.17, *SD*=1.48, Med=2 |
| Class percentages are reported in final models | 285 (85.3%) |
| Classes are labeled in final models | 208 (62.3%) |
| Profile plots are presented | 129 (38.6%) |
| Fit indices reported for model selection | 322 (96.4%) |
| Multiple fit values applied for model selection | 286 (85.6%) |
| Interpretability and theory considered for model selection | 116 (34.7%) |
| AIC/BIC difference used for model selection | 35 (10.5%) |
| Elbow plots used for model selection | 9 (2.7%) |
| Class sizes considered for model selection | 42 (12.6%) |
| Applied model fit values | |
| BIC | 296 (88.6%) |
| AIC | 180 (53.9%) |
| Entropy | 165 (49.4%) |
| SABIC | 161 (48.2%) |
| LMR-LRT | 99 (29.6%) |
| BLRT | 81 (24.3%) |
| LRT/aLRT | 52 (15.6%) |
| VLMR | 44 (13.2%) |
| CAIC | 20 (6.0%) |
| Others | 61 (18.3%) |
| Most frequent covariates included | |
| Psychological and behavioral disorders | 117 (35.0%) |
| Gender | 104 (31.1%) |
| Age | 92 (27.5%) |
| Education level | 39 (11.7%) |
| Ethnicity | 20 (6.0%) |
| Marital status | 18 (5.4%) |
| Income | 12 (3.6%) |
| SES | 11 (3.3%) |
| BMI | 10 (3.0%) |
| Employment status | 7 (2.1%) |
| Language | 4 (1.2%) |

**Table 1 Continued**

| Study Characteristics | *N* (%), or *M* (SD), Median |
|---|---|
| Statistical analyses after FMM | |
| Chi-square test | 48 (14.4%) |
| Regression | 39 (11.7%) |
| Logistic regression | 34 (10.2%) |
| t-test or non-parametric versions | 26 (7.8%) |
| ANOVA | 26 (7.8%) |
| Correlation | 13 (3.9%) |
| Multiple comparison/Mean comparison | 13 (3.9%) |
| MANOVA | 11 (3.3%) |
| R3STEP | 9 (2.7%) |
| ANCOVA | 7 (2.1%) |
| Odds ratio | 6 (1.8%) |
| ROC Curves | 5 (1.5%) |
| Wald test | 4 (1.2%) |
| Cross-tabs | 4 (1.2%) |
| Kappa classification agreement | 4 (1.2%) |
| SEM/ESEM | 4 (1.2%) |
| Fisher's exact test | 3 (0.9%) |
| Others | 33 (9.9%) |
| Not reported | 134 (40.1%) |
| Use of FMM | |
| Identifying the latent classes/clusters/profiles/patterns | 157 (47%) |
| Investigating the latent structure | 89 (27%) |
| Model comparison | 17 (5%) |
| Analyzing population heterogeneity | 14 (4%) |
| Determining the best fitting model | 14 (4%) |
| Exploring the response process/styles | 11 (3%) |
| Examining the validity | 6 (2%) |
| Testing measurement invariance/equivalence | 6 (2%) |
| Others | 20 (6%) |

## Use of FMM

Based on preliminary analysis, FMM applications in Table 1 were divided into nine subcategories. These included identifying the latent classes/clusters/profiles/patterns (47%), investigating the latent structure (27%), analyzing population heterogeneity (4%), model comparison (5%), determining the best fitting model (4%), exploring the response process or response styles (3%), examining the validity (2%), and testing measurement invariance or equivalence (2%). The remaining six percent of the studies included applications examining measurement assumptions, evaluating the appropriateness of latent class factor analysis, model building, investigating the covariate effect, investigating Spearman's law of diminishing returns, and testing for performance and structural differences.

## Outcome Measured

Different topics covered in these studies included alcohol use disorder (*n*=14), posttraumatic stress disorder (*n*=10), anxiety sensitivity (*n*=7), panic attack symptoms (*n*=7), schizotypal personality disorders (*n*=5), tobacco dependence (*n*=5), autism spectrum disorder (*n*=5), borderline personality disorder (*n*=4), life satisfaction (*n*=4), job stress and job resources (*n*=4), mathematics (*n*=4), and reading (*n*=4).

## Number of items and item type

There were five different item types in the studies reviewed. As can be seen in Table 1, the most frequent item type was the Likert item (*n*=209, 62.6%), followed by the dichotomous (*n*=81, 23.4%), the continuous (*n*=24, 7.2%), mixed (*n*=11, 3.3%), and rating scale items (*n*=2, 0.6%). Item type was not specified in 3% (*n*=10) of the studies. The number of items (or indicators) used varied greatly from *k*=1

(Ma et al., 2022) to *k*=165 (Grove et al., 2015). Only four studies used measurement tools with more than 100 items. The average number of items across 334 applications was 17.34, with a median of 12, and an SD of 17.71.

### Sample size and population type

The studies reviewed included samples of participants from a variety of populations. These populations were grouped into 10 categories. Frequencies and percentages for each category are presented in Table 1. Populations in the reviewed studies were identified as students (18.9%), adults (15.9%), patients (11.4%), general population (9.9%), children (7.2%), adolescents (6.6%), employees (5.7%), twins (3.9%), and young adults and teens (2.4%). Almost twenty-two percent of the studies included different types of participants including veterans, soldiers, dancers, gamers, athletic performers, educators, households, immigrants, job applicants, current smokers or drinkers, and individuals with autism spectrum disorder.

Sample size is an important consideration in applications of FMM. Of the studies reviewed, sample sizes varied greatly from *N*=50 to *N*=261,747. Apart from four studies with very large sample sizes (261747, 212674, 177480, and 116543), the remaining studies had sample sizes of less than 50,000. Only two studies used samples of fewer than 100 individuals, and fifteen studies had sample sizes between 100 and 200. The mean sample size across 330 applications was 3,562, with a median of 888, and an SD=7,662.53, excluding the four outlier studies with the very large sample sizes.

### Missing data

Researchers often have to deal with missing data. When this occurs, there are a number of different methods that can sometimes be used to deal with missing data. These include data deletion (pairwise or listwise), imputation (single or multiple), or FIML (full information maximum likelihood) methods (see Enders, 2022). In handling missing data, most studies (*n*=68, 20.4%) preferred the FIML estimation method. Missing data were excluded or deleted pairwise or listwise in 58 studies (17.4%). Imputation was used in 12 studies (3.6%), including single imputation, multiple imputation, nonparametric imputation, mean imputation, and recording missingness as zero. Only 16 of the studies (4.8%) reviewed reported using a complete data set. The remaining 180 studies with missing data (53.9%) did not report how missing data was addressed.

### Analyses applied before FMM application

Clark et al. (2013, p. 691) recommend the following for the initial step (Step 0) of the FMM analysis: "*Fit latent class analysis and factor analysis models for later comparison and to determine the ending point combination of number of class and factors when fitting factor mixture models*". Additional analyses prior to the FMM included confirmatory factor analysis (CFA; *n*=141, 42.2%) and LCA (*n*=109, 32.6%), which were used in most often, followed by EFA (*n*=79, 23.7%), LPA (*n*=23, 6.9%), exploratory structural equational model (ESEM; *n*=15, 4.5%), SEM (*n*=13, 3.9%), IRT (*n*=13, 3.9%), LTA (*n*=11, 3.3%), and others (*n*=14, 4.2%). In 87 studies (26.0%), the type of analysis conducted before FMM was not reported. More than one preliminary analysis was performed in 144 (43.1%) of the studies. EFA and LCA were used together in 50 (15.0%) studies, CFA and LCA were used together in 39 (11.7%) studies, and EFA-CFA was used together in 15 (4.5%) studies.

### Modeling strategy

An exploratory approach to determine the number of latent classes was used in 275 of the 334 studies reviewed (82.3%). For the remaining six studies, a single latent class solution (i.e., 2 or 3 latent classes) was used (1.78%). A model with a single latent class solution was the final model in 53 studies (4.5%). In these latter 53 studies, no information was provided as to the number of different latent class solutions

tried. A number of different models were reported in the studies reviewed. Different labels were used for the FMM analysis: 67.9% used the label FMM; 1.4% used the label exploratory FMM. Of those, 32.6% used FMM-1 or latent class factor analysis; 13.8% used FMM-2; 9.9% used FMM-3; and 5.7% used FMM-4. The specific type of FMM used in 46.4% of the studies could not be understood from the information presented. Multilevel extensions of FMM were used in 3.0% of the studies. In the remaining studies, different FMM labels were used including non-normal FMM, twin FMM, multimodality FMM, discrete FMM, constrained FMM, confirmatory FMM, MTMM mixture modeling, repeated measures LCA, and mixtures of factor analyzer.

## Estimation methods and software

FMM analyses can be conducted with several statistical software packages, including Mplus (L. K. Muthén & Muthén, 2017), Latent GOLD (Vermunt & Magidson, 2003), mdltm (von Davier, 2006), Mx (Neale et al., 2006), WINBUGS (Spiegelhalter et al., 2003), and R packages such as FactMixtAnalysis (Viroli, 2011) and mclust (Scrucca et al., 2016). Of the papers included in this review, Mplus (*n*=259, 77.5%) was the most commonly reported statistical software package for FMM applications followed by Latent GOLD (*n*=37, 11.1%) and two R packages, FactMixtAnalysis and mclust (*n*=12, 3.6%). Eight (2.4%) studies reported other statistical software packages, including mdltm, Mx, OpenMx, and WINBUGS. There were studies (*n*=19, 5.7%) that did not report the name of the software used.

In the present review, we distinguish between two types of parameter estimation methods for FMMs: frequentist and Bayesian estimation. Most of the included studies (*n*=245; 73.4%) were conducted with frequentist estimation methods. Only one study (Cho et al., 2014) reported using Bayesian estimation. Eighty-eight studies (26.3%) did not report the estimation method used. For the frequentist estimation methods, we differentiate between maximum likelihood (ML), robust maximum likelihood (MLR), full maximum likelihood (FIML), marginal maximum likelihood (MML), linear approximation of ML, and weighted least squares (WLS)-based estimations (WLS, WLSM, and WLSMV). In this review, MLR (*n*=152, 45.5%) was the most commonly reported estimation method for FMM applications, followed by ML (*n*=72, 21.6%), FIML (*n*=9, 2.7%), and MML (*n*=4, 1.2%). Linear approximation ML was used in two studies. Apart from ML based methods, six studies used WLS (*n*=2, 0.6%), WLSM (*n*=1, 0.3%), and WLSMV (*n*=3, 0.9%) estimation methods.

Random starting values are another important issue to consider in ML estimation of FMM parameters due to the local maxima problem. In the present review, random starting values were explicitly mentioned in 87 studies (26.0%). Of these, software defaults were used in one study, and 20 of the reviewed studies reported that "random starting values were used" without reporting the number of random starting values. The number of random starting values used varied greatly in the reviewed studies, ranging from 2 to 200,000. The most commonly used random starting values were 500 (*n*=14), 100 (*n*=11), 5000 (*n*=8), 1000 (*n*=8), and 2000 (*n*=6).

## Model fit statistics

In the present review, at least one fit index was reported for model selection in most of the studies (*n*=322, 96.4%). All but 48 studies (*n*=286, 85.6%) reported multiple fit indices with a majority reporting between two and five indices (74.2%). For the studies reporting fit indices, BIC (*n*=296, 88.6%) was the most commonly reported fit index for FMM applications, followed by AIC (*n*=180, 53.9%), entropy (*n*=165, 49.4%), and SABIC (*n*=161, 48.2%). These four indices were followed by likelihood ratio-based tests, including LMR-LRT (*n*=99, 29.6%), BLRT (*n*=81, 24.3%), LRT/aLRT (*n*=52, 15.6%), and VLMR (*n*=44, 13.2%). CAIC was reported in 20 studies (6.0%). Apart from these indices, other indices were also reported in 61 studies (18.2%). These indices include values such as $L^2$ value (*n*=10), classification error (*n*=9), bivariate residuals (*n*=8), chi-square diff test/the Pearson chi-square (*n*=8), ICL-BIC (*n*=5), AICC (*n*=3), AIC3 (*n*=3), Cressie-Read (*n*=2), delta AIC/BIC (*n*=2), Akaike weight (*n*=2), bootstrap *p* (*n*=2), DIC (*n*=1), ACPP (*n*=1), IC1000 (*n*=1), the log penalty AIC (*n*=1), AWE (*n*=1), and ratio of distance measure (*n*=1). In the present review, only 169 studies (50.6%) reported BIC

and one of the likelihood ratio tests. Selecting best among several models should take into account theoretical factors in addition to fit indices (Muthén, 2006). Methods other than fit indices were also used in the studies reviewed. Studies also considered interpretability and theory (*n*=116, 34.7%), class sizes (*n*=42, 12.6%), AIC or BIC difference (*n*=35, 10.5%), and elbow plot (*n*=9, 2.7%) for model selection. Loglikelihood (LL) values and degrees of freedom (df) are other statistics to be reported with fit indices. LL was reported in 183 studies (54.8%) and df value was reported in 165 studies (49.4%). Only 44.9% of the studies (*n*=150) reported both statistics together. However, neither LL nor df were reported in 137 studies (40.7%).

**Numbers of factors and classes**

As suggested above, the best FMM model should be decided upon on the basis of model fit indices and theory. The number of factors and the number of latent classes also need to be reported for the final model. In the present review, the number of factors varied between 1 and 11, while the number of latent classes varied between 0 and 7. The mean number of classes across the 334 applications was 2.96, with a median of 3, and an SD=1.28. A majority of the studies reported a two-class (*n*=130), three-class (*n*=87) or four-class (*n*=62) FMM solution. The mean number of factors was 2.17, with a median of 2, and an SD=1.48. A majority of the studies reported one- (*n*=138), two- (*n*=104) or three-factor (*n*=48) models. Labeling for multiple latent classes in the final model, reporting the percentage or ratio of each latent class, and drawing a profile plot of the latent classes on the items are among the common practices in FMM analyses. In the present review, class percentages or proportions were reported in 85.3% (*n*=285), latent classes were labeled in 62.3% of the studies (*n*=208), and profile plots were drawn in 129 studies (38.6%).

**Covariates and further analyses**

Another important issue is the use of covariates in FMM analyses. Because latent classes are unobserved, variables (such as gender and race) are frequently linked to the latent class variable in order to better understand and characterize latent classes (Wang et al., 2022). A significant covariate effect would specifically mean that this covariate could explain the latent class membership. There are two options to examine the covariate effect in FMM analyses: adding the covariate variable directly to the model (sometimes referred to as a one-step approach) or performing a regression analysis with the latent classes obtained from the model (sometimes referred to as a three-step approach). In the three-step analysis, researchers first estimate an unconditional FMM without adding a covariate, then assign each respondent to one of the latent classes in the final model, after which a multinomial regression analysis is applied with the class membership and covariates (e.g., gender) specified by the researcher. In the present review, covariates in the estimation of FMMs were included in 199 studies (59.6% of the 334 applications). Only 28 studies added covariates directly to the FMM. The remaining 171 studies followed a two- or three-step approach. The most frequent covariates included were psychological and behavioral disorders (*n*=117, 35.0%), gender (*n*=104, 31.1%), age (*n*=92, 27.5%), and education level (*n*=39, 11.7%). Additional covariates included ethnicity (*n*=20, 6.0%), marital status (*n*=18, 5.4%), income (*n*=12, 3.6%), SES (*n*=11, 3.3%), body mass index (*n*=10, 3.0%), employment status (*n*=7, 2.1%), and language (*n*=4, 1.2%). Most studies included more than one covariate.

Once the best-fitting model was determined, classes were compared across different covariates. Applications in this review suggest that researchers were interested in analyzing the categorical latent variable (i.e., latent class) with further statistical analyses in order to investigate the relationship between class membership and auxiliary observed variables. A number of different analyses are used in FMM studies for covariate effect. The chi-square test (*n*=48, 14.4%), linear regression (*n*=39, 11.7%) and logistic regression (*n*=34, 10.2%) were the most commonly used analyses. Additional analyses included *t*-test and its non-parametric version (*n*=26, 7.8%), ANOVA (*n*=26, 7.8%), correlation (*n*=13, 3.9%), multiple comparison and comparison of means (*n*=13, 3.9%), MANOVA (*n*=11, 3.3%), R3STEP (*n*=9, 2.7%), ANCOVA (*n*=7, 2.1%), odds ratio (*n*=6, 1.8%), ROC curves (*n*=5, 1.5%), Wald test (*n*=4, 1.2%), cross-tabs (*n*=4, 1.2%), kappa classification agreement (*n*=4, 1.2%), SEM/ESEM (*n*=4, 1.2%), MIMIC

model (*n*=3, 0.9%), and Fisher's exact test (*n*=3, 0.9%). Other statistical methods included cluster analysis (*n*=2), taxometric (*n*=2), IFA (*n*=2), PLS (*n*=1), bifactor analysis (*n*=1), MAXCOV (*n*=1), latent factor (*n*=1), and confirmatory MIRT (*n*=1).

## Conclusion

In this study, a systematic review was conducted to first summarize the state of the use of FMMs as found in peer-reviewed journals and then to describe the trends in use based on these studies. There were a total of 304 peer-reviewed articles with 334 applications retrieved from databases including Web of Science, PubMed, ERIC, and PscyInfo. Relatively few studies using FMMs were published in fields other than psychology. In future studies, more emphasis may need to be placed on FMM analyses, particularly in the areas of health and education. Further, most studies were conducted in North American countries, including the U.S. (*n*=130), Canada (*n*=13), and European countries such as the Netherlands (*n*=19), Germany (*n*=14), Italy (*n*=12), and Finland (*n*=12). It is important that researchers in other countries take advantage of the FMM and seize the chance to answer new research questions.

Studies varied in their use of FMMs, outcome measured, methods for handling missing data, and reporting of methods and results. An interesting finding is that most studies used Mplus software and the MLR estimation method. Latent Gold was also used, particularly in LCFA studies. Not surprisingly, a majority of the reviewed studies used an exploratory approach, as FMMs are mainly exploratory in nature. As is the case with CFA, however, FMMs can also be applied in a confirmatory fashion. When multiple populations are believed to underlie the data, researchers may want to use a confirmatory version of the FMMs (Gagné, 2004) where restrictions are added in advance. The theoretically more grounded confirmatory approach may enable researchers to obtain more accurate findings in future studies. As Clark et al. (2013) has suggested, most of the FMM studies started with either EFA or CFA and LCA. In this review, however, more than a quarter of the studies did not apply these methods. Clark et al. (2013, p. 690) notes that "…the FMM-1 and FMM-2 often do not fit real data well because the specification of invariant factor loadings and thresholds are likely to be too restrictive for certain items." In this review, however, most of the studies relied on simple FMMs such as FMM-1 or LCFA, and FMM-2 methodology, and other models were not reported that often. One barrier to other models would be access to example code/syntax. As was noted in this review, FMMs can be applied with varying numbers of items and sample sizes. Typical applications include sample sizes of around 900 and approximately 12 items. The number of studies applying FMM analysis with small samples is not small. Studies with small samples can decide on the adequacy of the sample according to the power analysis that can be performed with Monte Carlo simulation analysis.

In the present review, 169 studies (50.3%) reported BIC and one of the likelihood ratio tests for model fit followed this suggestion. According to Muthén (2006), selecting among several models should take into account both theoretical and statistical factors. In the present review, 116 studies followed this suggestion (34.5%). Based on evidence from the studies in this review, it is recommended reporting multiple fit indices and also taking the theory into account, when selecting the final model. AIC, BIC, and SABIC were the most frequently reported IC indices. LMR and BLRT were the most reported LR based tests. This review also found that other considerations, including class proportions and entropy, were also considered for model selection. Although entropy is not a model selection index, it was the 3rd most reported index in the studies reviewed. Although previous studies (e.g., Lubke & Muthén, 2007) have indicated that entropy should not be used to determine the number of classes, there were a number of studies in this review that did use entropy in model selection. The least used indices for model selection were the AIC or BIC difference and the elbow plot methods. While most of the studies reported the percentage of latent classes in the final model, only 62% labeled the latent classes. Only 39% of the studies presented a profile plot over the final latent classes. This shows that after deciding on the optimal number of classes, researchers ignore some of the information they should give to inform the reader.

A majority of the studies reviewed included covariates in the estimation of FMMs. This approach can be useful for accounting for the uncertainty in class membership and for helping interpret FMM results. Findings of this review also showed that demographic variables such as gender, age, education level,

and marital status, were used often in addition to psychological and behavioral disorders. A possible limitation of this review is that the focus was solely on peer-reviewed research published in English language journals. This ignores the gray literature, which includes unpublished studies, especially theses.

It is clear from this review that not all studies using FMMs have provided the same level and standard of study details. In the literature, some researchers (Clark et al., 2013; Lubke, 2019) have made some suggestions on how mixture model and FMM studies should be reported. It is expected that both the suggestions of these researchers and the results found in this review study will improve the reporting quality of future FMM studies. We hope that this review will contribute to future FMM research.

## Declarations

**Author Contribution:** S.Ş. conceived of the presented idea. S.Ş. reviewed the studies and analyzed the data. A.S.C. verified the analytical methods. Both authors contributed to the final version of the manuscript.

**Conflict of Interest:** The authors have no conflicts of interest to declare.

**Ethical Approval:** Ethical approval was not required because this study retrieved and synthesized data from already published studies.

## References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723. https://doi.org/10.1109/TAC.1974.1100705

Baron, E., Bass, J., Murray, S. M., Schneider, M., & Lund, C. (2017). A systematic review of growth curve mixture modelling literature investigating trajectories of perinatal depressive symptoms and associated risk factors. *Journal of Affective Disorders, 223*, 194–208. https://doi.org/10.1016/j.jad.2017.07.046

Berlin, K. S., Williams, N. A., & Parra, G. R. (2014). An introduction to latent variable mixture modeling (part 1): Overview and cross-sectional latent class and latent profile analyses. *Journal of Pediatric Psychology, 39*(2), 174–187. https://doi.org/10.1093/jpepsy/jst084

Brown, T. A. (2013). Latent variable measurement models. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods* (Vol. 2, pp. 257–280). Oxford University Press.

Cho, S. J., Cohen, A. S., & Kim, S. H. (2014). A mixture group bifactor model for binary responses. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(3), 375–395. https://doi.org/10.1080/10705511.2014.915371

Clark, S. L., Muthén, B. O., Kaprio, J., D'Onofrio, B. M., Viken, R., & Rose, R. J. (2013). Models and strategies for factor mixture analysis: An example concerning the structural underlying psychological disorders. *Structural Equation Modeling, 20*, 681–703. https://doi.org/10.1080%2F10705511.2013.824786

Enders, C. K. (2022). *Applied missing data analysis*. Guilford Publications.

Gagné, P. E. (2004). *Generalized confirmatory factor mixture models: A tool for assessing factorial invariance across unspecified populations* [Unpublished doctoral dissertation]. University of Maryland, College Park.

Grove, R., Baillie, A., Allison, C., Baron-Cohen, S., & Hoekstra, R. A. (2015). Exploring the quantitative nature of empathy, systemising and autistic traits using factor mixture modelling. *The British Journal of Psychiatry, 207*(5), 400–406. https://doi.org/10.1192/bjp.bp.114.155101

Hofmans, J., Wille, B., & Schreurs, B. (2020). Person-centered methods in vocational research. *Journal of Vocational Behavior, 118*, 103398. https://doi.org/10.1016/j.jvb.2020.103398

Killian, M. O., Cimino, A. N., Weller, B. E., & Hyun Seo, C. (2019). A systematic review of latent variable mixture modeling research in social work journals. *Journal of Evidence-Based Social Work, 16*(2), 192–210. https://doi.org/10.1080/23761407.2019.1577783

Kim, E., Wang, Y., & Hsu, H. Y. (2023). A systematic review of and reflection on the applications of factor mixture modeling. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000630

Krawietz, C. E., & Pett, R. C. (2023). A systematic literature review of latent variable mixture modeling in communication scholarship. *Communication Methods and Measures, 17*(2), 83–110. https://doi.org/10.1080/19312458.2023.2179612

Lazarsfeld, P., & Henry, N. (1968). *Latent structure analysis*. Houghton Mifflin.

Lin, Y., & Mâsse, L. C. (2021). A look at engagement profiles and behavior change: A profile analysis examining engagement with the Aim2Be lifestyle behavior modification app for teens and their families. *Preventive Medicine Reports, 24*, 101565. https://doi.org/10.1016/j.pmedr.2021.101565

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*, 767–778. https://www.jstor.org/stable/2673445

Lubke, G. (2019). Latent variable mixture models. In G. R., Hancock, L. M., Stapleton, & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 202–213). Routledge.

Lubke, G., & Muthén, B. O. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling, 14*, 26–47. https://doi.org/10.1080/10705510709336735

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*, 21–39. https://doi.org/10.1037/1082-989X.10.1.21

Ma, X., Wang, M., Ma, J., Zhang, Z., Hao, Y., & Yan, N. (2022). The association between lifestyles and health conditions and the choice of traditional Chinese medical treatment in China: A latent class analysis. *Medicine, 101*(51), e32422. https://doi.org/10.1097/md.0000000000032422

Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots, and related graphical displays. *Sociological Methodology, 31*(1), 223–264. http://dx.doi.org/10.1111/0081-1750.00096

Masyn, K. E., Henderson, C. E., & Greenbaum, P. E. (2010). Exploring the latent structures of psychological constructs in social development using the dimensional–categorical spectrum. *Social Development, 19*(3), 470–493. https://doi.org/10.1111/j.1467-9507.2009.00573.x

McDonald, R. P. (2003). A review of multivariate taxometric procedures: Distinguishing types from continua. *Journal of Educational and Behavioral Statistics, 28*, 77–81. http://dx.doi.org/10.3102/10769986028001077

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. Wiley.

Mislevy, R.J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*(2), 195–215. https://psycnet.apa.org/doi/10.1007/BF02295283

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*, e1000097. https://doi.org/10.1136/bmj.b2535

Moors, G., Kieruj, N. D., & Vermunt, J. K. (2014). The effect of labeling and numbering of response scales on the likelihood of response bias. *Sociological Methodology, 44*(1), 369–399. https://doi.org/10.1177/0081175013516114

Morin, A. J., & Marsh, H. W. (2015). Disentangling shape from level effects in person-centered analyses: An illustration based on university teachers' multidimensional profiles of effectiveness. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(1), 39–59. https://doi.org/10.1080/10705511.2014.919825

Muthén, B. (2006). Should substance use disorders be considered as categorical or dimensional?. *Addiction, 101*, 6–16. https://doi.org/10.1111/j.1360-0443.2006.01583.x

Muthén, L. K., & Muthén, B. O. (1998/2017). *Mplus user's guide* (Eight ed.). Muthén & Muthén.

Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55*(2), 463–469. https://doi.org/10.1111/j.0006-341x.1999.00463.x

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H., (2006). *Mx: Statistical Modeling*, 7th ed. Medical College of Virginia, Richmond.

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*, 535–569. https://doi.org/10.1080/10705510701575396

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271–282. https://doi.org/10.1177/014662169001400305

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. https://www.jstor.org/stable/2958889

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R journal, 8*(1), 289–317.

Sen, S., & Cohen, A. S. (2019). Applications of mixture IRT models: A literature review. *Measurement: Interdisciplinary Research and Perspectives, 17*(4), 177–191. https://doi.org/10.1080/15366367.2019.1583506

Spearman, C. (1904). 'General intelligence' objectively determined and measured. *American Journal of Psychology, 5*, 201–293.

Spiegelhalter, D., Thomas, A., & Best, N. (2003). *WinBUGS (Version 1.4) [Computer software]*. Cambridge, UK: Biostatistics Unit, Institute of Public Health.

Spurk, D., Hirschi, A., Wang, M., Valero, D., & Kauffeld, S. (2020). Latent profile analysis: A review and "how to" guide of its application within vocational behavior research. *Journal of Vocational Behavior, 120*, 103445. https://doi.org/10.1016/j.jvb.2020.103445

Ulbricht, C. M., Chrysanthopoulou, S. A., Levin, L., & Lapane, K. L. (2018). The use of latent class analysis for identifying subtypes of depression: A systematic review. *Psychiatry Research, 266*, 228–246. https://doi.org/10.1016/j.psychres.2018.03.003

Vermunt, J. K., & Magidson, J. (2003). *Latent Gold 3.0*. Belmont, MA. URL http://www. statisticalinnovations.com.

Viroli, C. (2011). *FactMixtAnalysis: Factor Mixture Analysis with covariates*.

von Davier, M. (2006). *Multidimensional Latent Trait Modelling (MDLTM) [Computer Software]*. Educational Testing Service.

Wang, Y., Cao, C., & Kim, E. (2022). Covariate inclusion in factor mixture modeling: Evaluating one-step and three-step approaches under model misspecification and overfitting. *Behavior Research Methods*, 1–16. https://doi.org/10.3758/s13428-022-01964-8

Yung, Y. F. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika, 62*, 297–330. https://doi.org/10.1007/BF02294554

ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(2); 94-119

# The eTIMSS and TIMSS Measurement Invariance Study: Multigroup Factor Analyses and Differential Item Functioning Analyses with the 2019 Cycle

Murat Yalçınkaya* Hakan Atılgan** Selim Daşcıoğlu*** Burak Aydın****

**Abstract**

In this study, measurement invariance and differential item functioning (DIF) studies of the TIMSS 2019 4th and 8th-grade mathematics and science achievement tests were conducted for the country groups participating in both TIMSS and eTIMSS. The study sample consisted of 9560 responders of the first booklet of the 2019 cycle. Multiple Group Confirmatory Factor Analysis (MGCFA) was utilized to test measurement invariance, and Mantel-Haenszel (MH), Logistic Regression (LR), and SIBTEST were used for the DIF analyses. The measurement invariance results indicated strict invariance between groups for all tests which included 111 items in total. In the DIF analyses, for the 4th and 8th-grade mathematics tests, only three items showed moderate DIF with MH, and four items showed DIF with SIBTEST. For the 4th-grade science test, one item showed moderate DIF with both MH and SIBTEST. However, in the 8th-grade science test, no items showed DIF with MH and LR methods, while four items showed moderate DIF with SIBTEST. Overall, MH and SIBTEST techniques were in agreement, whereas LR method produced inconsistent results and showed disagreement with these two methods. The results of the measurement invariance analysis and the LR method were consistent and indicated equivalency of TIMSS and e-TIMSS scores.

***Keywords:*** *Multiple Group Confirmatory Factor Analysis, Differential Item Functioning, DIF, TIMSS, Computer-Based Assessments, Paper-Pencil Assessments*

## Introduction

In recent years the widespread use of technology in education and the measurement of psychometric properties have become more prevalent. The 1970s marked the first decade when tests started to be used in a computer-based environment (Drasgow, 2002). The widespread use of computers in homes and classrooms has played a significant role in improving the quality of tests and enabling the use of measurement tools in different forms. Before tests were transferred to electronic platforms, ensuring equivalence with traditional paper-pencil applications posed a significant problem. In the literature, there are numerous studies comparing computer-based systems with paper-pencil tests (Mills, Potenza, Fremer, Ward, 2002; Russel, Goldberg, O'Connor, 2003; Anakwe, 2008; Ergün, 2002; İlci, 2004; Maguire, Smith, Brallier, & Palm, 2009). However, it is observed that no such studies were conducted concerning the computer-based tests implemented in the Trends in International Mathematics and Science Study (TIMSS) 2019. During the TIMSS 2019 administration, approximately half of the participating countries chose to switch to eTIMSS, while the other half preferred paper-pencil-based administration (Mullis et al., 2020). Therefore, conducting studies that demonstrate whether computer-based and paper-pencil-based tests can

* MA., Ege University, Institute of Education Sciences, Measurement And Evaluation In Education, , Faculty of Education, İzmir-Turkey, muratyalcinkaya35@gmail.com, ORCID ID: 0000-0001-8564-3096
** Prof. Dr., Ege University, Faculty of Education, İzmir-Turkey, hakan.atilgan@ege.edu.tr, ORCID ID: 0000-0002-5562-3446
*** MA., Ege University, Institute of Education Sciences, Measurement And Evaluation In Education, , Faculty of Education, İzmir-Turkey, selimdascioglu@gmail.com, ORCID ID: 0000-0001-6820-4585
****Assoc. Prof., Ege University, Faculty of Education, İzmir-Turkey, burak.aydin@ege.edu.tr, ORCID ID:0000-0003-4462-1784

To cite this article:
Yalçınkaya, M., Atılgan, H., Daşçıoğlu S., & Aydın B. (2024). The eTIMSS and TIMSS measurement invariance study: Multigroup factor analyses and differential item functioning analyses with the 2019 cycle. *Journal of Measurement and Evaluation in Education and Psychology, 15*(2), 94-119. https://doi.org/10.21031/epod.1426739

*Received: 27.01.2024*
*Accepted: 30.05.2024*

be used interchangeably, and their measurement invariance and differential item functioning (DIF) is essential under these conditions (Gündoğmuş, 2017).

In general, validity, which forms the fundamental principle of this study, refers to the extent to which a measurement tool accurately measures the characteristic it intends to measure without confounding it with other attributes (Atılgan, Kan, & Aydın, 2017). It does not seem possible to refer to a more effective concept than validity in this sense (Rogers, 1995). In order to provide evidence for the construct validity of a measurement tool, studies on measurement invariance have gained prominence in the academic field. Measurement invariance is simply defined as evaluating the equality of measurement results for different groups (Moraes & Reichenheim, 2002). At the same time, measurement invariance stands out as a prerequisite in group comparisons (Meredith, 2006). Testing measurement invariance ensures that intergroup comparisons are meaningful. In cases where measurement invariance cannot be achieved, it is possible that one of the groups to be compared may have an advantage or disadvantage, leading to biased interpretations. Therefore, as in the present study, comparing countries and ranking them based on achievement scores increases the importance of measurement invariance analyses.

Furthermore, measurement invariance studies allow for interpreting data at the scale level between groups, and the determination of items showing DIF provides additional evidence for construct validity. Another positive aspect of DIF studies is that they contribute to identifying the reasons for the strengths and weaknesses of the compared groups (Klieme & Baumert, 2001). Thus, in-depth examinations at the item level in tests and subtests can provide insights into item bias and predict which group may have an advantage or disadvantage. Although different methods applied in DIF analysis generally yield similar results, they may not produce entirely consistent results due to their different matching criteria and cut-off values used for labeling items as DIF (Gök, Kelecioğlu, & Doğan, 2010). Therefore, considering all these factors, it is recommended that researchers use multiple methods in DIF analysis (Hambleton, 2006). In this study, three different DIF determination methods were utilized. While methods based on Item Response Theory (IRT) include separate structures for categorical items, this study will use MH, LR, and SIBTEST methods based fundamentally on CTT for dichotomous items. During the process of determining DIF, one group with equal ability level to the test-taking group is referred to as the reference group, while the other is referred to as the focal group (Holland & Wainer, 1993.

## Purpose and Significance of the Research

This study aims to analyze and interpret the findings regarding measurement invariance and DIF between paper-pencil tests and computer-based tests administered in TIMSS 2019. For this purpose, both scale-level Confirmatory Factor Analysis (CFA) for measurement invariance and item-level DIF analyses will be conducted for country groups participating in both paper-pencil and computer-based administrations. Additionally, it is believed that the data collected will provide insights for future similar test administrations and scientific studies.

In investigating the measurement invariance between computer-based and paper-pencil tests using the data obtained from the student achievement tests of TIMSS 2019, comparing the results from models without establishing measurement invariance would not be meaningful. It is essential to determine whether the items in the computer-based version provide advantages or disadvantages to test-takers compared to the items in the paper-pencil test.

TIMSS results, being one of the leading indicators in determining country's education policies, have been applied in our country in previous years using paper-pencil tests and in the latest administration using computer-based tests. Other countries are also gradually transitioning. Therefore, the purpose of this study is to evaluate the paper-pencil administration and computer-based administration in terms of measurement invariance and to identify whether DIF exists at the item level. This will contribute to the discussion of the

sustainability and feasibility of the transition to computer-based administration by examining its positive and negative aspects.

## Methods

The International Association for the Evaluation of Educational Achievement (IEA) conducts TIMSS every four years. In the TIMSS 2019 administration, 580,000 students from 64 countries participated, with the inclusion of seven more countries compared to TIMSS 2015. Among these countries, 32 opted for computer-based (eTIMSS) administration, while the other 32 preferred paper-pencil-based administration see Table 1.

**Table 1**

*Countries Participating in TIMSS 2019 Implementation*

| | | | |
|---|---|---|---|
| Germany * | Philippines | Japan | Sweetcorn |
| USA* | Finland* | Canada* | Norway* |
| Albania | France* | Montenegro | Pakistan |
| Australia | South Africa | Qatar* | Poland |
| Austria* | South Cyprus | Kazakhistan | Portugal* |
| Azerbaijan | Georgia* | South Korea* | Romania |
| Bahrain | Croatia* | Kosovo | Russia* |
| Belgium (Flemish Region) | Holland* | Kuwait | Serbia |
| UAE* | Hong Kong* | North Ireland | Singapore* |
| Bosnia and Herzegovina | England* | North Macedonia | Slovakia* |
| Bulgaria | Iranian | Latvia | Saudi Arabia |
| Czech Republic* | Ireland | Lithuania* | Chile* |
| Taiwan* | Spain* | Lebanon | Türkiye* |
| Denmark* | Israel* | Hungary* | Oman |
| Armenia | Sweden* | Malaysia* | Jordan |
| Morocco | Italy* | Malta* | New Zeland |

**Countries participating in eTIMSS (MEB,2020)*

In studies involving 4th-grade students, certain countries (Albania, Bosnia and Herzegovina, Kosovo, Kuwait, Montenegro, Morocco, North Macedonia, Pakistan, Philippines, Saudi Arabia, South Africa) have preferred to use "Less Difficult Mathematics" test versions, and therefore, they were not included in this study (Mullis et al., 2020).

As a result, in this study, 29 countries participated in the paper-pencil-based administration, and 30 countries participated in the computer-based administration for the 4th-grade mathematics test. Similarly, the countries Jordan, Romania, Israel, Malaysia, Egypt did not participate in the 4th and 8th-grade mathematics and science assessments. For the 8th-grade mathematics and science tests, 17 countries participated in the paper-pencil-based administration, while 22 countries opted for computer-based administration (MEB, 2020). In the studies, only one randomly selected test booklet was examined for all grade levels and tests (Table 2). The distribution frequency of this booklet among the students was similar or very close to the frequencies observed in all other booklets (7.2%).

**Table 2**

*Booklet Usage Rates for TIMSS 2019 Mathematics 4th-grade Test*

| Booklets | Frequency | Percentage | Current Percentage | Additive Percentage |
|---|---|---|---|---|
| Booklet 1 | 9560 | 7.2 | 7.2 | 7.2 |
| Booklet 2 | 9480 | 7.1 | 7.1 | 14.3 |
| Booklet 3 | 9505 | 7.1 | 7.1 | 21.4 |
| Booklet 4 | 9517 | 7.1 | 7.1 | 28.5 |
| Booklet 5 | 9543 | 7.2 | 7.2 | 35.7 |
| Booklet 6 | 9521 | 7.1 | 7.1 | 42.8 |
| Booklet 7 | 9586 | 7.2 | 7.2 | 50.0 |
| Booklet 8 | 9509 | 7.1 | 7.1 | 57.2 |
| Booklet 9 | 9506 | 7.1 | 7.1 | 64.3 |
| Booklet 10 | 9498 | 7.1 | 7.1 | 71.4 |
| Booklet 11 | 9517 | 7.1 | 7.1 | 78.6 |
| Booklet 12 | 9543 | 7.2 | 7.2 | 85.7 |
| Booklet 13 | 9514 | 7.1 | 7.1 | 92.9 |
| Booklet 14 | 9529 | 7.1 | 7.1 | 100.0 |
| Total | 133328 | 100.0 | 100.0 | |

Derived items (Annex 13) were scored by taking the integrated answer part (TIMSS, 2019).

The integrated response part of the derived items (Appendix 13) was scored in TIMSS 2019. The extensions of the derived items were not considered, and the responses to the binary items were coded as "1" if all sub-items were answered correctly, and "0" if not. Therefore, the number of items in the 8th-grade science test, for example, was 44 for the derived items, including their sub-items, but after arranging the dependent items, 31 items were included in the analysis. The table resulting from the item matching process and the corresponding number of students are presented in Table 3.

**Table 3**
*TIMSS 2019 Number of Items and Students*

| GROUP | NUMBER OF ITEMS | NUMBER OF STUDENTS |
|---|---|---|
| | *4th GRADE* | |
| TIMSS MATHEMATICS | 24 | 5373 |
| eTIMSS MATHEMATICS | 24 | 8917 |
| TIMSS SCIENCE | 25 | 9284 |
| eTIMSS SCIENCE | 25 | 9264 |
| | *8th GRADE* | |
| TIMSS MATHEMATICS | 31 | 7326 |
| eTIMSS MATHEMATICS | 31 | 7270 |
| TIMSS SCIENCE | 31 | 7224 |
| eTIMSS SCIENCE | 31 | 7930 |

In all booklets, care was taken to ensure an equal distribution of item types and numbers, and to distribute the booklets to as close to an equal number of students as possible. The data for the 4th and 8th grades included in the study were downloaded and organized from the official website of the TIMSS&PIRLS International Study Center.

## Analysis of Data

The evaluation of the TIMSS 2019 mathematics and science test items involved completing studies on missing data, followed by an examination of outliers. Among the main methods chosen by researchers for dealing with missing data are data deletion, estimation of missing data using imputation methods, and approximate value assignment to missing data (Büyüköztürk, Çokluk, & Şekercioğlu, 2014). Regarding the present study, due to the size of the data set and the missing data rate being less than 5% and considered random, data deletion method was selected as the most appropriate approach (Tabachnick & Fidell, 2007). During the examination of missing data, responses to items labeled as "9" in the data set, indicating that the student left the answer blank because they did not know the correct response, were coded as "0". Responses coded as "6", representing patterns where the student did not encounter the item due to technical issues or insufficient time during the exam, were removed from the data set.
Subsequently, CFA and Multiple Group Confirmatory Factor Analysis (MGCFA) were conducted. Given that the research data were categorical, the assumption of normality was not tested. Furthermore, the multicollinearity assumption was examined by assessing the tetrachoric correlation between items, and it was observed that all correlations were below .90. Additionally, Variance Inflation Factors (VIF), Tolerance Levels, and Condition Indices (CI) were examined, and it was found that CI values were below 30, VIF values were below 10, or tolerance values were above .10, indicating the absence of multicollinearity issues (Kline, 2016; Hair, Anderson, Tatham, & Black, 1998; Mertler & Vannatta, 2005; Tabachnick & Fidell, 2007). The VIF and tolerance values for each subscale are provided in Appendix 1 through Appendix 4; tetrachoric correlation coefficients are provided in Appendix 5 through Appendix 8.
The Weighted Least Squares Mean and Variance (WLSMV) method was employed as the parameter estimation method in CFA and MGCFA. It is noted in the literature that the asymptotically distribution-free estimator is used in conjunction with ordinal categorical data. WLSMV, utilized in analyses with ordinal categorical data, produces better results based on polychoric correlations, accuracy of parameter estimates, and estimated standard errors. In other words, polychoric correlations are reported to provide the

best estimates of model parameters (Joreskog & Sorbom, 1981). WLSMV can be considered as an alternative method for non-normally distributed, highly skewed, or platykurtic ordinal data (Muthén, 1993). In this study, the established models were confirmed through Confirmatory Factor Analysis for the entire data set, obtaining evidence for construct validity. The learning domains specified by TIMSS were used as the sub-dimensions in the analysis (Mullis et al., 2020). CFA was conducted using the M*plus* 7.4 program with the WLSMV estimation method (Jöreskog & Sörbom, 2006).

CFA analyses were conducted to confirm the subscales specified by TIMSS. Additionally, the path diagrams of the CFA analyses performed using the M*plus* 7 program are provided in Appendix 9 through 12. Table 4 illustrates how model-data fit is assessed based on the fit indices obtained from the CFA results based on $\chi^2$/df (Kline, 2016), CFI (Bentler, 1980), SRMR and RMSEA (Browne & Cudeck, 1993).

**Table 4**

*Cut off values to be used in the evaluation of CFA fit indices*

| Fit Index | Good Fit | Acceptable Fit |
|---|---|---|
| $\chi$2 | p>.05 | p>.05 |
| $\chi$2/df | 0≤ $\chi$2/df≤2 | 2≤ $\chi$2/df≤8 |
| CFI | .97≤CFI≤1.00 | .95≤CFI<.97 |
| TLI | .95≤TLI≤1.00 | .90≤TLI<.95 |
| RMSEA | 0≤RMSEA≤.05 | .05<RMSEA≤.08 |

For the 4th-grade science test, the three-factor model (life sciences, physical sciences, and earth sciences) demonstrated an acceptable fit ($\chi^2$/df =5.955, CFI=.990, TLI=.989, and RMSEA=.016). Similarly, for the 8th-grade science test, the four-factor model (physics, chemistry, biology, and earth sciences) showed an acceptable fit ($\chi^2$/df = 8.795, CFI=.981, TLI=.979, and RMSEA=.023). For the 4th-grade mathematics test, the three-factor model (numbers, data, measurement, and geometry) displayed a considerably lower $\chi$2/df (37.749) statistic, indicating an acceptable fit, while the CFI (.953) indicated a good fit, and the TLI (.947) and RMSEA (.051) showed an acceptable fit. For the 8th-grade mathematics test, the four-factor model (numbers, algebra, geometry, data, and probability) exhibited an acceptable fit with a $\chi$2/df (13.938) statistic below the acceptable limit, and a good fit based on the CFI (.981), TLI (.979), and RMSEA (.030) statistics. MGCFA based on structural equation modeling was used to assess measurement invariance. In the literature, there are different views among researchers regarding the number of steps and the nature of operations involved in evaluating measurement invariance. In this study, a 4-step hierarchical model, encompassing configural, metric, scalar, and strict invariance, will be employed (Steenkamp & Baumgartner, 1998; Wu, Li, & Zumbo, 2007; Byrne, 2008; Meredith & Teresi, 2006).

**Table 5**

*Parameters Used in Measurement Invariance Analysis*

| Invariance Model | Fixed Parameters | Tested Parameters |
|---|---|---|
| Configural Invariance | - | Item/Factor groups |
| Metric Invariance | Factor variances and covariances | Factor loadings |
| Scalar Invariance | + Factor and observed variable means | Intercepts or thresholds |
| Strict Invariance | + Observed Variances and Covariances | Residual variances |

As shown in Table 5, in each step, one additional parameter is added and fixed at each stage to the parameters kept constant (Gregorich, 2006). Moreover, with each step, one more parameter is added and fixed in the tested parameters. In measurement invariance studies categorical variables can be forced to fit these four steps (e.g., Li, Gooden & Toland, 2016) or the number of steps can be reduced based on the number of categories (e.g., Bagdu Soyler, Aydın & Atılgan, 2021; titina et al., 2020; Raykov et al., 2018). In our analyses we preferred to use the four-step approach given that it is more common with the TIMSS analyses.

**Fit Indices**

MGCFA is based on Structural Equation Modeling (SEM) and allows simultaneous testing of the model in multiple groups (Tabachnick & Fidell, 2007). In the first stage of the study, which is within the scope of the MGCFA technique, CFI, TLI, and RMSEA are used to evaluate the model-data fit. In each step of the invariance testing, differences between CFI and TLI are used to provide information about the relationship between latent scores and observed scores. It is noted that CFI, TLI, and RMSEA fit indices should fall within the desired range, with $.01 \geq \Delta CFI \geq -.01$ and $.01 \geq \Delta TLI \geq -.01$ for each step of the MGCFA data sets (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000). However, $\chi^2$ statistic, being influenced by sample size, is considered in large samples like this study by taking into account other fit indices (Brown, 2006; Büyüköztürk, 2010; Tabachnick & Fidel, 2007). In the literature, it has been stated that the $\chi^2$ difference used for measurement invariance analyses should not be used alone (Wu, Li, & Zumbo, 2007), and other findings have been reported (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000). Further for the estimators appropriate for categorical data regular $\chi^2$ tests might not be appropriate adjustments might be needed, in M*plus* this is handled with the DIFFTEST command, and its technical details are briefly studied by Kite, Johnson and Xing (2018).

After MGCFA, the derived test items were evaluated for DIF using the MH, LR, and SIBTEST procedures. While test-level CFA can be used to evaluate measurement invariance, DIF can be used for item and subtest level analyses, as observed in the literature (Cheung & Rensvold, 2002; Raju, Laffitte, & Byrne, 2002). DIF is defined as the differentiation of the probability of correctly answering a test item among different subgroups of individuals with equal abilities (Camilli & Shepard, 1994; Zumbo, 1999). DIF determination techniques based on the Classical Test Theory (CTT) are index-dependent sampling techniques (Camilli & Shepard, 1994). In the CTT-based methods, separate procedures are used for polytomous and dichotomous items. In this study, the MH, LR, and SIBTEST methods will be used for comparing the results of DIF obtained for dichotomous tests. Unlike the MGCFA, the DIF analyses were conducted separately for test dimensions. Even though it is possible to conduct multidimensional DIF (e.g., Bulut & Suh, 2017) our attempts to utilize *mirt* (Chalmers, 2012) package was unsuccessful probably due to the large sample size and relatively complex factor structure.

**Mantel-Haenszel (MH)**

William Haenszel and Nathan Mantel developed the DIF determination method based on the chi-square statistic in the 1950s. This technique is a method used in tests containing dichotomously scored items. The odds ratio ($\alpha$) calculates the degree of performance difference between the reference and focal groups, in other words, the ratio of individuals answering correctly and incorrectly in each ability level for both reference and focal groups, taking into account the total number of respondents (Mertler and Vannatta, 2005; Agresti, 1984). To express MH more effectively, the natural logarithm is obtained, and $\Delta MH$ (delta coefficient) is determined through a logarithmic transformation. When determining DIF with the MH technique, the following interpretations are made: if $\Delta MH=0$ or $\alpha=1$, there is no DIF in the item; if $\Delta MH<0$ or $\alpha>1$, there is DIF in favor of the reference group; if $\Delta MH>0$ or $\alpha<1$, there is DIF in favor of the focal group (Camilli and Shepard, 1994; Nandakumar, 1993). Additionally, if $|\Delta MH|<1$, DIF in the item is negligible (Level A); if $1 \leq |\Delta MH| < 1.5$, DIF in the item is moderate (Level B); if $|\Delta MH| \geq 1.5$, DIF in the item is significant (Level C) (Dorans & Holland, 1993; Zieky, 1993).

### Logistic Regression (LR)

LR is a regression model used when the dependent variable is binary (1-0). In other words, LR is used when it is expected that the dependent variable will exhibit responses in a non-linear relationship with one or more independent variables (Tabachnick & Fidell, 2007). LR is a non-parametric method.

The standardized regression coefficients are considered LR effect sizes (Gierl, Jodoin & Ackerman, 2000). The standardized regression coefficients ($R^2$) provide the degree of DIF (Differential Item Functioning), and they are determined in three levels. If $R^2<.035$ for the difference between Model 1 and Model 3, there is no DIF or it is negligible. If $.035\leq R^2<.070$, there is moderate-level (B) DIF. If $R^2\geq .070$, there is significant-level (C) DIF. For an item to be classified as having DIF (B or C level), the chi-square value must be statistically significant at the .05 level or less, and the $R^2$ value must be at least .035 (Zumbo, 1999). Additionally, for items with identified DIF, the presence of non-uniform DIF is examined by checking if the difference between the $R^2$ values of Model 2 and Model 3 is greater than .035. If it is greater, non-uniform DIF can be considered.

### SIBTEST

The SIBTEST method can statistically demonstrate whether one or more items exhibit DIF (Shealy & Stout, 1993). SIBTEST is used in DIF analyses for dichotomous data and can estimate the degree of DIF exhibited by an item. As a non-parametric method based on the IRT, SIBTEST provides a more precise synchronization of the focal and reference groups (Osterlind & Everson, 2009).

The β index primarily represents the effect size. A positive index value indicates DIF in favor of the reference group, while a negative value indicates DIF in favor of the focal group. If $\beta<|.059|$, the item is considered to have negligible DIF (Level A), if $|.059|\leq\beta<|.088|$, it has moderate DIF (Level B), and if $\beta\geq|.088|$, it has substantial DIF (Level C) (Rousses & Stout, 1996).

## Results

The first stage of measurement invariance, known as configural invariance, examines whether the structure is comparable across groups. When looking at the fit indices for the 4th grade mathematics test, as shown in Table 6, all values, including RMSEA (.051), CFI (.952), and TLI (.947), fall within an acceptable range of fit. The $\chi^2$/sd (19.720) value falls outside the specified intervals for the likelihood, as a result of biased results in large samples (Kline, 2016). Hence, as expected, all $\chi^2$ difference tests reported in the Table 6, including the one for the 4th grade mathematics are significant. However, all other values are within the permitted minimum level intervals, confirming that the structure is similar across all groups, and the model demonstrates invariance at all stages between the TIMSS 4th-grade mathematics test using paper and pencil and computer-based methods.

**Table 6**

*Measurement Invariance Results by TIMSS 2019 Tests Participation Pattern (eTIMSS/TIMSS)*

| Test | Invariance Type | $\chi^2$/sd | $\Delta\chi^2$ | RMSEA | CFI | TLI | ΔCFI | ΔTLI |
|---|---|---|---|---|---|---|---|---|
| 4th-grade Science | Configural | 3.993 | | 0.018 | 0.987 | 0.985 | | |
| | Weak | 5.636 | 438.32* | 0.022 | 0.979 | 0.977 | 0.008 | 0.008 |
| | Strong | 6.366 | 621.96* | 0.024 | 0.974 | 0.974 | 0.005 | 0.003 |
| | Strict | 5.269 | 459.12* | 0.021 | 0.980 | 0.979 | -0.006 | -0.005 |
| 4th-grade Mathematics | Configural | 19.720 | | 0.051 | 0.952 | 0.947 | | |
| | Weak | 15.716 | 264.95* | 0.045 | 0.961 | 0.959 | -0.009 | -0.012 |
| | Strong | 16.177 | 692.81* | 0.046 | 0.958 | 0.957 | 0.003 | 0.002 |
| | Strict | 19.677 | 312.34* | 0.051 | 0.951 | 0.947 | 0.007 | 0.010 |
| 8th-grade Science | Configural | 5.594 | | 0.025 | 0.975 | 0.973 | | |
| | Weak | 5.362 | 284.62* | 0.024 | 0.976 | 0.975 | -0.001 | -0.002 |
| | Strong | 5.701 | 543.88* | 0.025 | 0.973 | 0.973 | 0.003 | 0.002 |
| | Strict | 6.146 | 300.85* | 0.026 | 0.972 | 0.970 | 0.001 | 0.003 |
| 8th-grade Mathematics | Configural | 7.968 | | 0.031 | 0.978 | 0.976 | | |
| | Weak | 5.955 | 271.77* | 0.026 | 0.984 | 0.983 | -0.006 | -0.007 |
| | Strong | 7.122 | 1235.06* | 0.029 | 0.979 | 0.979 | 0.005 | 0.004 |
| | Strict | 8.920 | 344.67* | 0.033 | 0.974 | 0.973 | 0.005 | 0.006 |

Note: * p<.05

Similarly, when examining the 8th-grade mathematics test, during the stage of configural invariance, all values, including RMSEA (.031), CFI (.978), and TLI (.976), fall within the good fit range. Except for the $\chi^2$ tests, it can be observed that the structure is similar across groups, and the model demonstrates invariance at all stages based on the participation method for the 8th grade mathematics test.

Except for the $\chi^2$ tests, it is observed that strict invariance is achieved in the 4th and 8th grade science test. As a result, when examining Table 6 which show the goodness-of-fit indices as well as the differences between ΔCFI and ΔTLI values considered after structural invariance at all stages of measurement invariance for both 4th and 8th-grade mathematics and science tests, it is evident that the differences are within acceptable limits, indicating the achievement of strict invariance stages.

In the context of the TIMSS and eTIMSS samples, combined data sets were analyzed using MH, SIBTEST, and LR techniques to identify items exhibiting DIF based on the participation format. α, β, and $\Delta R^2$ coefficients were computed, and the directions and magnitudes of these coefficients were taken into account to determine the level of DIF for matched items between paper-pencil and computer-based formats. As mentioned earlier, DIF analyzes were performed separately for each sub-dimension of the tests.

**Table 7**

*DIF Status of 4th Grade Mathematics Test Items in Booklet No. 1 in TIMSS 2019 Implementation by Country Groups Participating in TIMSS/eTIMSS*

| Sub-dimension | Item | MH | | | | | LR | | | | SIBTEST | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | α | χ2 | p | ΔMH | DIF Level, Direction | Δχ2 | p | $\Delta R^2$ | DIF Level, Direction | β | χ2 | p | DIF Level, Direction |
| Number | M1 | .776 | 41.946 | <.001 | .595 | A | 101.280 | <.001 | <.035 | A | -.054 | 40.733 | <.001 | A |
| | M2 | 1.224 | 18.524 | <.001 | -.475 | A | 36.838 | <.001 | <.035 | A | .034 | 22.425 | <.001 | A |
| | M3 | .663 | 64.900 | <.001 | .967 | A | 81.177 | <.001 | <.035 | A | -.056 | 66.308 | <.001 | A |
| | M4 | 1.280 | 26.918 | <.001 | -.581 | A | 35.640 | <.001 | <.035 | A | .039 | 27.197 | <.001 | A |
| | M5 | 1.625 | 94.205 | <.001 | -1.141 | B- | 103.062 | <.001 | <.035 | A | .071 | 94.535 | <.001 | B- |
| | M6 | .846 | 13.008 | <.001 | .392 | A | 15.188 | .001 | <.035 | A | -.030 | 15.366 | <.001 | A |
| | M13 | 1.118 | 4.675 | .031 | -.263 | A | 7.710 | .021 | <.035 | A | .016 | 5.494 | .019 | A |
| | M14 | 1.065 | 2.032 | .154 | -.148 | A | 3.958 | .138 | <.035 | A | .012 | 2.130 | .144 | A |
| | M15 | .949 | .997 | .318 | .123 | A | 65.859 | <.001 | <.035 | A | -.012 | 3.351 | .067 | A |
| | M16 | .921 | 2.662 | .103 | .193 | A | 3.661 | .160 | <.035 | A | -.007 | .839 | .360 | A |
| | M17 | .958 | .504 | .478 | .102 | A | 24.160 | <.001 | <.035 | A | -.011 | 3.060 | .080 | A |
| Measurement and Geometry | M7 | 1.598 | 126.864 | <.001 | -1.101 | B- | 129.704 | <.001 | <.035 | A | .098 | 132.030 | <.001 | C- |
| | M8 | .685 | 69.199 | <.001 | .890 | A | 70.655 | <.001 | <.035 | A | -.065 | 77.081 | <.001 | B+ |
| | M9 | 1.456 | 64.410 | <.001 | -.882 | A | 73.766 | <.001 | <.035 | A | .059 | 62.752 | <.001 | B- |
| | M10 | .782 | 32.395 | <.001 | .577 | A | 34.202 | <.001 | <.035 | A | -.041 | 26.267 | <.001 | A |
| | M18 | 1.285 | 36.013 | <.001 | -.590 | A | 47.811 | <.001 | <.035 | A | .051 | 37.031 | <.001 | A |
| | M19 | .902 | 5.671 | .017 | .242 | A | 7.400 | .025 | <.035 | A | -.022 | 6.672 | .010 | A |
| | M20 | 1.070 | 1.585 | .208 | -.158 | A | 6.253 | .044 | <.035 | A | .006 | .741 | .389 | A |
| | M21 | .588 | 123.179 | <.001 | 1.249 | B+ | 131.190 | <.001 | <.035 | A | -.086 | 140.335 | <.001 | B+ |
| Data | M11 | 1.178 | 11.102 | .001 | -.386 | A | 13.130 | .001 | <.035 | A | .031 | 10.908 | .001 | A |
| | M12 | 1.082 | 1.631 | .202 | -.185 | A | 1.621 | .445 | <.035 | A | .010 | 1.541 | .215 | A |
| | M22 | 1.200 | 10.429 | .001 | -.428 | A | 11.626 | .003 | <.035 | A | .023 | 8.032 | .005 | A |
| | M23 | .712 | 44.072 | <.001 | .799 | A | 44.553 | <.001 | <.035 | A | -.057 | 41.349 | <.001 | A |
| | M24 | .958 | .661 | .416 | .102 | A | 7.639 | .022 | <.035 | A | -.006 | 0.452 | .501 | A |

+/-: DIF favors focal/reference group.

Based on the MH results, out of the 24 items in the 4th grade mathematics test of TIMSS 2019, 21 exhibited negligible levels of DIF (Level A), while 3 items showed moderate DIF (level B). Item 21 favors students taking the paper-pencil version, whereas item 5 and 7 favor students taking the computer-based version (see Table 7). On the other hand, the LR results indicated that all items in the 4th grade mathematics test exhibited negligible levels of DIF (Level A). As for the SIBTEST results, 19 items were found to have negligible levels of DIF (level A), 4 items showed DIF at Level B, and 1 item showed DIF at Level C (see Table 7). Based on the SIBTEST analyses, items 8 and 21 favor students taking the paper-pencil version, items 5, 7 and 9 favor students taking the computer-based version.

**Table 8**

*DIF Status of 8th Grade Mathematics Test Items in Booklet No. 1 in TIMSS 2019 Implementation by Country Groups Participating in TIMSS/eTIMSS*

| Subest | Item | MH | | | | | LR | | | | SIBTEST | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | α | χ2 | p | ΔMH | DIF Level, Direction | Δχ2 | p | $\Delta R^2$ | DIF Level, Direction | β | χ2 | p | DIF Level, Direction |
| Number | M1 | .982 | .167 | .683 | .043 | A | 2.574 | .276 | <.035 | A | .011 | 1.907 | .167 | A |
| | M2 | 1.762 | 148.537 | <.001 | -1.331 | B- | 145.679 | <.001 | <.035 | A | .098 | 152.008 | <.001 | C- |
| | M3 | 1.066 | 1.190 | .275 | -.151 | A | 13.552 | .001 | <.035 | A | -.010 | 2.308 | .129 | A |
| | M4 | 1.078 | 3.010 | .083 | -.177 | A | 12.194 | .002 | <.035 | A | .027 | 10.701 | .001 | A |
| | M5 | .489 | 275.735 | <.001 | 1.679 | C+ | 309.959 | <.001 | <.035 | A | -.118 | 203.741 | <.001 | C+ |
| | M17 | 1.070 | 1.386 | .239 | -.159 | A | 8.871 | .012 | <.035 | A | -.003 | 0.216 | .643 | A |
| | M18 | 1.196 | 17.494 | <.001 | -.420 | A | 17.702 | <.001 | <.035 | A | .044 | 28.392 | <.001 | A |
| | M19 | .739 | 25.779 | <.001 | .711 | A | 28.878 | <.001 | <.035 | A | -.041 | 47.593 | <.001 | A |
| | M20 | 1.094 | 4.510 | .034 | -.211 | A | 17.759 | <.001 | <.035 | A | .038 | 20.277 | <.001 | A |
| Algebra | M6 | 1.055 | 1.556 | .212 | -.125 | A | 5.014 | .082 | <.035 | A | .010 | 1.490 | .222 | A |
| | M7 | .860 | 11.903 | .001 | .354 | A | 28.760 | <.001 | <.035 | A | -.028 | 12.584 | <.001 | A |
| | M8 | .693 | 81.117 | <.001 | .863 | A | 89.486 | <.001 | <.035 | A | -.063 | 56.990 | <.001 | B+ |
| | M9 | .432 | 172.488 | <.001 | 1.974 | C+ | 194.924 | <.001 | <.035 | A | -.084 | 225.012 | <.001 | B+ |
| | M10 | .896 | 6.384 | .012 | .258 | A | 9.926 | .007 | <.035 | A | -.012 | 2.340 | .126 | A |
| | M21 | 1.531 | 98.762 | <.001 | -1.001 | B- | 102.884 | <.001 | <.035 | A | .078 | 94.783 | <.001 | B- |
| | M22 | .895 | 6.846 | .009 | .261 | A | 7.176 | .028 | <.035 | A | -.019 | 5.617 | .018 | A |
| | M23 | 1.220 | 16.806 | <.001 | -.467 | A | 17.532 | <.001 | <.035 | A | .021 | 8.533 | .004 | A |
| | M24 | 1.434 | 40.271 | <.001 | -.846 | A | 40.972 | <.001 | <.035 | A | .027 | 18.143 | <.001 | A |
| | M25 | 1.341 | 52.384 | <.001 | -.689 | A | 61.148 | <.001 | <.035 | A | .069 | 69.199 | <.001 | B- |
| Geometry | M11 | .592 | 148.233 | <.001 | 1.232 | B+ | 167.335 | <.001 | <.035 | A | -.069 | 62.568 | <.001 | B+ |
| | M12 | 1.379 | 53.325 | <.001 | -.755 | A | 47.148 | <.001 | <.035 | A | .068 | 59.885 | <.001 | B- |
| | M13 | .961 | .764 | .382 | .094 | A | 6.522 | .038 | <.035 | A | -.011 | 1.444 | .230 | A |
| | M26 | .705 | 52.960 | <.001 | .822 | A | 70.726 | <.001 | <.035 | A | -.080 | 82.549 | <.001 | B+ |
| | M27 | 1.543 | 108.160 | <.001 | -1.019 | B- | 101.924 | <.001 | <.035 | A | .104 | 127.927 | <.001 | C- |
| | M28 | 1.127 | 5.435 | .020 | -.281 | A | 6.991 | .030 | <.035 | A | -.005 | 0.321 | .571 | A |
| Data and Probability | M14 | 1.080 | 2.497 | .114 | -.180 | A | 40.866 | <.001 | <.035 | A | .030 | 11.378 | .001 | A |
| | M15 | .819 | 17.918 | <.001 | .470 | A | 49.633 | <.001 | <.035 | A | -.021 | 6.115 | .013 | A |
| | M16 | .907 | 4.839 | .028 | .231 | A | 8.239 | .016 | <.035 | A | -.005 | 0.265 | .607 | A |
| | M29 | 1.765 | 114.973 | <.001 | -1.335 | B- | 117.194 | <.001 | <.035 | A | .065 | 65.862 | <.001 | B- |
| | M30 | .978 | .241 | .623 | .053 | A | 13.436 | .001 | <.035 | A | .020 | 4.772 | .029 | A |
| | M31 | .714 | 28.362 | <.001 | .792 | A | 32.810 | <.001 | <.035 | A | -.042 | 43.260 | <.001 | A |

+/-: DIF favors focal/reference group.

In the TIMSS 2019 8th grade mathematics test, MH results shows that 5 items have DIF at Level B, and 2 items have DIF at Level C, as reported in Table 8. Item 5, 9, and 11 favor students taking the paper-pencil

version, while item 2, 22, 27 and 29 favor students taking the computer-based version. However, based on the LR results, all items showed negligible levels of DIF (Level A). As for the SIBTEST results, 8 items were found to have DIF at Level B, and 3 items exhibited DIF at Level C. Similarly, item 5, 8, 9, 11 and 26 favored students taking the paper-pencil version, while item 2, 12, 21, 25, 27 and 29 favored students taking the computer-based version in terms of DIF.

**Table 9**

*DIF Status of 4th Grade Science Subtest Items in Booklet No. 1 in TIMSS 2019 Implementation by Country Groups Participating in eTIMSS/TIMSS*

| Subtest | Item | MH | | | | | LR | | | | SIBTEST | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | α | χ2 | p | ΔMH | DIF Level, Direction | Δχ2 | p | $\Delta R^2$ | DIF Level, Direction | β | χ2 | p | DIF Level, Direction |
| Life | M1 | 1.389 | 73.972 | <.001 | -.771 | A | 88.636 | <.001 | <.035 | A | .059 | 76.872 | <.001 | B- |
| | M2 | 1.032 | .734 | .392 | -.073 | A | 21.921 | <.001 | <.035 | A | .004 | .273 | .602 | A |
| | M3 | 1.181 | 13.957 | <.001 | -.390 | A | 30.329 | <.001 | <.035 | A | .013 | 5.102 | .024 | A |
| | M4 | .637 | 77.031 | <.001 | 1.060 | B+ | 77.856 | <.001 | <.035 | A | -.047 | 92.492 | <.001 | A |
| | M5 | .719 | 90.511 | <.001 | .774 | A | 84.115 | <.001 | <.035 | A | -.067 | 86.005 | <.001 | B+ |
| | M6 | .941 | 2.598 | .107 | .144 | A | 2.194 | .334 | <.035 | A | -.008 | 1.395 | .238 | A |
| | M13 | .667 | 134.066 | <.001 | .952 | A | 152.259 | <.001 | <.035 | A | -.082 | 127.134 | <.001 | B+ |
| | M14 | 1.059 | 2.172 | .141 | -.135 | A | 5.944 | .051 | <.035 | A | .009 | 1.868 | .172 | A |
| | M15 | 1.261 | 42.470 | <.001 | -.545 | A | 55.523 | <.001 | <.035 | A | .040 | 31.547 | <.001 | A |
| | M16 | 1.109 | 7.122 | .008 | -.242 | A | 33.526 | <.001 | <.035 | A | -.002 | .049 | .824 | A |
| | M17 | 1.067 | 3.646 | .056 | -.153 | A | 22.452 | <.001 | <.035 | A | .036 | 24.023 | <.001 | A |
| | M18 | 1.217 | 28.456 | <.001 | -.462 | A | 52.723 | <.001 | <.035 | A | .023 | 10.757 | .001 | A |
| Physycal | M7 | .944 | 2.694 | .101 | .135 | A | 18.744 | <.001 | <.035 | A | .001 | 0.009 | .926 | A |
| | M8 | .993 | .029 | .866 | .016 | A | 1.981 | .371 | <.035 | A | .001 | 0.035 | .851 | A |
| | M9 | 1.257 | 36.680 | <.001 | -.538 | A | 42.914 | <.001 | <.035 | A | .026 | 12.489 | <.001 | A |
| | M10 | .926 | 4.479 | .034 | .181 | A | 8.146 | .017 | <.035 | A | -.008 | 1.264 | .261 | A |
| | M19 | .854 | 21.015 | <.001 | .370 | A | 29.607 | <.001 | <.035 | A | -.024 | 9.595 | .002 | A |
| | M20 | 1.389 | 83.992 | <.001 | -.773 | A | 84.483 | <.001 | <.035 | A | .066 | 79.369 | <.001 | B- |
| | M21 | .840 | 23.399 | <.001 | .410 | A | 25.570 | <.001 | <.035 | A | -.027 | 13.287 | <.001 | A |
| | M22 | .956 | 1.372 | .242 | .105 | A | 2.434 | .296 | <.035 | A | -.020 | 7.441 | .006 | A |
| Earth | M11 | .941 | 2.770 | .096 | .143 | A | 4.793 | .091 | <.035 | A | .006 | .540 | .463 | A |
| | M12 | 1.516 | 111.321 | <.001 | -.978 | A | 117.528 | <.001 | <.035 | A | .081 | 98.431 | <.001 | B- |
| | M23 | .654 | 119.245 | <.001 | 1.000 | A | 192.285 | <.001 | <.035 | A | -.096 | 144.527 | <.001 | C+ |
| | M24 | .987 | .082 | .775 | .030 | A | 3.174 | .205 | <.035 | A | -.027 | 12.278 | .001 | A |
| | M25 | 1.094 | 6.205 | .013 | -.212 | A | 90.601 | <.001 | <.035 | A | .047 | 32.750 | <.001 | A |

+/-: DIF favors focal/reference group.

Based on the MH results reported in Table 9, in the TIMSS 2019 4th grade science test consisting of 25 items only 1 item exhibited DIF at Level B favors students taking the paper-pencil version, and no items showed DIF at Level C. Based on the LR results, all items showed negligible levels of DIF (Level A). For the SIBTEST results, 5 items exhibited DIF at Level B, indicating that 1 item showed DIF at this level.

Therefore, based on the SIBTEST results, item 5, 13 and 23 favored students taking the paper-pencil version, while item 1 and 12 favored students taking the computer-based version in terms of DIF.

**Table 10**

*DIF Status of 8th-grade Science Subtest Items in Booklet No. 1 in TIMSS 2019 Implementation by Country Groups Participating in eTIMSS/TIMSS*

| Subtest | Item | MH | | | | | LR | | | | SIBTEST | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | α | χ2 | p | ΔMH | DIF Level, Direction | Δχ2 | p | $\Delta R^2$ | DIF Level, Direction | β | χ2 | p | DIF Level, Direction |
| Biology | M1 | 1.280 | 37.819 | <.001 | -.581 | A | 40.502 | <.001 | <.035 | A | .048 | 39.996 | <.001 | A |
| | M2 | .926 | 4.270 | .039 | .180 | A | 9.260 | .010 | <.035 | A | .003 | .164 | .686 | A |
| | M3 | 1.497 | 109.744 | <.001 | -.948 | A | 109.631 | <.001 | <.035 | A | .084 | 113.852 | <.001 | B- |
| | M4 | .896 | 5.957 | .015 | .257 | A | 10.021 | .007 | <.035 | A | -.029 | 16.962 | <.001 | A |
| | M5 | .847 | 18.905 | <.001 | .390 | A | 65.766 | <.001 | <.035 | A | -.030 | 13.562 | <.001 | A |
| | M15 | 1.043 | 1.162 | .281 | -.099 | A | 1.832 | .400 | <.035 | A | .014 | 2.950 | .086 | A |
| | M16 | 1.041 | 1.157 | .282 | -.094 | A | 9.258 | .010 | <.035 | A | .024 | 8.795 | .003 | A |
| | M17 | .898 | 6.551 | .011 | .253 | A | 7.557 | .023 | <.035 | A | -.034 | 2.173 | <.001 | A |
| | M18 | .761 | 53.260 | <.001 | .642 | A | 54.437 | <.001 | <.035 | A | -.047 | 33.027 | <.001 | A |
| | M19 | 1.206 | 16.830 | <.001 | -.440 | A | 20.574 | <.001 | <.035 | A | .017 | 6.203 | .013 | A |
| | M20 | .793 | 19.784 | <.001 | .545 | A | 24.984 | <.001 | <.035 | A | -.031 | 27.736 | <.001 | A |
| Chemistry | M6 | 1.066 | 2.176 | .140 | -.149 | A | 14.593 | .001 | <.035 | A | .022 | 5.370 | .021 | A |
| | M21 | .814 | 23.915 | <.001 | .484 | A | 35.221 | <.001 | <.035 | A | -.043 | 19.908 | <.001 | A |
| | M22 | 1.006 | .012 | .914 | -.014 | A | .666 | .717 | <.035 | A | .004 | .255 | .614 | A |
| | M23 | 1.136 | 7.513 | .006 | -.300 | A | 7.936 | .019 | <.035 | A | .018 | 4.708 | .030 | A |
| | M24 | 1.038 | .701 | .403 | -.088 | A | 2.655 | .265 | <.035 | A | .009 | 1.045 | .307 | A |
| Physics | M7 | .995 | .011 | .918 | .011 | A | .427 | .808 | <.035 | A | .010 | 1.499 | .221 | A |
| | M8 | .733 | 61.356 | <.001 | .731 | A | 67.089 | <.001 | <.035 | A | -.062 | 57.906 | <.001 | B+ |
| | M9 | .695 | 80.003 | <.001 | .855 | A | 90.701 | <.001 | <.035 | A | -.073 | 83.041 | <.001 | B+ |
| | M10 | .906 | 6.808 | .009 | .232 | A | 9.634 | .008 | <.035 | A | -.002 | .034 | .854 | A |
| | M11 | 1.351 | 59.731 | <.001 | -.707 | A | 69.915 | <.001 | <.035 | A | .072 | 78.148 | <.001 | B- |
| | M25 | 1.160 | 14.258 | <.001 | -.349 | A | 21.779 | <.001 | <.035 | A | .029 | 13.055 | <.001 | A |
| | M26 | 1.187 | 14.025 | <.001 | -.403 | A | 18.518 | <.001 | <.035 | A | .009 | 1.613 | .204 | A |
| | M27 | .928 | 2.768 | .096 | .175 | A | 16.139 | <.001 | <.035 | A | -.022 | 8.622 | .003 | A |
| | M28 | 1.323 | 43.694 | <.001 | -.657 | A | 44.481 | <.001 | <.035 | A | .046 | 36.452 | <.001 | A |
| Earth | M12 | .852 | 15.190 | <.001 | .376 | A | 19.419 | <.001 | <.035 | A | -.013 | 2.103 | .147 | A |
| | M13 | .932 | 2.176 | .140 | .165 | A | 4.815 | .090 | <.035 | A | -.034 | 18.700 | <.001 | A |
| | M14 | 1.081 | 3.039 | .081 | -.182 | A | 11.566 | .003 | <.035 | A | -.008 | 1.011 | .315 | A |
| | M29 | 1.067 | 2.688 | .101 | -.153 | A | 4.258 | .119 | <.035 | A | .046 | 26.647 | <.001 | A |
| | M30 | .997 | .002 | .966 | .006 | A | 12.598 | .002 | <.035 | A | -.019 | 5.042 | .025 | A |
| | M31 | 1.075 | 3.385 | .066 | -.169 | A | 2.336 | .311 | <.035 | A | .035 | 14.760 | <.001 | A |

+/-: DIF favors focal/reference group.

Based on the MH and LR results reported in Table 10, in the TIMSS 2019 8th grade science test consisting of 31 items, all items showed negligible levels of DIF (Level A). However, according to the SIBTEST results, 4 items exhibited DIF at Level B. Item 8 and 9 favored students taking the paper-pencil version, while items 3 and 11 favored students taking the computer-based version in terms of DIF, see Table 10.

## Discussion

In this study, measurement invariance based on the participation format in paper-pencil TIMSS and computer-based eTIMSS mathematics and science achievement tests in TIMSS 2019 is examined, along with whether the items exhibit DIF. The stages of measurement invariance are tested hierarchically. Following the findings from the stages of measurement invariance, DIF analyses are conducted using three different approaches, namely MH, LR, and SIBTEST, to determine the items exhibiting DIF for mathematics and science subtests between paper-pencil and computer-based groups. These analyses also indicate whether DIF favors the focal or reference groups.

The results of the analyses indicate that in TIMSS 2019, at both 4th and 8th grade levels, the stages of measurement invariance, including configural, metric, scalar, and strict invariance, are established for all subtests in mathematics and science based on the ΔCFI and ΔTLI. But $\chi^2$ difference tests indicated lack of invariance, as expected with large sample sizes. The variables in the mathematics and science achievement test models, including item and factor loadings, item intercepts, and error variances, are considered to be invariant across paper-pencil and computer-based groups for all subtests and grade levels, indicating measurement invariance. In other words, the observed differences between paper-pencil and computer-based groups for all subtests seem to stem from genuine ability differences between the groups. Consequently, it can be concluded that the computer-based eTIMSS and paper-pencil TIMSS assessments conducted for the first time in 2019 are comparable across all four subtests. This finding is considered to be particularly significant, and it is suggested that countries participating in the paper-pencil administration should expedite the transition to computer-based assessment procedures once they complete the necessary infrastructure work.

Most of the measurement invariance studies conducted for large-scale exams in the literature involve the hierarchical stages and results reached through MGCFA analyses for variables such as gender, school environment, and achievement vary and their outcomes differ (Arim & Ercikan, 2014; Gündoğmuş, 2017; Wruster, 2022). In line with this research, Wu, Li, and Zumbo (2007) present the results of binary comparisons of 21 countries selected for TIMSS 1999 mathematics and science tests. The results obtained for all tests included in our study are consistent with the conclusion of measurement invariance at the level of strong invariance. Ercikan and Koh (2009) find strong invariance in three out of eight test booklets for TIMSS 2003 cycle science and mathematics tests between Canada-England and France. In contrast, similar uniformity is not observed in the others. In this sense, it can be said that the results are consistent. Similarly, in Akyıldız's (2009) study, the MGCFA comparisons of 35 countries in the PIRLS 2001 achievement tests provide evidence of strong invariance, which is consistent with the results obtained for all tests included in this study. In Eriştiren's (2021) study, the measurement invariance achieved at all stages in the analyses conducted with binary categorical data for the Turkish language achievement test in the LGS 2018, inclusive of 3000 students, is in line with this study.

The MGCFA results at the scale level were also evaluated in terms of DIF at the item level. The results of the analyses conducted with three different methods for item-level analysis and MGCFA at the scale level were compared and evaluated in line with the examples in the literature. The items in the mathematics and science subtests at the 4th and 8th grade levels were analyzed using the MH, LR, and SIBTEST methods, depending on the mode of test administration (paper-pencil/computer-based).

For the 4th grade mathematics subtest, based on the MH method, a total of three items showed DIF at the B level, while the SIBTEST method showed five items with DIF, and the LR method did not reveal any DIF

items. When comparing the MH and SIBTEST methods, three similar items with DIF were found in both methods, and two items showed DIF in the SIBTEST method but not in the MH method. Among the three DIF items identified in both the MH and SIBTEST methods, two items favored students taking the paper-pencil test (focus group), and two items favored students taking the computer-based test (reference group). These findings support Yörü and Atar's (2019) recommendation to use at least two methods to identify DIF, as the results obtained from the three DIF methods in the 4th grade mathematics test were qualitatively different. Additionally, in the study by Eriştiren (2021), it was observed that MH and SIBTEST techniques showed consistency, but LR method did not exhibit the same level of consistency, which aligns with the current study's results.

Regarding the 8th grade mathematics subtest, based on the MH method, seven items showed DIF, while the LR method did not reveal any DIF items, and the SIBTEST method showed 11 items with DIF. Among the DIF items in the SIBTEST method, four items were not present in the MH method. Four items among the DIF items in both the MH and SIBTEST methods favored the focus group, and three items favored the reference group. However, of the four other items marked DIF by SIBTEST, two favor focal and two favor reference group.

In the 4th grade science subtest, the MH method revealed one item with DIF, the LR method showed no DIF items, and the SIBTEST method showed six items with DIF. Among the DIF items, item 4 showed DIF only in the MH method and favored the focal group at the B level. The SIBTEST method flagged three items favor focal and the rest favor reference group. These results align with previous studies by Gök, Kelecioğlu, and Doğan (2010) and Ercikan and Koch (2009), indicating a low level of agreement between the MH and LR methods for DIF detection. Furthermore, similar findings were observed between this study and Eriştiren's (2021) study on measurement invariance using the results from the entrance exam for secondary education.

When examining the DIF results of the 8th grade science subtest, no items showed DIF in the MH and LR methods, while four items exhibited DIF in the SIBTEST method. Among the DIF items identified in the SIBTEST method, two favored the focal group, and two favored the reference group. However, the SIBTEST method revealed DIF in four items, indicating its lack of alignment with the other two methods. Overall, the DIF analyses conducted in this study suggest that using multiple methods, such as MH, LR, and SIBTEST, can enhance the accuracy of identifying DIF in educational assessments.

In terms of the DIF analyses conducted using the MH and SIBTEST techniques showed some agreement, for the disagreements SIBTEST flagged more items than the MH method. However, the LR approach did not agree with SIBTEST and MH, and did not flag any B or C level DIF in our analysis. In other words, no set of items was consistently advantageous or disadvantageous to either the reference or focus group across all subtest results based on the LR approach.

Overall, the MGCFA conclusions based on the ΔCFI and ΔTLI are in agreement with the LR approach, and they provide evidence for the measurement invariance. The MGCFA conclusions based on the $\chi^2$ difference tests are in agreement with the SIBTEST and MH conclusions and they can arguably be considered as concerns about the invariance. These findings are inconsistent with some literature (Çepni, 2011; Wiberg, 2009) while being consistent with others (Doğan, 2008; Gök, 2010). Similarly, Eriştiren's (2021) study on measurement invariance and DIF in entrance exams to secondary education also presents similar findings to this study. While measurement invariance was largely achieved across all stages in the tests, discrepancies in DIF were observed, particularly concerning achievement levels based on school type, where the MH and SIBTEST analyses showed converging results, but the LR method exhibited incongruent results. Additionally, the discrepancies observed in the results of the study by Özdemir (2003) comparing two-category and partial credit scoring methods for multiple-choice items in a Turkish reading comprehension test support the outcomes of this study.

It should be noted that MGCFA analyses took into account the factor structure while the DIF analyses were conducted separately for each dimension. Despite our efforts to conduct multidimensional DIF our attempts to utilize R was unsuccessful probably due to the large sample size and relatively complex factor structure.

Our final attempt was to run DIF analyses for the entire test, assuming unidimensionality; with this assumption the number of flagged items were less compared to what we reported in this paper. To be on the conservative side, we reported the DIF analyses that conducted separately for each dimension. Future studies are needed to address this limitation.

## Declarations

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

**Ethical Approval:** I declare that all ethical guidelines for the author have been followed. This study does not require any ethics committee approval as it includes open-access data.

### References

Agresti, A. (1984). *Analysis of ordinal categorical data*. New York: John Wiley & Sons.

Akyıldız, M. (2009). Pırls 2001 testinin yapı geçerliliğinin ülkelerarası karşılaştırılması. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi*, *6*(1)

Anakwe, B. (2008). Comparison of student performance in paper-based versus ccomputer-based testing. *Journal of Education for Business.* September-October, 13-17.

Arım, G, R., Ercikan, K. (2014). Comparability between the American and Turkish versions of the TIMSS mathematics test results*. Eğitim ve Bilim*. *39*(172), 33- 48.

Atılgan, H., Kan, A., Aydın, B. (2017). *Eğitimde ölçme ve değerlendirme.* Onuncu Baskı. Ankara: Anı Yayıncılık.

Bağdu Söyler, P., Aydın, B., & Atılgan, H. (2021). PISA 2015 Reading Test Item Parameters Across Language Groups: A measurement Invariance Study with Binary Variables. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 112-128. https://doi.org/10.21031/epod.800697

Büyüköztürk, Ş. (2010). *Sosyal bilimler için veri analizi el kitabı*. Pegem Akademi:Ankara.

Büyüköztürk, Ş., Çokluk, Ö., & Şekercioğlu, G. (2014). Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL uygulamaları. Ankara: Pegem Akademi.

Camilli G. Shepard L. A. (1994). *Methods for Identifying Biased Test Items. Volume 4.* California: SAGE Publications. Inc.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of statistical Software*, *48*, 1-29.

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

Çepni, Z. (2011). *Değişen madde fonksiyonlarının SIBTEST, Mantel-Haenzsel,lojistik regresyon ve madde tepki kuramı yöntemleriyle incelenmesi* (doktora tezi). Hacettepe Üniversitesi, Ankara

Doğan, N; Öğretmen, T. (2008). *Değişen madde fonksiyonunu belirlemede Mantel - Haenszel, ki-kare ve lojistik regresyon tekniklerinin karşılaştırılması.* Eğitim ve Bilim Dergisi. 33(148).

Dorans, N. J., & Holland, P. W. (1993). *DIF detection and description: Mantel-Haenszel and standardization.* In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 35–66). Lawrence Erlbaum Associates, Inc.

Drasgow, F (2002). The work ahead: a psychometric infrastructure for computerized adaptive tests. In C.N. Mills, M.T. Potenza, J.J. Fremer, & W.C. Ward (Eds.), Computer-based testing: Building the foundation for future assessments (pp. 67–88). Hillsdale, NJ: Lawrence Erlbaum.

Ercikan, K; Koh, K. (2009). *Examining the construct comparability of the English and French versions of TIMSS.* Internatıonal Journal Of Testıng, 5(1), 23–35

Ergün, E. (2002). *Üniversite öğrencilerinin bilgisayar destekli ölçmeden elde ettikleri eaşarının kalem-kâğıt testi başarısı, bilgisayar kaygısı ve bilgisayar tecrübeleri açısından* i*ncelenmesi.* Yayımlanmamış yüksek lisans tezi. Anadolu Üniversitesi Eğitim Bilimleri Enstitüsü, Eskişehir.

Eriştiren, İ. (2021). *Ortaöğretime Geçiş Sınavlarında ölçme değişmezliği ve DIF'nin incelenmesi* (Yüksek Lisans Tezi). Haccettepe Üniversitesi, Ankara.

Gök, B., Kelecioğlu, H. ve Doğan, N. (2010). Değişen madde fonksiyonunu belirlemede Mantel- Haenzsel ve lojistik regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim*, *35*, 3-16.

Gierl, M. J., Jodoin, M. G., & Ackerman, T. A. (2000). American Educational Research Association (AERA) New Orleans, Louisiana, USA April 24-27, 2000.

Gündoğmuş, İ. (2017). *Kâğıt-kalem, bilgisayar ve tablet ortamında gerçekleştirilen sınavlar için ölçme değişmezliğinin ve öğrenci görüşlerinin incelenmesi*. Hacettepe Üniversitesi, Ankara

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups?: Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care, 44*(1),

Hair, J.F. Jr., Anderson, R.E., Tatham, R.L., and Black, W.C. (1998). *Multivariate data analysis*, (5th Edition). Upper Saddle River, NJ: Prentice Hall.

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. Medical Care, 44, 182-188.

İlci, B. (2004). *Geleneksel kâğıt-kalem yöntemi ile ve bilgisayarda online uygulanan çoktan seçmeli sayısal yetenek ve sözel yetenek testlerine ait madde ve test* i*statistiklerinin karşılaştırılması.* Yüksek lisans tezi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Jöreskog, K. G. ve Sörbom, D. (2006). LISREL (Version 8.8) [computer software]. Chicago: Scientific Softare International Inc.

Kite, B. A., Johnson, P. E., & Xing, C. (2018, January 28). Replicating the Mplus DIFFTEST Procedure. https://pj.freefaculty.org/guides/crmda_guides/44.difftest/44.difftest.html

Klieme E., Baumert J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of Education, 16:3, 385-402.*

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.b.). New York & London: The Guilford Press.

Li, Z., Gooden, C. J., & Toland, M. D. (2016). Measurement invariance with categorical indicators. *Applied Psychometric Strategies Lab, Applied Quantitative and Psychometric Series. Presentation conducted at the University of Kentucky, Lexington, KY. Retrieved from https://education. uky. edu/edp/apslab/events*.

MEB. (2020). *TIMSS 2019 ulusal matematik ve fen bilimleri ön raporu: 4. ve 8. sınıflar.* Ankara.

Meredith, W., & Teresi, J. A. (2006). *An essay on measurement and factorial invariance.* Medical care, 44(11), 69-S77.

Mertler, C. A. & Vannatta, R. A. (2005). *Advanced and multivariate statistical methods: Practical application and interpretation* (3rd ed.). Los Angeles: Pyrczak.

Mills, C. N., Potenza, M.T., Fremer, J.J., Ward, W.C. (2001). *Computer Based Testing: Building the Foundation for Future Assessment.* Lawrance Erlbaum Associates, Publishers: Londra

Moraes, C.L & Reichenheim, M.E. (2002). Cross-cultural measurement equivalence of the revised conflict tactics scales (cts2) portuguese version used to identify violence within couples. Cad. Saúde Pública, 18 (3).

Muthén, B. O. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 205–243). Newbury Park, CA: Sage.

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D., & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science*. Boston College, TIMSS & PIRLS International Study Center.

Nandakumar, R. (1993). A fortran 77 program for detecting differential item functioning through the mantel-haenszel statistic. *Educational and Psychological Measurement, 53,* 679–684.

Osterlind S. J. Everson H. T. (2009). *Differential Item Functioning: Second Edition.* California: SAGE Publications. Inc.

Özdemir, D. (2003). Çoktan seçmeli testlerde iki kategorili ve önsel ağırlıklı puanlamanın değişen madde fonksiyonuna etkisi ile ilgili bir araştırma. *Eğitim ve Bilim, 28(*129), 37-43.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517-529.

Raykov, T., Dimitrov, D. M., Marcoulides, G. A., Li, T., & Menold, N. (2018). Examining measurement invariance and differential item functioning with discrete latent construct indicators: A note on a multiple testing procedure. *Educational and Psychological Measurement*, *78*(2), 343-352.

Rogers, T. B. (1995). *The psychological testing enterprise: An introduction.* Pasific Grove, California: Brooks/Cole.

Russel, M., Goldberg, A., O'Connor, K. (2003). Computer based test and validity: A look back into the future. *Assessment in Education.* 10, 279- 293.

Shealy, R. and Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detect test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194. doi: 10.1007/BF02294572

Steenkamp, B., E., M. and Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. journal of consumer research, 25(1),78-107.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics*. Boston: Pearson Education.

Wiberg, M. (2009). Differential item functioning in mastery tests: A comparison of three methods using real data. International *Journal of Testing, 9,* 41–59

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

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

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF) logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Canada: Ottowa, Directorate of Human Resources Research and Evaluation National Defense Headquarters: Author.

# Appendix

## Appendix 1. 4th Grade Science VIF Analysis Results

| ITEMS | Tolerance | VIF |
|---|---|---|
| NEW KIND OF MAMMAL DISCOVERED (A) | 0,810 | 1,234 |
| COVER YOUR MOUTH THOUGH NOT SICK (1) | 0,862 | 1,160 |
| HAMAD'S GARDEN: WHICH SURVIVE (1) | 0,873 | 1,146 |
| HAMAD'S GARDEN: PLANT STRUCTURE (1) | 0,915 | 1,092 |
| TWO THINGS ANIMALS NEED (1) | 0,893 | 1,120 |
| CELERY STALK LEAVES TURN RED (B) | 0,821 | 1,217 |
| WOODEN AND METAL CUBES ON BALANCE (B) | 0,902 | 1,109 |
| TWO METAL BARS (C) | 0,858 | 1,166 |
| DROPS OF WAX ON A METAL FRAME (1) | 0,722 | 1,385 |
| OBJECT INSIDE A WOODEN BOX (C) | 0,897 | 1,115 |
| AMOUNT OF WATER AND LAND ON EARTH (D) | 0,892 | 1,121 |
| WHAT MAKES UP SOLAR SYSTEM (C) | 0,809 | 1,236 |
| LIVING AND NON-LIVING THINGS IN A DESERT (1) | 0,863 | 1,159 |
| HUMAN ORGAN WITH SAME FUNCTION AS GILLS (B) | 0,789 | 1,267 |
| CHARACTERISTICS OF LIVING AND TOY DUCK (DERIVED) (1) | 0,811 | 1,233 |
| EXPLAIN DECREASE IN INSECT POPULATION (1) | 0,727 | 1,376 |
| WHAT MAKES VENUS FLYTRAP DIFFERENT FROM OTHER PLANTS (B) | 0,904 | 1,107 |
| WHY GROUND SQUIRREL HOLDS TAIL OVERHEAD (1) | 0,763 | 1,311 |
| CHANGE WHERE MATERIALS IN OBJECTS STAY THE SAME (A) | 0,911 | 1,098 |
| CAUSE OF SKYDIVER'S FALL (C) | 0,822 | 1,217 |
| ENERGY CHANGE IN A FLASHLIGHT (A) | 0,889 | 1,125 |
| WHY MARY'S BOX IS EASIER TO MOVE (D) | 0,817 | 1,225 |
| ADVANTAGES TO FARMING NEAR A RIVER (1) | 0,843 | 1,186 |
| DISADVANTAGES TO FARMING NEAR A RIVER (1) | 0,809 | 1,236 |
| POSITION OF THE EARTH WHEN IT IS SUMMER IN CITY A (C) | 0,920 | 1,087 |

## Appendix 2. 4th Grade Mathematics VIF Analysis Results

| ITEMS | Tolerance | VIF |
|---|---|---|
| NUMBERS WITH 6 AS A FACTOR (DERIVED) (1) | 0,898 | 1,114 |
| FIGURE WITH THREE QUARTERS SHADED (A) | 0,856 | 1,168 |
| WHO PAID LESS FOR EACH BOTTLE (1) | 0,756 | 1,323 |
| FRACTION WATERED ON MONDAY (1) | 0,404 | 2,475 |
| FRACTION WATERED ON TUESDAY (1) | 0,373 | 2,682 |
| NEXT 2 NUMBERS IN THE PATTERN (DERIVED) (1) | 0,686 | 1,458 |
| STREET PARALLEL TO GREEN STREET (A) | 0,839 | 1,192 |
| PERPENDICULAR TO APPLE STREET (B) | 0,940 | 1,064 |
| NUMBER OF TRIANGLES NEEDED (B) | 0,908 | 1,101 |
| SHAPE THAT FOLDS INTO A BOX (D) | 0,940 | 1,064 |
| MOST FREQUENT SCORE ON QUIZ (1) | 0,818 | 1,223 |
| SCORE OF 4 OR MORE ON QUIZ (1) | 0,728 | 1,374 |
| NUMBER WITH 7 HUNDREDS AND 6 ONES (C) | 0,876 | 1,141 |
| DISTANCE TRAVELED EACH DAY ON BICYCLE (B) | 0,756 | 1,323 |
| FRACTIONS GREATER THAN 1/2 (DERIVED) (1) | 0,726 | 1,378 |
| EXPRESSION FOR STICKERS GIVEN TO EACH FRIEND (D) | 0,745 | 1,343 |
| COST BANANAS AND PLUMS (DERIVED) (2) | 0,828 | 1,208 |
| UNITS FOR MEASUREMENTS (DERIVED) (1) | 0,882 | 1,134 |
| WEIGHT OF 1 PEAR (C) | 0,807 | 1,240 |
| NUMBER OF SHAPES TO COVER SQUARE (DERIVED) (2) | 0,763 | 1,311 |
| COMPLETE FIGURE WITH LINE OF SYMMETRY (1) | 0,867 | 1,154 |
| WATER LEVEL IN DAM - WEEK 8 (1) | 0,811 | 1,233 |
| PICTOGRAPH OF ANIMAL WEIGHTS (DERIVED) (1) | 0,738 | 1,355 |
| BAR GRAPH OF CARS EACH MORNING (DERIVED) (1) | 0,669 | 1,495 |

## Appendix 3. 8th Grade Science VIF Analysis Results

| ITEMS | Tolerance | VIF |
|---|---|---|
| PENGUIN BEHAVIOR AND SURVIVAL (2) | 0,859 | 1,164 |
| ORGANISM WITH CELL WALLS (C) | 0,898 | 1,114 |
| HOW DECOMPOSERS GET ENERGY (B) | 0,821 | 1,217 |
| ORGANISM THAT COMPETES WITH HUMANS (1) | 0,760 | 1,317 |
| GARDEN WITH BIRD FEEDER (DERIVED) (1) | 0,869 | 1,151 |
| WHY SOLUTION 2 IS PALER THAN 1 (1) | 0,796 | 1,256 |
| WHICH IS A PHYSICAL CHANGE (D) | 0,896 | 1,116 |
| MODEL FLASHLIGHT: BULB WON'T LIGHT (1) | 0,840 | 1,190 |
| MODEL FLASHLIGHT: 2 PARALLEL BULBS (1) | 0,814 | 1,229 |
| MODEL FLASHLIGHTS: COMPARISON (C) | 0,923 | 1,083 |
| TWO BAR MAGNETS REPELLING (A) | 0,818 | 1,223 |
| PLANETS: SHORTEST DAY LENGTH (D) | 0,887 | 1,128 |
| PLANETS: DISTANCE FROM SUN (1) | 0,759 | 1,318 |
| TEMPERATURE OUTSIDE AN AIRPLANE (A) | 0,769 | 1,300 |
| RELATIONSHIP BETWEEN INSECTS AND FLOWERING PLANTS (D) | 0,827 | 1,210 |
| WHERE IN A CELL DNA REPLICATION OCCURS (B) | 0,902 | 1,108 |
| INCREASE GREEN SPACE AS CARBON DIOXIDE INCREASES (1) | 0,689 | 1,451 |
| WHY LEAVES' MASSES DECREASED (C) | 0,901 | 1,110 |
| CLASSIFY ANIMALS BASED ON A SINGLE CHARACTERISTIC (1) | 0,762 | 1,312 |
| IDENTIFY THE CHARACTERISTIC USED TO CLASSIFY ANIMALS (1) | 0,863 | 1,158 |
| LOCATION OF SUBATOMIC PARTICLES (1) | 0,831 | 1,203 |
| ORDER ELEMENTS FROM SMALLEST TO LARGEST ATOMIC NUM (1) | 0,804 | 1,244 |
| ACIDIC, BASIC, OR NEUTRAL SOLUTION (DERIVED) (1) | 0,814 | 1,229 |
| MIXING AN ACID AND BASE SOLUTION (D) | 0,837 | 1,195 |
| GAS MOLECULES IN AN EXPANDING BALLOON (A) | 0,850 | 1,177 |
| THINGS TOM SHOULD DO (DERIVED) (1) | 0,612 | 1,633 |
| VEHICLE WITH DIFFERENT WEIGHTS ON DIFFERENT PLANETS (D) | 0,747 | 1,338 |
| CELL PHONE IN A VACUUM (1) | 0,743 | 1,346 |
| WHY BALLOON GETS BIGGER AS IT RISES (B) | 0,923 | 1,083 |
| EVIDENCE OF GLOBAL WARMING (A) | 0,749 | 1,335 |
| NATURAL RESOURCE FORMATION SHOWN IN DIAGRAMS (C) | 0,866 | 1,154 |

## Appendix 4. 8th Grade Mathematics VIF Analysis Results

| ITEMS | Tolerance | VIF |
|---|---|---|
| OCTAGON WITH EQUIVALENT SHADING (B) | 0,740 | 1,352 |
| TIME WHEN PAT FINISHES LAST LAP (1) | 0,677 | 1,476 |
| PERCENTAGE OF LAPS FINISHED (1) | 0,633 | 1,581 |
| MULTIPLES OF 3 (D) | 0,745 | 1,342 |
| CONVERT DECIMAL TO A FRACTION (1) | 0,725 | 1,378 |
| EXPRESSION FOR AREA OF RECTANGLE (C) | 0,738 | 1,355 |
| EXPRESSION WITH EXPONENTS OF Y (B) | 0,725 | 1,380 |
| NUMBER OF MATCHES FOR FIGURE 10 (1) | 0,768 | 1,303 |
| RULE FOR NUMBER OF MATCHES (1) | 0,652 | 1,534 |
| GRAPH OF Y = 2X (A) | 0,884 | 1,132 |
| ROTATION AND REFLECTION (D) | 0,921 | 1,086 |
| SURFACE AREA OF THE PRISM (C) | 0,805 | 1,242 |
| VALUE OF ANGLE X OUTSIDE TRIANGLE (C) | 0,740 | 1,351 |
| NUMBER OF BALLS IN A BAG (B) | 0,753 | 1,327 |
| LIV'S SMARTPHONE USE (D) | 0,720 | 1,389 |
| SMARTPHONE USE LISTENING TO MUSIC (A) | 0,769 | 1,300 |
| STATEMENTS FOR ALL VALUES OF INTEGER A (DERIVED) (2) | 0,752 | 1,329 |
| ARROW TO SHOW 5/12 ON NUMBER LINE (B) | 0,743 | 1,345 |
| VALUE OF FRACTION X IN SQUARE (1) | 0,681 | 1,469 |
| NUMBER OF BLUE BEADS ON BRACELET (1) | 0,762 | 1,312 |
| VALUE OF 2(6X - 3Y) WHEN X = 3 AND Y = 2 (C) | 0,752 | 1,329 |
| EXPRESSION EQUIVALENT TO 2Y + 6XY2 (A) | 0,761 | 1,315 |
| FORMULA FOR STOPPING DISTANCE (1) | 0,624 | 1,601 |
| VALUE OF X GIVEN PERIMETER OF TRIANGLE ABC (1) | 0,542 | 1,844 |
| ADDITIONAL POINT ON A STRAIGHT LINE (D) | 0,776 | 1,288 |
| VALUE OF ANGLE X IN A QUADRILATERAL (1) | 0,634 | 1,578 |
| METHODS OF FOLDING PAPER (DERIVED) (1) | 0,846 | 1,182 |
| COORDINATES TO COMPLETE KLMN (DERIVED) (1) | 0,623 | 1,606 |
| MEAN TEMPERATURE FOR 5 DAYS (1) | 0,587 | 1,704 |
| BEST GRAPH FOR TOWN INFORMATION (DERIVED) (1) | 0,774 | 1,292 |
| BAR GRAPH OF NEWSPAPER SALES (1) | 0,764 | 1,309 |

## Appendix 5. 4th Grade Science Tetrachoric Correlation Analysis Results

| | M1 | M2 | M3 | M4 | M5 | M6 | M7 | M8 | M9 | M10 | M11 | M12 | M13 | M14 | M15 | M16 | M17 | M18 | M19 | M20 | M21 | M22 | M23 | M24 | M25 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| M1 | 1 | | | | | | | | | | | | | | | | | | | | | | | | |
| M2 | 0.34 | 1 | | | | | | | | | | | | | | | | | | | | | | | |
| M3 | 0.35 | 0.31 | 1 | | | | | | | | | | | | | | | | | | | | | | |
| M4 | 0.3 | 0.23 | 0.32 | 1 | | | | | | | | | | | | | | | | | | | | | |
| M5 | 0.26 | 0.25 | 0.24 | 0.25 | 1 | | | | | | | | | | | | | | | | | | | | |
| M6 | 0.36 | 0.34 | 0.32 | 0.27 | 0.27 | 1 | | | | | | | | | | | | | | | | | | | |
| M7 | 0.22 | 0.2 | 0.24 | 0.21 | 0.18 | 0.24 | 1 | | | | | | | | | | | | | | | | | | |
| M8 | 0.29 | 0.25 | 0.26 | 0.23 | 0.21 | 0.31 | 0.25 | 1 | | | | | | | | | | | | | | | | | |
| M9 | 0.44 | 0.4 | 0.41 | 0.36 | 0.3 | 0.43 | 0.33 | 0.36 | 1 | | | | | | | | | | | | | | | | |
| M10 | 0.26 | 0.2 | 0.23 | 0.22 | 0.2 | 0.27 | 0.21 | 0.27 | 0.32 | 1 | | | | | | | | | | | | | | | |
| M11 | 0.24 | 0.19 | 0.21 | 0.23 | 0.2 | 0.21 | 0.21 | 0.24 | 0.29 | 0.19 | 1 | | | | | | | | | | | | | | |
| M12 | 0.34 | 0.26 | 0.31 | 0.27 | 0.22 | 0.34 | 0.26 | 0.31 | 0.43 | 0.25 | 0.32 | 1 | | | | | | | | | | | | | |
| M13 | 0.27 | 0.27 | 0.29 | 0.29 | 0.29 | 0.27 | 0.19 | 0.24 | 0.34 | 0.21 | 0.25 | 0.29 | 1 | | | | | | | | | | | | |
| M14 | 0.4 | 0.29 | 0.33 | 0.29 | 0.28 | 0.39 | 0.25 | 0.3 | 0.44 | 0.28 | 0.27 | 0.4 | 0.31 | 1 | | | | | | | | | | | |
| M15 | 0.38 | 0.27 | 0.29 | 0.28 | 0.25 | 0.34 | 0.26 | 0.28 | 0.42 | 0.26 | 0.24 | 0.36 | 0.29 | 0.42 | 1 | | | | | | | | | | |
| M16 | 0.47 | 0.33 | 0.4 | 0.36 | 0.29 | 0.41 | 0.27 | 0.36 | 0.49 | 0.3 | 0.3 | 0.41 | 0.34 | 0.45 | 0.39 | 1 | | | | | | | | | |
| M17 | 0.26 | 0.19 | 0.24 | 0.2 | 0.17 | 0.23 | 0.2 | 0.21 | 0.29 | 0.16 | 0.17 | 0.24 | 0.2 | 0.27 | 0.27 | 0.3 | 1 | | | | | | | | |
| M18 | 0.4 | 0.33 | 0.36 | 0.33 | 0.3 | 0.37 | 0.23 | 0.31 | 0.45 | 0.29 | 0.28 | 0.36 | 0.3 | 0.41 | 0.36 | 0.5 | 0.29 | 1 | | | | | | | |
| M19 | 0.24 | 0.2 | 0.21 | 0.22 | 0.16 | 0.22 | 0.18 | 0.21 | 0.28 | 0.16 | 0.21 | 0.23 | 0.18 | 0.23 | 0.25 | 0.28 | 0.16 | 0.24 | 1 | | | | | | |
| M20 | 0.32 | 0.3 | 0.3 | 0.21 | 0.23 | 0.35 | 0.2 | 0.27 | 0.38 | 0.26 | 0.25 | 0.38 | 0.27 | 0.4 | 0.35 | 0.39 | 0.2 | 0.37 | 0.21 | 1 | | | | | |
| M21 | 0.3 | 0.18 | 0.21 | 0.2 | 0.19 | 0.25 | 0.18 | 0.23 | 0.31 | 0.21 | 0.21 | 0.24 | 0.17 | 0.29 | 0.28 | 0.31 | 0.18 | 0.3 | 0.25 | 0.26 | 1 | | | | |
| M22 | 0.31 | 0.31 | 0.32 | 0.25 | 0.25 | 0.35 | 0.24 | 0.32 | 0.42 | 0.27 | 0.26 | 0.34 | 0.27 | 0.37 | 0.35 | 0.41 | 0.25 | 0.39 | 0.25 | 0.36 | 0.3 | 1 | | | |
| M23 | 0.29 | 0.28 | 0.28 | 0.28 | 0.22 | 0.26 | 0.17 | 0.24 | 0.34 | 0.2 | 0.19 | 0.27 | 0.29 | 0.31 | 0.28 | 0.36 | 0.21 | 0.35 | 0.18 | 0.29 | 0.22 | 0.3 | 1 | | |
| M24 | 0.33 | 0.28 | 0.33 | 0.29 | 0.24 | 0.32 | 0.26 | 0.29 | 0.4 | 0.28 | 0.24 | 0.3 | 0.29 | 0.37 | 0.31 | 0.41 | 0.25 | 0.4 | 0.24 | 0.34 | 0.27 | 0.35 | 0.47 | 1 | |
| M25 | 0.22 | 0.17 | 0.2 | 0.17 | 0.13 | 0.22 | 0.19 | 0.2 | 0.27 | 0.18 | 0.19 | 0.27 | 0.17 | 0.23 | 0.23 | 0.27 | 0.19 | 0.25 | 0.14 | 0.25 | 0.16 | 0.24 | 0.15 | 0.23 | 1 |

## Appendix 6. 4th Grade Mathematics Tetrachoric Correlation Analysis Results

| | M1 | M2 | M3 | M4 | M5 | M6 | M7 | M8 | M9 | M10 | M11 | M12 | M13 | M14 | M15 | M16 | M17 | M18 | M19 | M20 | M21 | M22 | M23 | M24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| M1 | 1 | | | | | | | | | | | | | | | | | | | | | | | |
| M2 | 0.1 | 1 | | | | | | | | | | | | | | | | | | | | | | |
| M3 | 0.33 | 0.29 | 1 | | | | | | | | | | | | | | | | | | | | | |
| M4 | 0.19 | 0.49 | 0.43 | 1 | | | | | | | | | | | | | | | | | | | | |
| M5 | 0.2 | 0.54 | 0.46 | 0.93 | 1 | | | | | | | | | | | | | | | | | | | |
| M6 | 0.32 | 0.37 | 0.54 | 0.5 | 0.53 | 1 | | | | | | | | | | | | | | | | | | |
| M7 | 0.17 | 0.29 | 0.33 | 0.38 | 0.43 | 0.36 | 1 | | | | | | | | | | | | | | | | | |
| M8 | 0.14 | 0.17 | 0.26 | 0.25 | 0.27 | 0.26 | 0.21 | 1 | | | | | | | | | | | | | | | | |
| M9 | 0.14 | 0.24 | 0.26 | 0.32 | 0.35 | 0.32 | 0.24 | 0.15 | 1 | | | | | | | | | | | | | | | |
| M10 | 0.13 | 0.16 | 0.24 | 0.22 | 0.23 | 0.25 | 0.15 | 0.12 | 0.17 | 1 | | | | | | | | | | | | | | |
| M11 | 0.17 | 0.26 | 0.35 | 0.37 | 0.4 | 0.39 | 0.28 | 0.17 | 0.26 | 0.17 | 1 | | | | | | | | | | | | | |
| M12 | 0.23 | 0.35 | 0.46 | 0.47 | 0.52 | 0.5 | 0.4 | 0.24 | 0.31 | 0.2 | 0.69 | 1 | | | | | | | | | | | | |
| M13 | 0.25 | 0.29 | 0.32 | 0.36 | 0.38 | 0.41 | 0.24 | 0.2 | 0.27 | 0.17 | 0.28 | 0.34 | 1 | | | | | | | | | | | |
| M14 | 0.31 | 0.23 | 0.55 | 0.39 | 0.4 | 0.5 | 0.31 | 0.25 | 0.25 | 0.17 | 0.31 | 0.41 | 0.33 | 1 | | | | | | | | | | |
| M15 | 0.24 | 0.47 | 0.5 | 0.55 | 0.62 | 0.52 | 0.45 | 0.26 | 0.3 | 0.21 | 0.37 | 0.5 | 0.35 | 0.43 | 1 | | | | | | | | | |
| M16 | 0.3 | 0.31 | 0.54 | 0.46 | 0.5 | 0.51 | 0.35 | 0.24 | 0.29 | 0.2 | 0.33 | 0.46 | 0.39 | 0.54 | 0.46 | 1 | | | | | | | | |
| M17 | 0.28 | 0.28 | 0.5 | 0.4 | 0.42 | 0.48 | 0.32 | 0.25 | 0.25 | 0.23 | 0.31 | 0.38 | 0.31 | 0.44 | 0.5 | 0.45 | 1 | | | | | | | |
| M18 | 0.26 | 0.16 | 0.31 | 0.25 | 0.27 | 0.32 | 0.25 | 0.14 | 0.19 | 0.12 | 0.2 | 0.28 | 0.21 | 0.33 | 0.27 | 0.35 | 0.29 | 1 | | | | | | |
| M19 | 0.23 | 0.27 | 0.43 | 0.37 | 0.41 | 0.44 | 0.29 | 0.2 | 0.24 | 0.21 | 0.29 | 0.38 | 0.29 | 0.39 | 0.44 | 0.4 | 0.43 | 0.29 | 1 | | | | | |
| M20 | 0.28 | 0.33 | 0.47 | 0.45 | 0.49 | 0.5 | 0.38 | 0.25 | 0.3 | 0.29 | 0.33 | 0.44 | 0.35 | 0.43 | 0.48 | 0.45 | 0.47 | 0.35 | 0.44 | 1 | | | | |
| M21 | 0.22 | 0.25 | 0.34 | 0.34 | 0.39 | 0.38 | 0.24 | 0.19 | 0.26 | 0.23 | 0.26 | 0.29 | 0.27 | 0.26 | 0.39 | 0.34 | 0.34 | 0.24 | 0.28 | 0.38 | 1 | | | |
| M22 | 0.22 | 0.34 | 0.37 | 0.45 | 0.48 | 0.44 | 0.27 | 0.23 | 0.31 | 0.17 | 0.37 | 0.42 | 0.34 | 0.36 | 0.36 | 0.45 | 0.32 | 0.3 | 0.34 | 0.38 | 0.36 | 1 | | |
| M23 | 0.32 | 0.33 | 0.49 | 0.44 | 0.48 | 0.52 | 0.34 | 0.22 | 0.29 | 0.22 | 0.34 | 0.46 | 0.38 | 0.45 | 0.5 | 0.49 | 0.45 | 0.35 | 0.43 | 0.48 | 0.4 | 0.47 | 1 | |
| M24 | 0.25 | 0.43 | 0.51 | 0.53 | 0.57 | 0.55 | 0.42 | 0.28 | 0.34 | 0.27 | 0.4 | 0.52 | 0.37 | 0.43 | 0.57 | 0.48 | 0.49 | 0.31 | 0.46 | 0.55 | 0.42 | 0.52 | 0.56 1.00 | 1 |

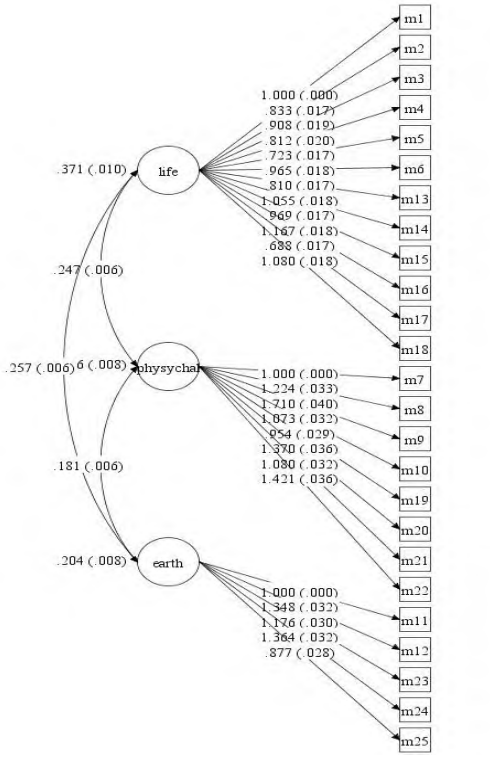
## Appendix 7. 8th Grade Science Tetrachoric Correlation Analysis Results

| | M1 | M2 | M3 | M4 | M5 | M6 | M7 | M8 | M9 | M10 | M11 | M12 | M13 | M14 | M15 | M16 | M17 | M18 | M19 | M20 | M21 | M22 | M23 | M24 | M25 | M26 | M27 | M28 | M29 | M30 | M31 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| M1 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| M2 | 0.21 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| M3 | 0.3 | 0.26 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| M4 | 0.33 | 0.25 | 0.38 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| M5 | 0.25 | 0.14 | 0.25 | 0.34 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| M6 | 0.36 | 0.2 | 0.29 | 0.37 | 0.31 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | |
| M7 | 0.21 | 0.22 | 0.26 | 0.31 | 0.19 | 0.22 | 1 | | | | | | | | | | | | | | | | | | | | | | | | |
| M8 | 0.25 | 0.18 | 0.27 | 0.34 | 0.24 | 0.34 | 0.21 | 1 | | | | | | | | | | | | | | | | | | | | | | | |
| M9 | 0.26 | 0.24 | 0.29 | 0.38 | 0.26 | 0.36 | 0.23 | 0.39 | 1 | | | | | | | | | | | | | | | | | | | | | | |
| M10 | 0.2 | 0.14 | 0.19 | 0.21 | 0.18 | 0.24 | 0.14 | 0.2 | 0.2 | 1 | | | | | | | | | | | | | | | | | | | | | |
| M11 | 0.25 | 0.18 | 0.31 | 0.38 | 0.25 | 0.32 | 0.23 | 0.28 | 0.31 | 0.2 | 1 | | | | | | | | | | | | | | | | | | | | |
| M12 | 0.23 | 0.2 | 0.26 | 0.25 | 0.22 | 0.27 | 0.17 | 0.2 | 0.23 | 0.16 | 0.27 | 1 | | | | | | | | | | | | | | | | | | | |
| M13 | 0.36 | 0.24 | 0.36 | 0.43 | 0.29 | 0.41 | 0.26 | 0.34 | 0.38 | 0.27 | 0.38 | 0.32 | 1 | | | | | | | | | | | | | | | | | | |
| M14 | 0.34 | 0.19 | 0.34 | 0.4 | 0.33 | 0.42 | 0.23 | 0.27 | 0.3 | 0.24 | 0.35 | 0.32 | 0.44 | 1 | | | | | | | | | | | | | | | | | |
| M15 | 0.25 | 0.13 | 0.28 | 0.41 | 0.29 | 0.28 | 0.23 | 0.26 | 0.25 | 0.19 | 0.34 | 0.17 | 0.35 | 0.36 | 1 | | | | | | | | | | | | | | | | |
| M16 | 0.18 | 0.21 | 0.23 | 0.28 | 0.16 | 0.2 | 0.22 | 0.18 | 0.21 | 0.14 | 0.23 | 0.15 | 0.25 | 0.21 | 0.23 | 1 | | | | | | | | | | | | | | | |
| M17 | 0.37 | 0.26 | 0.41 | 0.49 | 0.33 | 0.45 | 0.31 | 0.41 | 0.42 | 0.24 | 0.39 | 0.3 | 0.48 | 0.45 | 0.36 | 0.31 | 1 | | | | | | | | | | | | | | |
| M18 | 0.18 | 0.13 | 0.21 | 0.26 | 0.23 | 0.23 | 0.17 | 0.24 | 0.23 | 0.15 | 0.25 | 0.17 | 0.26 | 0.25 | 0.26 | 0.16 | 0.3 | 1 | | | | | | | | | | | | | |
| M19 | 0.29 | 0.25 | 0.35 | 0.44 | 0.31 | 0.33 | 0.26 | 0.29 | 0.29 | 0.22 | 0.37 | 0.27 | 0.4 | 0.38 | 0.41 | 0.29 | 0.44 | 0.3 | 1 | | | | | | | | | | | | |
| M20 | 0.21 | 0.19 | 0.27 | 0.37 | 0.26 | 0.27 | 0.19 | 0.25 | 0.28 | 0.13 | 0.29 | 0.18 | 0.35 | 0.29 | 0.33 | 0.24 | 0.37 | 0.24 | 0.51 | 1 | | | | | | | | | | | |
| M21 | 0.24 | 0.22 | 0.27 | 0.33 | 0.23 | 0.31 | 0.26 | 0.29 | 0.31 | 0.18 | 0.28 | 0.23 | 0.35 | 0.28 | 0.28 | 0.23 | 0.43 | 0.2 | 0.32 | 0.3 | 1 | | | | | | | | | | |
| M22 | 0.28 | 0.27 | 0.32 | 0.39 | 0.24 | 0.33 | 0.27 | 0.29 | 0.33 | 0.21 | 0.33 | 0.29 | 0.39 | 0.36 | 0.22 | 0.27 | 0.41 | 0.23 | 0.38 | 0.31 | 0.37 | 1 | | | | | | | | | |
| M23 | 0.22 | 0.27 | 0.31 | 0.33 | 0.16 | 0.24 | 0.24 | 0.28 | 0.29 | 0.12 | 0.3 | 0.2 | 0.32 | 0.27 | 0.27 | 0.23 | 0.39 | 0.24 | 0.33 | 0.3 | 0.32 | 0.34 | 1 | | | | | | | | |
| M24 | 0.21 | 0.25 | 0.27 | 0.3 | 0.16 | 0.24 | 0.21 | 0.25 | 0.25 | 0.14 | 0.26 | 0.19 | 0.28 | 0.26 | 0.22 | 0.22 | 0.34 | 0.18 | 0.28 | 0.26 | 0.27 | 0.31 | 0.48 | 1 | | | | | | | |
| M25 | 0.25 | 0.22 | 0.25 | 0.31 | 0.23 | 0.27 | 0.25 | 0.24 | 0.29 | 0.16 | 0.29 | 0.21 | 0.33 | 0.31 | 0.28 | 0.2 | 0.37 | 0.22 | 0.3 | 0.25 | 0.32 | 0.34 | 0.31 | 0.26 | 1 | | | | | | |
| M26 | 0.42 | 0.34 | 0.43 | 0.51 | 0.36 | 0.5 | 0.31 | 0.38 | 0.43 | 0.34 | 0.44 | 0.38 | 0.54 | 0.53 | 0.38 | 0.31 | 0.56 | 0.33 | 0.49 | 0.38 | 0.4 | 0.5 | 0.4 | 0.37 | 0.41 | 1 | | | | | |
| M27 | 0.35 | 0.25 | 0.36 | 0.42 | 0.34 | 0.42 | 0.27 | 0.34 | 0.4 | 0.22 | 0.34 | 0.33 | 0.46 | 0.43 | 0.33 | 0.21 | 0.48 | 0.28 | 0.36 | 0.32 | 0.36 | 0.39 | 0.34 | 0.36 | 0.36 | 0.62 | 1 | | | | |
| M28 | 0.31 | 0.27 | 0.36 | 0.44 | 0.27 | 0.36 | 0.27 | 0.35 | 0.37 | 0.21 | 0.39 | 0.28 | 0.44 | 0.43 | 0.35 | 0.27 | 0.52 | 0.28 | 0.45 | 0.35 | 0.34 | 0.37 | 0.38 | 0.34 | 0.32 | 0.52 | 0.46 | 1 | | | |
| M29 | 0.15 | 0.09 | 0.2 | 0.23 | 0.17 | 0.18 | 0.13 | 0.15 | 0.15 | 0.13 | 0.22 | 0.17 | 0.24 | 0.28 | 0.23 | 0.14 | 0.22 | 0.12 | 0.26 | 0.19 | 0.15 | 0.25 | 0.18 | 0.16 | 0.19 | 0.27 | 0.19 | 0.26 | 1 | | |
| M30 | 0.35 | 0.24 | 0.38 | 0.39 | 0.3 | 0.37 | 0.22 | 0.31 | 0.3 | 0.25 | 0.36 | 0.29 | 0.45 | 0.46 | 0.34 | 0.21 | 0.45 | 0.25 | 0.43 | 0.32 | 0.31 | 0.39 | 0.3 | 0.28 | 0.3 | 0.58 | 0.45 | 0.44 | 0.25 | 1 | |
| M31 | 0.25 | 0.26 | 0.32 | 0.33 | 0.17 | 0.24 | 0.21 | 0.2 | 0.25 | 0.16 | 0.25 | 0.22 | 0.31 | 0.27 | 0.22 | 0.2 | 0.32 | 0.17 | 0.29 | 0.22 | 0.24 | 0.29 | 0.3 | 0.29 | 0.22 | 0.36 | 0.32 | 0.33 | 0.13 | 0.28 | 1 |

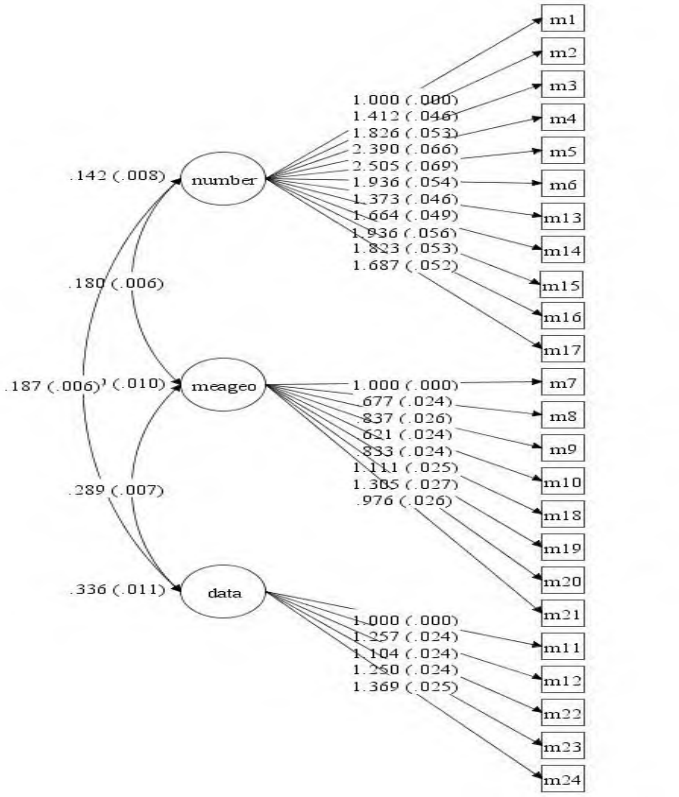
## Appendix 8. 8th Grade Mathematics Tetrachoric Correlation Analysis Results

| | M1 | M2 | M3 | M4 | M5 | M6 | M7 | M8 | M9 | M10 | M11 | M12 | M13 | M14 | M15 | M16 | M17 | M18 | M19 | M20 | M21 | M22 | M23 | M24 | M25 | M26 | M27 | M28 | M29 | M30 | M31 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| M1 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| M2 | 0.46 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| M3 | 0.61 | 0.71 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| M4 | 0.41 | 0.46 | 0.51 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| M5 | 0.38 | 0.45 | 0.52 | 0.41 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| M6 | 0.36 | 0.38 | 0.45 | 0.44 | 0.48 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | |
| M7 | 0.41 | 0.47 | 0.5 | 0.45 | 0.49 | 0.48 | 1 | | | | | | | | | | | | | | | | | | | | | | | | |
| M8 | 0.35 | 0.42 | 0.45 | 0.39 | 0.4 | 0.36 | 0.38 | 1 | | | | | | | | | | | | | | | | | | | | | | | |
| M9 | 0.51 | 0.51 | 0.6 | 0.52 | 0.56 | 0.52 | 0.52 | 0.72 | 1 | | | | | | | | | | | | | | | | | | | | | | |
| M10 | 0.27 | 0.26 | 0.37 | 0.3 | 0.29 | 0.28 | 0.29 | 0.24 | 0.42 | 1 | | | | | | | | | | | | | | | | | | | | | |
| M11 | 0.26 | 0.19 | 0.31 | 0.25 | 0.2 | 0.19 | 0.2 | 0.22 | 0.34 | 0.2 | 1 | | | | | | | | | | | | | | | | | | | | |
| M12 | 0.41 | 0.36 | 0.48 | 0.35 | 0.33 | 0.36 | 0.33 | 0.31 | 0.46 | 0.27 | 0.24 | 1 | | | | | | | | | | | | | | | | | | | |
| M13 | 0.4 | 0.48 | 0.5 | 0.4 | 0.4 | 0.37 | 0.42 | 0.35 | 0.47 | 0.26 | 0.21 | 0.36 | 1 | | | | | | | | | | | | | | | | | | |
| M14 | 0.44 | 0.54 | 0.58 | 0.39 | 0.38 | 0.39 | 0.43 | 0.38 | 0.46 | 0.23 | 0.2 | 0.36 | 0.41 | 1 | | | | | | | | | | | | | | | | | |
| M15 | 0.46 | 0.46 | 0.59 | 0.43 | 0.4 | 0.38 | 0.41 | 0.34 | 0.51 | 0.28 | 0.28 | 0.39 | 0.4 | 0.41 | 1 | | | | | | | | | | | | | | | | |
| M16 | 0.42 | 0.43 | 0.54 | 0.39 | 0.35 | 0.32 | 0.36 | 0.33 | 0.48 | 0.25 | 0.23 | 0.35 | 0.35 | 0.37 | 0.52 | 1 | | | | | | | | | | | | | | | |
| M17 | 0.42 | 0.44 | 0.52 | 0.48 | 0.51 | 0.52 | 0.47 | 0.37 | 0.56 | 0.38 | 0.25 | 0.42 | 0.43 | 0.39 | 0.48 | 0.41 | 1 | | | | | | | | | | | | | | |
| M18 | 0.46 | 0.46 | 0.56 | 0.4 | 0.39 | 0.38 | 0.41 | 0.32 | 0.49 | 0.27 | 0.21 | 0.41 | 0.38 | 0.44 | 0.47 | 0.4 | 0.46 | 1 | | | | | | | | | | | | | |
| M19 | 0.51 | 0.5 | 0.62 | 0.51 | 0.53 | 0.47 | 0.46 | 0.45 | 0.65 | 0.38 | 0.32 | 0.45 | 0.45 | 0.45 | 0.5 | 0.48 | 0.54 | 0.52 | 1 | | | | | | | | | | | | |
| M20 | 0.41 | 0.51 | 0.55 | 0.38 | 0.36 | 0.31 | 0.37 | 0.38 | 0.48 | 0.23 | 0.2 | 0.35 | 0.38 | 0.42 | 0.4 | 0.39 | 0.39 | 0.41 | 0.52 | 1 | | | | | | | | | | | |
| M21 | 0.37 | 0.45 | 0.49 | 0.42 | 0.45 | 0.43 | 0.47 | 0.34 | 0.49 | 0.26 | 0.18 | 0.34 | 0.38 | 0.41 | 0.39 | 0.36 | 0.45 | 0.4 | 0.48 | 0.36 | 1 | | | | | | | | | | |
| M22 | 0.32 | 0.36 | 0.44 | 0.41 | 0.42 | 0.51 | 0.46 | 0.32 | 0.49 | 0.26 | 0.17 | 0.31 | 0.35 | 0.34 | 0.38 | 0.31 | 0.49 | 0.37 | 0.45 | 0.32 | 0.41 | 1 | | | | | | | | | |
| M23 | 0.47 | 0.52 | 0.58 | 0.5 | 0.54 | 0.5 | 0.53 | 0.43 | 0.58 | 0.34 | 0.25 | 0.4 | 0.44 | 0.47 | 0.5 | 0.45 | 0.52 | 0.49 | 0.59 | 0.47 | 0.58 | 0.5 | 1 | | | | | | | | |
| M24 | 0.55 | 0.6 | 0.67 | 0.56 | 0.6 | 0.58 | 0.58 | 0.48 | 0.67 | 0.39 | 0.31 | 0.51 | 0.54 | 0.53 | 0.58 | 0.54 | 0.62 | 0.57 | 0.68 | 0.57 | 0.6 | 0.55 | 0.71 | 1 | | | | | | | |
| M25 | 0.37 | 0.4 | 0.49 | 0.38 | 0.36 | 0.34 | 0.36 | 0.3 | 0.46 | 0.28 | 0.26 | 0.35 | 0.34 | 0.36 | 0.4 | 0.37 | 0.42 | 0.39 | 0.47 | 0.36 | 0.36 | 0.35 | 0.47 | 0.53 | 1 | | | | | | |
| M26 | 0.45 | 0.57 | 0.57 | 0.46 | 0.52 | 0.45 | 0.53 | 0.43 | 0.56 | 0.24 | 0.21 | 0.36 | 0.62 | 0.49 | 0.44 | 0.39 | 0.45 | 0.45 | 0.53 | 0.47 | 0.48 | 0.45 | 0.58 | 0.66 | 0.42 | 1 | | | | | |
| M27 | 0.31 | 0.38 | 0.38 | 0.34 | 0.31 | 0.28 | 0.32 | 0.29 | 0.39 | 0.21 | 0.17 | 0.31 | 0.3 | 0.32 | 0.32 | 0.31 | 0.32 | 0.31 | 0.37 | 0.32 | 0.32 | 0.28 | 0.38 | 0.45 | 0.32 | 0.38 | 1 | | | | |
| M28 | 0.49 | 0.53 | 0.56 | 0.5 | 0.52 | 0.48 | 0.5 | 0.46 | 0.6 | 0.39 | 0.3 | 0.46 | 0.47 | 0.5 | 0.47 | 0.43 | 0.54 | 0.48 | 0.59 | 0.48 | 0.53 | 0.48 | 0.62 | 0.67 | 0.57 | 0.59 | 0.44 | 1 | | | |
| M29 | 0.55 | 0.59 | 0.63 | 0.53 | 0.53 | 0.5 | 0.53 | 0.45 | 0.6 | 0.36 | 0.28 | 0.48 | 0.52 | 0.57 | 0.55 | 0.51 | 0.52 | 0.55 | 0.61 | 0.51 | 0.55 | 0.48 | 0.63 | 0.7 | 0.5 | 0.64 | 0.47 | 0.66 | 1 | | |
| M30 | 0.38 | 0.48 | 0.49 | 0.39 | 0.36 | 0.31 | 0.38 | 0.35 | 0.44 | 0.21 | 0.23 | 0.31 | 0.37 | 0.44 | 0.37 | 0.37 | 0.35 | 0.36 | 0.43 | 0.41 | 0.39 | 0.34 | 0.46 | 0.48 | 0.37 | 0.45 | 0.31 | 0.51 | 0.52 | 1 | |
| M31 | 0.46 | 0.51 | 0.56 | 0.48 | 0.46 | 0.43 | 0.48 | 0.4 | 0.55 | 0.34 | 0.29 | 0.42 | 0.42 | 0.49 | 0.5 | 0.47 | 0.48 | 0.49 | 0.53 | 0.44 | 0.43 | 0.44 | 0.53 | 0.59 | 0.44 | 0.5 | 0.34 | 0.54 | 0.56 | 0.45 | 1 |

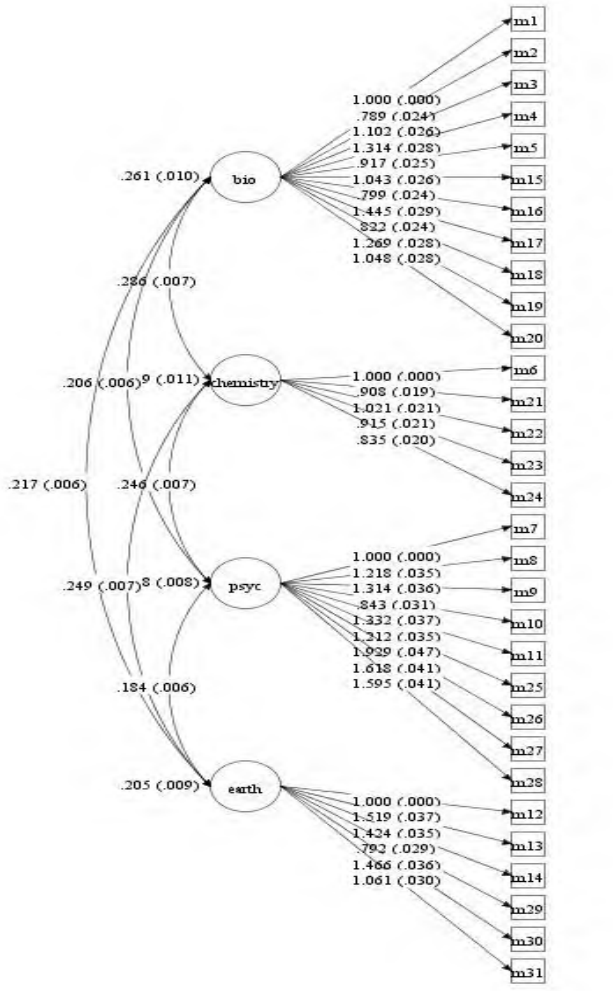
## Appendix 9. 4th Grade Science CFA Path Diagram



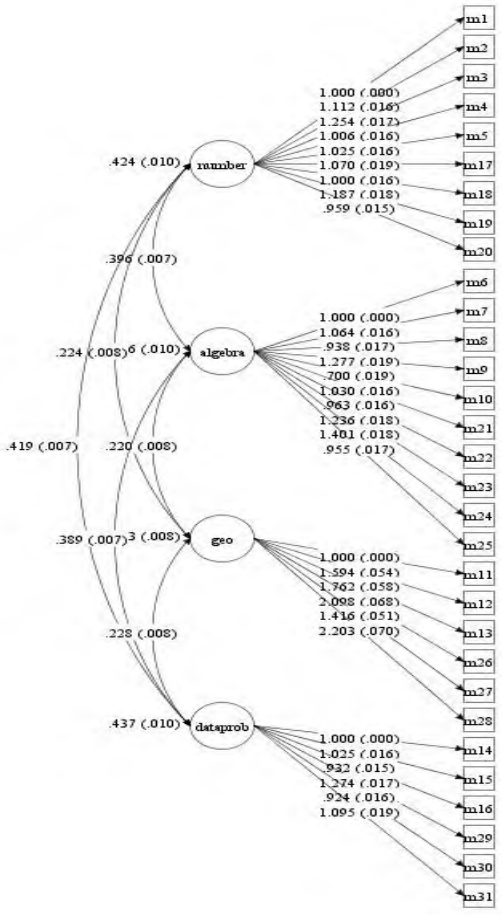
Appendix 10. 4th Grade Mathematics CFA Path Diagram



Appendix 11. 8th Grade Science CFA Path Diagram



Appendix 12. 8th Grade Mathematics CFA Path Diagram



Appendix 13. *Derived Items in TIMSS 2019*


## Appendix 10F: Derived Items in TIMSS 2019

| Grade 4 Mathematics |
|---|
| **M01_01 – ME51043:** Item parts A, B, C, D, E, F, G, and H are combined to create a 1-point item, where 1 score point is awarded if all parts are correct |
| **M01_05 – ME51508:** Item parts A and B are combined to create a 1-point item, where 1 score point is awarded if both parts are correct |
| **M02_03 – ME71167:** Item parts A, B, C, D, E, and F are combined to create a 1-point item, where 1 score point is awarded if all parts are correct |
| **M02_05 – ME71162, MP71162:** Item parts A and B are combined to create a 2-point item, where 2 score points are awarded if both parts are correct and 1 score point is awarded if 1 part is correct |
| **M02_06 – ME71078:** Item parts A, B, and C are combined to create a 1-point item, where 1 score point is awarded if all parts are correct |
| **M02_08 – ME71151, MP71151:** Item parts A, B, and C are combined to create a 2-point item, where 2 score points are awarded if all parts are correct and 1 score point is awarded if 2 parts are correct |
| **M02_11 – ME71142:** Item parts A and B are combined to create a 1-point item, where 1 score point is awarded if both parts are correct |
| **M02_12 – ME71204, MP71024:** Item parts A, B, and C are combined to create a 1-point item, where 1 score point is awarded if all parts are correct |
| **M04_03 – ME71036, MP71036:** Item parts A and B are combined to create a 1-point item, where 1 score point is awarded if both parts are correct |
| **M04_09 – ME71178, MP71176:** Item parts A, B, and C are combined to create a 1-point item, where 1 score point is awarded if all parts are correct |
| **M04_12 – ME71175, MP71175:** Item parts A, B, and C are combined to create a 2-point item, where 2 score points are awarded if all parts are correct and 1 score point is awarded if 1 or 2 are correct |
| **M06_01 – ME61018, MP61018:** Item parts A, B, C, and D are combined to create a 1-point item, where 1 score point is awarded if all parts are correct |
| **M06_10 – ME61266:** Item parts A, B, C, D, E, and F are combined to create a 2-point item, where 2 score points are awarded if all parts are correct and 1 score point is awarded if 5 parts are correct |
| **M08_11 – ME71141, M08_10 – MP71141:** Item parts A, B, C, and D are combined to create a 1-point item, where 1 score point is awarded if all parts are correct |
| **M08_12 – ME71194:** Item parts A and B are combined to create a 1-point item, where 1 score point is awarded if both parts are correct |
| **M08_13 – ME71193, M08_12 – MP71193:** Item parts A and B are combined to create a 2-point item, where 2 score points are awarded if both are correct and 1 score point is awarded if 1 part is correct |
| **M10_05 – ME71213:** Item parts A, B, and C are combined to create a 1-point item, where 1 score point is awarded if all parts are correct |
| **M10_08 – ME71179, MP71179:** Item parts A, B, and C are combined to create a 1-point item, where 1 score point is awarded if all parts are correct |
| **M10_12A – ME71187A:** Item parts A, B, C, and D are combined to create a 1-point item, where 1 score point is awarded if all parts are correct |

ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(2); 120-137

# Comparing Differential Item Functioning Based on Multilevel Mixture Item Response Theory, Mixture Item Response Theory and Manifest Groups

Ömer DOĞAN* Burcu ATAR**

**Abstract**

Studies on the differential item functioning (DIF) are usually considered in the context of manifest groups. Recently, with the increase in the number of analyses conducted with mixture models, investigating the situations that cause differences between groups has come to the forefront. In addition, it is considered important to examine the DIF with mixture models in which levels are also handled. In this study, it is aimed to compare the results of the multilevel mixture item response theory (MMIRT) model and the mixture item response theory (MIRT) model and the results of the DIF analyses based on the manifest groups. The research sample consists of students who answered the second booklet in the electronic Trends in International Mathematics and Science Study (eTIMSS) 2019 and coded their gender. The answers given to 15 items were analyzed with the Mantel Haenszel (MH) method for the gender variable according to the manifest groups, and with the selection of the most appropriate models by varying the number of groups and the number of levels according to the MIRT model and the MMIRT model. DIF analyses of the obtained latent groups were also performed with the MH method. In the light of the findings, the number of items displaying DIF in both the MIRT model and the MMIRT model is higher than the manifest groups. While only one item displayed DIF in the analysis according to gender, 14 items displayed DIF according to the MIRT model and seven items displayed DIF according to the MMIRT model. There is not a complete overlap in the number of DIF items and DIF effect sizes found as a result of the MIRT model and MMIRT model analyses. For this reason, a level analysis should be conducted before the analyses and if there is multi-levelness, the analyses should be conducted by taking this situation into consideration.

*Keywords:* multilevel mixture item response theory model, mixture item response theory model, manifest groups

## Introduction

In education, various tests are applied to determine the level of acquisition of the skills desired to be gained by individuals, to identify learning deficiencies and to place individuals in various institutions. In order to prevent errors in the tasks to be carried out through the scores obtained from these tests, several precautions are taken within the scope of measurement and evaluation. The fact that the scores of a test are valid and reliable contributes to the fairness of the decisions to be made using the scores. Validity, which is the first of these two important concepts, also includes reliability. Validity is a concept whose definition and content are constantly renewed according to the point of view in the historical process. Standards (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) define validity as the degree to which interpretations of test scores are supported by evidence and theory. Accordingly, validity is not a characteristic of the test, but is related to the inferences made from the test scores. The validity process also involves gathering the necessary evidence for a sound scientific basis for the proposed score interpretations. One of the evidences that should be obtained in this process can be obtained by analyzing differential item functioning (DIF), which is one of the evidences about the internal structure of the test. According to Kelderman and Macready (1990), test items exhibit DIF if the item scores of equal ability

* PhD Student., Hacettepe University, Faculty of Education, Ankara-Turkey, e-mail: 64omerdogan64@gmail.com, ORCID ID: 0000-0001-5169-520X
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, e-mail: burcua@hacettepe.edu.tr, ORCID ID: 0000-0003-3527-686X

To cite this article:
Doğan, Ö., & Atar, B. (2024). Comparing differential item functioning based on multilevel mixture item response theory, mixture item response theory and manifest groups. *Journal of Measurement and Evaluation in Education and Psychology, 15*(2), 120-137. https://doi.org/10.21031/epod.1457880

*Received: 24.03.2024*
*Accepted: 6.06.2024*

test takers from different groups (e.g., different gender, race, region or age) are significantly different. If a number of items on a test display DIF in favor of a specific group, there may be an unfair advantage for that group in terms of the assessed level of performance when compared to individuals from other groups. Items of test that display DIF are one of the important reasons for reducing the validity of the scores (Kristanjansonn et al., 2005; Messcik, 1995). DIF is an important indicator of test quality because it is directly related to the fairness and validity of the test. There are many methods for determining DIF, including Mantel-Haenszel (Holland & Thayer, 1988; Mantel & Haenszel, 1959), logistic regression, analysis of variance, transformed item difficulty and SIBTEST (Shealy & Stout, 1993) within the framework of classical test theory (CTT), Lord's (1980) chi-square method, Raju's (1988, 1990) field measurements and likelihood ratio test (Thissen et al.,1988) within the framework of item response theory (IRT). In DIF detection, the above-mentioned methods are compared with groups that are considered to be homogeneous within themselves, namely focal and reference groups. These groups are formed by gender, ethnicity, nationality, etc. and are referred to as manifest groups.

DIF detection methods in the context of CTT and IRT are very useful for detecting DIF in test administration, but they have made little progress in understanding the possible causes of DIF. This is because manifest group characteristics are typically only marginally related to the cause of DIF (Choi et al., 2015; Roussos & Stout, 1996). Several studies have shown that the homogeneity assumption is not always met in DIF analysis of manifest groups (e.g., Cohen & Bolt, 2005; de Ayala et al., 2002). Moreover, when differences between groups are found, it is not easily understood who is primarily advantaged or disadvantaged by DIF items (de Ayala et al., 2002).

Methods for DIF detection that have been mentioned in the context of IRT include comparisons of item parameters or areas between item response functions. However, efforts to understand why some test takers respond differently to these items are often conducted outside of the IRT context. Mixture IRT (MIRT) models have been proposed as a useful tool to investigate how differences in qualitative test takers, such as differences resulting from the use of different problem solving strategies, can lead to differences in responses to test items (Embretson & Reise, 2000). The use of the MIRT model, which is an integration of the IRT and latent class models, is typically exemplified by comparisons of item profiles across different latent groups or latent classes (Paek & Cho, 2015).

MIRT model is similar to a multigroup item response model, but the group of interest is not predetermined, but is determined based on the results obtained from model parameter estimation. As in multigroup item response models, item parameters and latent variable(s) may be different across latent groups in MIRT models (Cho et al., 2015). In MIRT models, individuals are assigned to non-predetermined classes with the highest within-group homogeneity and highest between-group heterogeneity in terms of the latent trait. Item parameters are estimated independently of the manifest group to which the individuals belong and specific to each group. Differences in group-specific estimated parameters suggest that DIF may be caused by a latent trait (De Ayala et al., 2002). De Boeck et al. (2011, p. 584) list four a priori reasons to consider implicit DIF analysis instead of manifest DIF analysis:

1. Lack of opinion (no idea about which group membership is interesting, or incomplete knowledge of group membership),

2. Unobservability (the group membership of interest is not observable),

3. Reliability (observed group membership may not be completely reliable) and

4. Validity (observed group membership may not be a completely valid indicator of actual group membership).

In the context of DIF models, Cohen and Bolt (2005) described a mixture Rasch model (MRM) approach to detecting uniform DIF, which differs from previous methods in some fundamental respects. This MRM is expressed as follows:

$$P(y_{ij} = 1|\theta_i) = \sum_{g=1}^{G} \pi_g \frac{\exp[(\theta_{ig} - b_{ig})]}{1+\exp[(\theta_{ig} - b_{ig})]} \quad (1)$$

$g$= 1,.....,$G$: Index indicating the latent class

$j$= 1,.....,$J$: Index indicating respondents

$\theta jg = j$. Individual's latent ability in latent class g

$\beta$= item difficulty parameter of item i in class $g$

Besides the MRM, there are also 2-parameter and 3-parameter models for mixture models. The two-parameter Mixture IRT model is shown as follows (Finch & French, 2012):

$$P(y_{ij} = 1|\theta_i) = \sum_{g=1}^{G} \pi_g \frac{\exp\left[a_{jg}\left(\theta_{ig} - b_{ig}\right)\right]}{1+\exp\left[a_{jg}\left(\theta_{ig} - b_{ig}\right)\right]} \quad (2)$$

The three-parameter Mixture IRT model, which includes item parameters and chance parameter for each grade, is shown as follows (Choi et al., 2015):

$$P(y_{ij} = 1|\theta_i) = \sum_{g=1}^{G} \pi_g [c_{jg} + (1 - c_{jg}) \frac{\exp\left[a_{jg}\left(\theta_{ig} - b_{ig}\right)\right]}{1+\exp\left[a_{jg}\left(\theta_{ig} - b_{ig}\right)\right]} \quad (3)$$

It can be said that MIRT models are important factors in the estimation of item parameters. In their study, Cohen and Bolt (2005) used mixture models to decompose the secondary dimension expressed by Ackerman (1992) and aimed to better understand the differences between test takers who were disadvantaged or advantaged by DIF items. In Study 1, they showed that the conventional approach to studying DIF does not contribute much to understanding the causes of DIF. They concluded that using explicit gender categories to identify those affected by gender DIF is likely to be misleading. Study 2 extended the analysis of DIF, showing how mixture models can be used to identify latent groups where some form of DIF may be present in the first place. In the case of the groups in Study 2, it was explained that there is a cognitive interpretation of the secondary dimension and thus the cause of the DIF can be more easily interpreted. As a result, in the case of gender DIF, it was clear that not all members of a gender group responded in the same way to items that were allegedly biased for or against their group, with some men being disadvantaged by items that were found to advantage men and some women being advantaged by items that were found to disadvantage women. Therefore, when it is accepted that DIF items do not universally advantage or disadvantage all members of a group, this practice becomes questionable. Similarly, Samuelsen (2005) based the basic premise of his study on the fact that it is not advisable to use open groups in DIF analyses. She argued that distinctions based on external characteristics of test takers are not useful and that the groups that emerge are neither homogeneous nor cognitively meaningful. Instead, by examining the latent dimensions underlying student performance, it is possible to identify and interpret the reasons behind DIF. By using the latent class perspective, individual differences in human behavior can be attributed to potentially meaningful dimensions rather than external characteristics, and when this happens, it is possible to truly explain why items work differently. In their study, Jiao and Chen (2014) addressed the problems arising from the use of the DIF approach based on traditional observed groups and analyzed both background and cognitive covariates that are effective in the characterization of latent class membership. The results of the study showed that a sole manifest group variable is insufficient to fully predict the sources of implicit DIF and that the implicit class-based DIF approach is a possible method for screening for potential DIF items arising from the intervening effects of multiple variables. The aforementioned studies and others (Cho & Cohen, 2010; Dras, 2023; Zhang, 2017) have shown that the MRM approach can provide more insight into the antecedents of DIF than methods that rely on assessing DIF in relation to manifest groups. In addition, this approach to DIF assessment has the potential to provide more comprehensive analyses that do not rely on a predetermined ranking of individuals, which itself may be biased in some respects (Finch & Finch, 2013). The mixture model is used to define latent classes of test takers who are homogeneous in terms of their item response patterns. Members of each latent class differ in ability and response strategies differ across classes. However, an important limitation of the mixture model is that it essentially ignores the underlying multilevel structure that exists beyond the student level in most educational test data (Cho & Cohen, 2010).

If the analysis is restricted to the traditional linear model, the basic assumptions are normality, homoscedasticity, linearity, and independence. It is desirable to preserve normality and linearity in the analyses, but the assumption of homoscedasticity and especially the assumption of independence need to be adapted. The general idea behind such adaptations is that persons in the same group are closer or more similar than persons in different groups. Thus, individuals in different classes may be independent, but individuals in the same class share values on many more variables (Raudenbush & Bryk, 2002). The biggest threat to the local independence assumption is the nested data structure (Jiao et al., 2012). For example, the multilevel data structure manifest in achievement tests is a structure in which students are nested to teachers and teachers are nested to schools. In addition to the mixture model, a fairly recent contribution to the DIF literature has been the emergence of methods for dealing with the multilevel data structure that is common in such assessments (French & Finch, 2010). For instance, especially in large-scale assessments, data for DIF detection studies are often collected from test takers nested within schools. In such cases, schools should be assumed to influence test item responses, at least to some extent. This influence will be expressed in the form of non-trivial intracluster correlation (ICC) values. When such multilevel data structure is ignored and ICCs are non-zero (or very close to it), the resulting analyses are likely to yield erroneous estimates of item parameters and their associated standard errors, leading to erroneous DIF detection results. Researchers (e.g. Finch &French, 2012) have continued to develop and adapt multilevel methods for DIF detection in the context of manifest groups (Finch & Finch, 2013).

Cho and Cohen (2010) described the MMIRT model, which allows for the simultaneous detection of differences in latent class structure at both test taker and school levels. Student-level latent classes capture the relationship between responses in the student-level unit. The MIRT model assumes that there may be heterogeneity in response patterns at the first level that should not be ignored (Mislevy & Verhelst, 1990; Rost, 1990). However, the MMIRT model also takes into account the possibility that there may not be latent classes at the first level. (Cho, 2007).

In the MMIRT model, dependency is taken into account by including latent variables at higher-level continuous and/or categorical latent variables. Vermunt (2007) proposed eight possible versions of two-level (e.g., students nested within schools) MMIRT models. Latent variables at each level of mixture models can be categorical, continuous, or both categorical and continuous, as mixture models include categorical latent variables and item response models include continuous latent variables (as cited in Lee et al., 2018).

Cho and Cohen (2010) showed in their study that it is possible to obtain grade-specific item difficulties for each level 1 and 2 and express them on the same scale. In the empirical example they examined, the mixed groups at the student and school level that emerged in the data were similarly clearly distinguishable in terms of ability levels, item difficulty profiles, student and school demographics, and response patterns, but when more than one factor characterizes a class, it can be difficult to find factors that potentially cause DIF. Gurkan (2021) used Programme for International Student Assessment (PISA) 2012 data to investigate the correlation patterns of the multidimensional and multilevel MIRT model and to improve the model, and aimed to investigate the variance between within-country correlations based on traditional estimates and to determine to what extent this variance is due to heterogeneity in the amount of measurement error and the clustered nature of the data. As required by the characteristics of the PISA data, the multidimensional MMIRT models used in the study not only appropriately accounted for measurement error and clustering in the data, but also took into account the possibility of different subpopulations within countries.

Another international study of the PISA type is TIMSS. TIMSS is an international comparative study that measures student achievement in mathematics and science worldwide. Conducted in a four-year assessment cycle since 1995, TIMSS has assessed student achievement in fourth and eighth grades seven times - 1995, 1999, 2003, 2007, 2011, 2015 and 2019 - and accumulated 24 years of trend measurements. In 2019, TIMSS began transitioning to computer-based assessment by introducing a digital version of the paper-and-pencil assessment called “eTIMSS”. Within the scope of the research, the use of real data was planned and eTIMSS 2019 data was utilized. This is because the DIF studies conducted with MIRT

models, which have been increasing recently, are mostly conducted using simulative data (e.g. Cho, 2007; Cho & Cohen, 2010; de Ayala et al. 2002; Sırgancı, 2019; Uyar, 2015). In the current studies, deficiencies such as disregarding the levelness (Choi et al., 2015; Toker & Green, 2021; Yalcin, 2018, etc.), conducting DIF analysis based only on manifest groups (Aydemir, 2023; Bayram, 2024; Unal, 2023, etc.), lack of using real data (Sırgancı, 2019; Uyar, 2015, etc.) and ignoring the source of DIF were found. The aim of this study is to compare the results of DIF analyses based on the MMIRT model and manifest groups and to investigate what may cause performance differences in eTIMSS. In the study, DIF analysis was performed on the data in eTIMSS booklet 2 with the MMIRT model and the results obtained were compared with the results of DIF analysis based on the MMIRT model and manifest groups. In the light of the findings obtained, the number of latent classes, items with DIFs and changes in the number of items with DIFs were examined when multi-levelness and the differentiation of manifest groups and latent groups were included in the analyses in studies such as TIMSS prepared for cross-country comparisons in education. Thus, by comparing mixture models and manifest groups methods, the differences in determining the source of DIF were revealed and it was investigated whether the addition of multi-levelness to the mixture model had a positive effect on the complexity of the model.

## Methods

### Sample

In this study, the typical case sampling method of purposive sampling was used. Since the models used in the DIF analysis (MMIRT and MIRT) are based on the IRTTIMSS items developed according to this model were used. In 2019, TIMSS started to move to computer-based assessment by introducing a digital version of the paper-and-pencil assessment called "eTIMSS". This included 22 countries at the eighth grade level and five participants from regions or cities of some countries as benchmark participants.

In the study, the second booklet was selected because it is suitable for multilevel data structure and the number of multiple-choice items is higher than the other booklets. For the study, the responses of eighth grade students from 22 countries in the eTIMSS 2019 data to 15 dichotomously scored mathematics and science items in the second booklet were used. Within the scope of the research, the answers of 8167 individuals were analyzed and 123 individuals were excluded from the study because their gender was not specified. Finally, the data of 8044 individuals were analyzed. Li et al., (2009) stated that a sample size of 600 individuals would be appropriate for MIRT models when the number of items is between 15 and 30. In addition, Li et al. (2009) stated that for a 15-item test, a sample size of 600 would be sufficient in a model with 1 to 4 classes for both MIRT 2PL and MIRT 3PL models. Cho et al., (2013) state that a sample size of more than 360 can be used for the MRM. Cohen and Bolt (2005) successfully applied the MIRT 3PL model with a sample size of 1000. Demographic information is presented in Table 1.

**Table 1**

*22 eTIMSS participant countries, number of participants and average scores*

| Country | Number of Participants | Mean Score | Gender(F/M) |
|---|---|---|---|
| United Arab Emirates (UAE) | 1584 | 6.42 | 774/810 |
| Chile | 289 | 5.42 | 141/148 |
| England | 222 | 7.00 | 120/102 |
| Finland | 347 | 7.27 | 178/169 |
| France | 266 | 5.37 | 139/127 |
| Georgia | 244 | 5.82 | 118/126 |
| Hong Kong | 228 | 8.64 | 109/119 |
| Hungary | 328 | 7.65 | 181/147 |
| Israel | 267 | 7.36 | 141/126 |
| Italy | 257 | 6.12 | 132/125 |
| Korea Rep. of | 273 | 10.79 | 137/136 |

**Table 1 (continued)**

| Country | Number of Participants | Mean Score | Gender(F/M) |
|---|---|---|---|
| Lithuania | 259 | 6.95 | 125/134 |
| Malaysia | 499 | 7.50 | 258/241 |
| Norway | 335 | 6.96 | 156/159 |
| Portugal | 238 | 6.45 | 122/116 |
| Qatar | 278 | 6.06 | 129/149 |
| Russian Federation | 278 | 8.33 | 125/153 |
| Singapore | 352 | 10.53 | 178/174 |
| Sweden | 280 | 7.38 | 127/153 |
| Türkiye | 289 | 7.11 | 137/152 |
| Chinese Taipei | 349 | 10.58 | 178/171 |
| United States | 602 | 7.55 | 278/324 |
| General | 8044 | 7.33 | 3983/4061 |

As seen in Table 1, the number of male and female students is close to each other. The highest number of participants was from the UAE, while the lowest number of participants was from England. Looking at the mean scores, the three highest scores belong to the states located in Asia.

## Data Analysis

The eTIMSS 2019 application consists of 14 booklets. The booklets contain mathematics and science items with certain common items. The items are prepared as multiple-choice, open-ended and short-answer. Within the scope of the study, 31 items from mathematics and science courses were selected, all of which were four-choice multiple-choice items. ICC values and dimensionality structure of the items were examined for the planned MMIRT model. Students were identified as level 1 and countries as level 2. The fact that the ICC values are close to zero indicates that there is no nested structure. For this reason, items with ICC values close to zero were removed and the analyses continued with the remaining 15 items. The average of the ICC values of the selected items is approximately .15. In other words, approximately 15% of the variance is due to country differences. Muthen (1997) suggested that multilevel modeling should definitely be taken into account when group sizes exceed 15 if the ICC>.10, and Julian (2001) and Selig et al., (2008) suggested that the hierarchical structure should not be ignored even when the ICC values are lower than .10 (as cited in Şen, 2022). The ICC values for 15 items are given in Table 2.

**Table 2.**

*ICC values for the selected 15 items*

| Item Numbers | ICC Values |
|---|---|
| 1 | .12 |
| 2 | .11 |
| 3 | .17 |
| 4 | .15 |
| 5 | .14 |
| 6 | .13 |
| 7 | .17 |
| 8 | .16 |
| 9 | .20 |
| 10 | .10 |
| 11 | .11 |
| 12 | .12 |
| 13 | .17 |
| 14 | .15 |
| 15 | .19 |
| Mean | .15 |

According to Table 2, the lowest ICC value is .10 while the highest value is .19 These values indicate that at least 10% of the variance of each item is due to country differences. Regarding the 15 items used in the study, it was examined whether there was a unidimensional structure. In order to determine this, the suitability of the data for factor analysis was examined using the 'fa' function in the 'psych' package of R software (Revelle, 2023) and exploratory factor analysis (EFA) based on the tetrachoric correlation matrix was performed on the data. The adequacy of the correlation matrix between the items and its comparison with the unit matrix were examined with Kaiser-Meyer-Olkin (KMO) coefficient and Barlett's test of sphericity. For factorization, the KMO value is expected to be higher than 0.60 and the Barlett test is expected to be significant (Büyüköztürk, 2018). For 15 items, the KMO value was found to be 0.850 and the Barlett test was significant ($p<.001$). Therefore, it was interpreted that the data was appropriate for factorization. The eigenvalues obtained in the analyses for dimensionality are shown in Table 3 and the slope accumulation graph is shown in Figure 1. According to the values obtained, it is understood that the data shows a unidimensional structure.

**Table 3.**

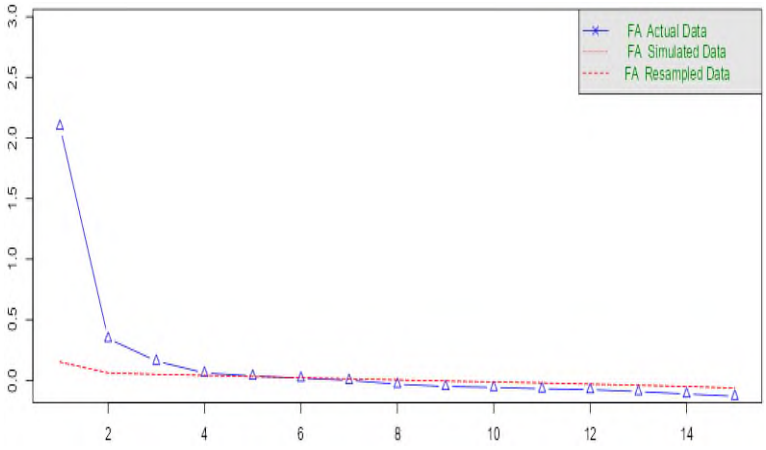
*Eigenvalues obtained in dimensionality analyses for 15 items*

| Factor Number | Eigenvalue |
|---|---|
| 1 | 2.10 |
| 2 | .35 |
| 3 | .16 |
| 4 | .06 |
| 5 | .04 |

According to Table 3, only the eigenvalue for the first dimension is greater than 1, the others are less than 1 and the ratio between the first two eigenvalues is six times. According to Kaiser's (1960) K1 rule, the construct is unidimensional.

**Figure 1.**

*Slope deposition graph for 15 items*




The courses, subject areas and cognitive domains of the 15 items selected for the analysis are given in Table 4. Accordingly, the subject areas of the items selected from seven mathematics and eight science courses consist of numbers, algebra, geometry, biology, chemistry, earth science, and physics. In addition, there are items from all three cognitive domains of eTIMSS: knowing, applying, and reasoning.

**Table 4.**

*Courses, subject areas and cognitive domains of the data*

| Item Number | Course | Subject Area | Cognitive Domaim |
|---|---|---|---|
| 1 | Math | Numbers | Applying |
| 2 | Math | Algebra | Knowing |
| 3 | Math | Algebra | Knowing |
| 4 | Math | Algebra | Knowing |
| 5 | Math | Algebra | Knowing |
| 6 | Math | Geometry | Applying |
| 7 | Math | Geometry | Reasoning |
| 8 | Science | Biology | Knowing |
| 9 | Science | Chemistry | Knowing |
| 10 | Science | Earth Science | Knowing |
| 11 | Science | Earth Science | Reasoning |
| 12 | Science | Earth Science | Applying |
| 13 | Science | Chemistry | Applying |
| 14 | Science | Biology | Applying |
| 15 | Science | Physics | Knowing |

According to Table 4 in the math section includes one item on numbers, four items on algebra and two items on geometry. In the science section, there are two items each from biology and chemistry, one item from physics and three items from earth science. Three information criterion indices are used to determine the appropriate model for parameter estimation based on the MIRT and MMIRT models. Akaike's (1974) information criteria (AIC), Schwarz's (1978) Bayesian information criterion (BIC) and the sample size-adjusted version of BIC (SABIC; Sclove, 1987). Within the scope of the research, BIC value is used in accordance with the literature (Choi et al., 2015; Li et al. 2009; Şen & Toker, 2021).

The Mplus software package was used to determine the appropriate model based on the DIF according to the MIRT and MMIRT (Muthén & Muthén, 2017). The robust version of the marginal maximum likelihood estimation technique (MLR) was used in Mplus parameter estimation. In addition, the number of iterations was increased in the analyses as the models became more complex. For the DIF analysis for manifest groups, the MH technique was chosen and the "difR" package in the R software language was used (Magis et al., 2015). Latent classes were characterized in terms of item difficulty parameter estimates and descriptive characteristics of test takers. As suggested by Cho and Cohen (2010), test taker-level DIF analyses were conducted separately for each second-level latent class, while uniform country-level DIF was determined by comparing school latent class item difficulty estimates across test taker levels. It was decided to use the standardized MH test when there were two latent classes and the Generalized Mantel Haenszel (GMH) when there were more latent groups. If $\Delta MH>0$, DIF is interpreted as DIF in favor of the focus group, $\Delta MH<0$ as DIF in favor of the reference group, and $\Delta MH\cong0$ as no DIF (Holland & Thayer, 1986).

## Results

Within the scope of the study, 9 different models were analyzed for MIRT and MMIRT with the data set consisting of 15 items in eTIMSS booklet 2. Model fit statistics for these nine models are presented in Table 5.

**Table 5**

*Model fit statistics for 9 models*

| Model | LogL | np | AIC | BIC | SABIC |
|---|---|---|---|---|---|
| L0-G2 | -73891.28 | 61 | 147904.55 | **148331.11** | 148137.26 |
| L0-G3 | -73764.92 | 91 | 147711.84 | 148348.17 | 14805899 |
| L0-G4 | -73764.96 | 92 | 147713.91 | 148357.24 | 148064.88 |
| L1-G2 | -72759.79 | 91 | 145701.58 | 146337.92 | 146048.75 |
| L1-G3 | -72492.14 | 137 | 145258.28 | 146216.28 | 145780.92 |
| L1-G4 | -72372.38 | 183 | 145110.75 | 146390.41 | 145808.87 |
| L2-G2 | -71745.28 | 167 | 144138.24 | **145306.02** | 144775.33 |

**Table 5 (continued)**

| Model | LogL | np | AIC | BIC | SABIC |
|---|---|---|---|---|---|
| L2-G3 | -71591.89 | 243 | 143669.77 | 145368.99 | 144596.79 |
| L2-G4 | -71388.76 | 319 | 143415.56 | 145646.18 | 144632.46 |

LogL: Log-likelihood; np: Number of Parameter; AIC: Akaike's Information Criteria ; BIC: Bayesian Information Criterion; SABIC: Sample Size-Adjusted Version of BIC

As shown in Table 5, the level 0 and number of groups 2 (L0-G2) model has the smallest BIC value among the MIRT models. The level 2 and number of groups 2 (L2-G2) model has the smallest BIC value among the MMIRT models. As mentioned in the data analysis section, it is in the literature that BIC is more appropriate in the selection of mixture models. Therefore, in the light of these results, the L0-G2 model among the MIRT models and the L2-G2 model among the MMIRT models are used in the analyses. Based on the L0-G2 model, students are divided into two latent student classes, and based on the L2-G2 model into two latent student classes and two latent country classes. Table 6 and Table 7 present the final class numbers and proportions for each latent class variable based on the estimated posterior probabilities for the MIRT and MMIRT models. Student-level latent class 2 is the dominant group (.73) in the MIRT model. Note that the sum of the proportions reported in Table 6 is equal to 1. In the MMIRT model, the second level student level class 1 is the dominant group (.45).

**Table 6**.

*Final Class Numbers and Ratios for Each Student Level Latent Classroom for the MIRT Model*

| Latent Class | Number of Individuals (Female/Male) | Ratio |
|---|---|---|
| 1 | 2233(1149/1084) | .28 |
| 2 | 5811(2834/2977) | .72 |

**Table 7.**

*Final Class Numbers and Ratios for the Student and Country Level Latent Class for the MMIRT Model*

| | Student Level Latent Group | |
|---|---|---|
| **Country Level Latent Group** | 1 | 2 |
| 1 | 1499(741/758) (.19) | 240(125/115) (.03) |
| 2 | 5319(2640/2679) (.66) | 986(477/509) (.12) |

Table 6 and Table 7 present the final class numbers and proportions for each latent class variable based on the estimated posterior probabilities for the MIRT and MMIRT models. Student-level latent class 2 is the dominant group (.72) in the MIRT model. Note that the sum of the proportions reported in Table 6 is equal to 1. In the MMIRT model, the second level student level class 1 is the dominant group (.66).

The item parameter estimations of the final model are reported in Table 8, Table 9 and Table 10. The Mplus output provides separate slope and intercept or threshold parameters for within-group and between-groups for the MMIRT models. For this reason, the subscripts W (within-group) and B (between-group) are used to distinguish between the two levels. As illustrated in Table 6, slope ($\alpha$) parameters are reported for each class at both levels. But thresholds were obtained only for the between-levels part. As described by Sen et al. (2020), the IRT discrimination parameters are equal to the slope parameters provided in the Mplus output. Nevertheless, item difficulty parameters can be obtained by dividing the threshold values for each item by the slope values. In the MIRT model, item difficulty parameters for latent class 2 appear to be higher than latent class 1.

**Table 8.**

*Item Parameter Estimations of the Final Model for MIRT*

| | Latent Class 1 | | Latent Class 2 | |
|---|---|---|---|---|
| **Item** | **$\alpha_1$** | **$\beta_1$** | **$\alpha 2$** | **$\beta 2$** |
| 1 | 1.18 | -1.34 | .55 | 1.01 |
| 2 | .89 | -3.49 | .66 | -.28 |
| 3 | .13 | -1.57 | .51 | 2.19 |
| 4 | .71 | .21 | -.09 | -3.63 |
| 5 | .39 | -.43 | .16 | 4.01 |
| 6 | 1.31 | -.99 | .10 | 5.35 |
| 7 | 1.13 | .35 | -.14 | -1.76 |
| 8 | .94 | -.81 | .71 | .03 |
| 9 | 1.05 | -.44 | .74 | .86 |
| 10 | .86 | -1.64 | 1.25 | -.39 |
| 11 | .92 | -.79 | .79 | -.09 |
| 12 | .94 | -2.06 | 1.01 | -.76 |
| 13 | .84 | -.11 | .66 | .80 |
| 14 | .99 | -.21 | .56 | 1.36 |
| 15 | .49 | -4.31 | .88 | -.50 |

When item difficulty indices are analyzed, items 1,2,3,5,6,8,9,10,11,12,13,14 and 15 are lower for latent class 1 than latent class 2, i.e. they are easier. The remaining two items, items 4 and 7, are easier for latent class 2.

**Table 9.**

*Mean scores and standard deviations for the latent classrooms in the MIRT model*

| Latent Class | Mean Score (Female/Male) | Standard Deviation |
|---|---|---|
| 1 | 10.67(10.30/11.05) | 2.38(2.35/2.35) |
| 2 | 6.05(5.95/6.15) | 2.39(2.32/2.45) |

According to the MIRT model, the averages of male students in both implicit groups are higher than the averages of female students. In general averages, latent class 1 has a higher average than latent class 2 and it can be said that latent class 1 is more successful.

**Table 10.**

*Item Parameter Estimates of the Final Model for Student Level*

| | Latent Class 1 | | | Latent Class 2 | | |
|---|---|---|---|---|---|---|
| **Item** | **$\alpha_{1W}$** | **$\alpha_{1B}$** | **$\beta_1$** | **$\alpha_{2W}$** | **$\alpha_{2B}$** | **$\beta_2$** |
| 1 | 1.41 | 1.82 | -.38 | .28 | 1.08 | .40 |
| 2 | 1.60 | 1.83 | -1.26 | .01 | -.69 | 1.23 |
| 3 | 1.33 | 1.58 | -.65 | .26 | -.70 | -.54 |
| 4 | 1.10 | 1.02 | .51 | .48 | -.40 | -5.04 |
| 5 | .54 | 2.37 | .58 | .01 | -.76 | -.96 |
| 6 | 1.47 | 2.01 | -.56 | .34 | .90 | .60 |
| 7 | .96 | 2.37 | .37 | .31 | 1.32 | 1.25 |
| 8 | .85 | -.05 | 4.54 | -1.15 | -1.19 | .69 |
| 9 | .79 | 1.87 | -0.03 | -1.01 | 1.11 | .70 |
| 10 | 1.05 | .18 | -3.96 | -3.32 | -1.75 | .94 |
| 11 | .76 | 1.35 | -.13 | -1.11 | 1.56 | -.04 |
| 12 | .75 | .66 | -3.22 | -4.70 | 4.46 | -.13 |
| 13 | .83 | -.43 | .66 | -1.06 | .35 | .28 |
| 14 | .81 | .92 | -.38 | -.79 | .16 | 4.14 |
| 15 | .94 | 1.30 | -1.50 | -1.45 | -.05 | 3.21 |

When item difficulty indexes are analyzed, items 1,2, 6, 7, 9, 10, 12, 14, and 15 are lower for latent class 1 than latent class 2, i.e. they are easier. The remaining six items, items 3, 4, 5, 8, 11, and 13 are easier for latent class 2. Table 11 presents the item parameter estimates of the final model for the country level of the MMIRT model.

**Table 11.**

*Item Parameter Estimates of the Final Model for the Country Level of the MMIRT Model*

| | Latent Class 1 | | | Latent Class 2 | | |
|---|---|---|---|---|---|---|
| **Item** | **$\alpha_1$W** | **$\alpha_1$B** | **$\beta_1$** | **$\alpha_2$W** | **$\alpha_2$B** | **$\beta_2$** |
| 1 | 1.00 | -.42 | -1.38 | 1.00 | .55 | -3.35 |
| 2 | 1.31 | .17 | -1.02 | 1.31 | .15 | -2.36 |
| 3 | .88 | .25 | 3.90 | .88 | .84 | -1.57 |
| 4 | -.08 | .23 | 3.93 | -.08 | .82 | .48 |
| 5 | .51 | .51 | 1.21 | .51 | 1.23 | -.60 |
| 6 | .07 | -.11 | -5.83 | .07 | -.80 | 1.35 |
| 7 | -.09 | .14 | 4.51 | -.09 | .35 | 1.78 |
| 8 | 1.18 | -.26 | -.59 | 1.18 | .05 | -3.16 |
| 9 | 1.32 | -1.08 | -.64 | 1.32 | -1.79 | .40 |
| 10 | 2.26 | -.87 | .53 | 2.26 | .41 | -4.15 |
| 11 | 1.50 | -.37 | .09 | 1.50 | -.27 | 4.11 |
| 12 | 2.30 | .06 | -.80 | 2.30 | .86 | -2.13 |
| 13 | .96 | -.80 | -.80 | .96 | -.71 | .79 |
| 14 | .70 | -.95 | -.93 | .70 | -2.00 | .24 |
| 15 | 2.03 | .70 | -.55 | 2.03 | 1.74 | -1.16 |

When item difficulty indexes are analyzed, items 6, 9, 11, 12, 13, and 14 are lower for latent class 1 than latent class 2, i.e. they are easier. The remaining nine items, items 1, 2, 3, 4, 5, 7, 8, 10, and 15 are easier for latent class 2. Table 12 shows the mean scores and standard deviations for the latent classes in the MMIRT model.

**Table 12.**

*Mean scores and standard deviations for the latent classes in the MMIRT model*

| **Latent Class** | **Mean Scores** | **Standard Deviations** |
|---|---|---|
| 1-1 | 6.85 | 2.89 |
| 1-2 | 6.82 | 2.92 |
| 2-1 | 7.43 | 3.19 |
| 2-2 | 7.67 | 3.31 |
| 1 | 6.84 | 2.89 |
| 2 | 7.47 | 3.21 |

Of the two latent classes for country level 1, latent class 1 has a higher mean than students in latent class 2. For country level 2, of the two latent classes, latent class 2 has a higher mean than students in latent class 1. Table 13 presents the countries included in the country-level latent classes, which is the second level in the MMIRT model.

**Table 13**

*Countries included in the country-level latent classes of the MMIRT model*

| Country Level Latent Classes | |
|---|---|
| **Latent Class 1** | **Latent Class 2** |
| France | Hungary |
| Georgia | Türkiye |
| UAE | Italy |
| Norway | Portugal |
| Malaysia | Russian Federation |
| Finland | Israel |
| England | Lithuania |
| | Qatar |
| | Sweden |
| | Chile |
| | United States |
| | Chinese Taipei |
| | Hong Kong |
| | Korea Rep. of |
| | Singapore |

Of the 7 countries in country level latent class 1, five are located in Europe and two in Asia. Of the 15 countries in latent class 2, eight are located in Europe, five in Asia and two in the Americas. Table 14 shows the mean scores and standard deviations for the manifest group model by gender.

**Table 14.**

*Manifest group model mean scores and standard deviations by gender*

| Gender | Mean Scores | Standard Deviations |
|---|---|---|
| Female | 7.21 | 3.05 |
| Male | 7.46 | 3.25 |

When the averages for the gender variable are analyzed, the averages of male students are higher than those of female students. In the MH method according to the manifest groups, analysis was made in the context of gender variable and the results obtained are shown in Table 15.

**Table 15.**

*MH test results according to gender variable*

| Item | Chi Square | Alpha MH | Delta MH | Effect Size |
|---|---|---|---|---|
| 1 | 4.62* | 1.12 | -.26 | A |
| 2 | 12.54*** | .82 | .46 | A |
| 3 | 63.20*** | .65 | 1.01 | B |
| 4 | 8.04** | .84 | .40 | A |
| 5 | 3.26 | 1.10 | -.22 | A |
| 6 | .02 | 1.08 | -.02 | A |
| 7 | 132 | 1.07 | -.16 | A |
| 8 | .03 | 1.01 | -.03 | A |
| 9 | 7.99** | .86 | .35 | A |
| 10 | .44 | 1.04 | -.09 | A |
| 11 | 16.15*** | 1.22 | -.48 | A |
| 12 | 12.27*** | .82 | .46 | A |
| 13 | 34.45*** | 1.35 | -.70 | A |
| 14 | 22.71*** | 1.29 | -.60 | A |
| 15 | 3.09 | .91 | .23 | A |

Annotation: ***:0.001, **:0.01, *:0.05 significance level. A: Negligible effect, B: Moderate effect, C: Large effect, indicates effect level magnitudes.

According to Table 15, it can be said that only item 3 shows DIF as a result of the MH test conducted according to the gender variable. If ΔMH>0, DIF is interpreted as DIF in favor of the focus group, ΔMH<0 as DIF in favor of the reference group, and ΔMH≅0 as no DIF (Holland & Thayer, 1986). Item 3, which had a moderate DIF effect size, displayed DIF in favor of the focal group of females. Other items show negligible level of DIF. Item 3 is a knowledge level item about finding another algebraic expression that is equivalent to an algebraic expression in algebra in mathematics. The results of the DIF analysis of the latent classes created based on item difficulties for the L0-G2 model in the MIRT model are shown in Table 16.

**Table 16.**

*Results of the MH test for the two latent classes in the MIRT model*

| Item | Chi Square | Alpha MH | Delta MH | Effect Size |
|---|---|---|---|---|
| 1 | 121.06*** | .42 | 2.03 | C |
| 2 | 344.65*** | .06 | 6.78 | C |
| 3 | 1341.28*** | .03 | 8.51 | C |
| 4 | 211.66*** | .30 | 2.80 | C |
| 5 | 8.17** | 1.23 | -.49 | A |
| 6 | 98.12*** | .49 | 1.68 | C |
| 7 | 27.03*** | .63 | 1.07 | B |
| 8 | 245.79*** | 3.55 | -2.98 | C |
| 9 | 131.14*** | 2.51 | -2.16 | C |
| 10 | 230.28*** | 3.82 | -3.15 | C |
| 11 | 319.63*** | 4.59 | -3.58 | C |
| 12 | 83.86*** | 2.40 | -2.06 | C |
| 13 | 273.79*** | 3.96 | -3.23 | C |
| 14 | 107.05*** | 2.29 | -1.95 | C |
| 15 | 50.40*** | .48 | 1.72 | C |

Annotation: ***:0.001, **:0.01, *:0.05 significance level. A: Negligible effect, B: Moderate effect, C: Large effect, indicates effect level magnitudes.

As a result of the MH test conducted for the two latent groups in the MIRT model, it can be said that all items except item 5 displayed DIF. While item 7 displayed level B DIF, the remaining 13 items displayed level C DIF. Item 5 is an algebra question at the knowledge domain and is about defining a curve with a positive slope belonging to the subject of algebra in mathematics.

The DIF analysis was evaluated at both student and country level using the MH tests. The student-level results of the MH method for the DIF analysis conducted through the item difficulty parameters in the latent groups obtained according to the L2-G2 model in the MMIRT model are given in Table 17 and the country-level results are given in Table 18.

**Table 17**

*Results of the DIF analysis for the student-level MMIRT model*

| Item | Chi Square | Alpha MH | Delta MH | Effect Size |
|---|---|---|---|---|
| 1 | 4.51* | .73 | .74 | A |
| 2 | 40.60*** | .38 | 2.28 | C |
| 3 | 63.46*** | .31 | 2.73 | C |
| 4 | 26.25*** | .35 | 2.49 | C |
| 5 | 22.17*** | 2.08 | -1.72 | C |
| 6 | 44.85*** | .38 | 2.27 | C |
| 7 | .63 | .87 | .33 | A |
| 8 | 1.93 | 1.29 | -.60 | A |
| 9 | 4.31* | .69 | .88 | A |
| 10 | 0.42 | .88 | .29 | A |
| 11 | 26.69*** | 2.52 | -2.17 | C |

**Table 17 (continued)**

| Item | Chi Square | Alpha MH | Delta MH | Effect Size |
|---|---|---|---|---|
| 12 | 96.14*** | .16 | 4.34 | C |
| 13 | 15.94*** | 1.94 | -1.56 | C |
| 14 | 15.59*** | .50 | 1.61 | C |
| 15 | .79 | .82 | .46 | A |

Annotation: ***:0.001, **:0.01, *:0.05 significance level. A: Negligible effect, B: Moderate effect, C: Large effect, indicates effect level magnitudes.

As a result of the MH test conducted for the students in the two latent groups at the first level of the MMIRT model, it was concluded that items 1, 6, 7, 8, 9, and 15 displayed DIF at a negligible effect level and items 2, 3, 4, 5, 6, 11, 12, 13, and 14 displayed DIF at a large effect level.

**Table 18.**

*Results of the DIF analysis for the country-level MMIRT model*

| Item | Chi Square | Alpha MH | Delta MH | Effect Size |
|---|---|---|---|---|
| 1 | 267.66*** | 11.32 | -5.70 | C |
| 2 | 176.89*** | 19.03 | -6.92 | C |
| 3 | 220.47*** | 7.79 | -4.82 | C |
| 4 | 112.22*** | 4.01 | -3.27 | C |
| 5 | 58.05*** | 2.43 | -2.08 | C |
| 6 | 140.92*** | 4.69 | -3.63 | C |
| 7 | 59.24*** | 2.76 | -2.38 | C |
| 8 | 7.50** | 1.34 | -.70 | A |
| 9 | 1.62 | 1.16 | -.34 | A |
| 10 | 25.25*** | .52 | 1.52 | C |
| 11 | 5.30* | .78 | .59 | A |
| 12 | 2.46 | .85 | .39 | A |
| 13 | 7.93** | 1.36 | -.72 | A |
| 14 | 10.14** | 1.49 | -.94 | A |
| 15 | 1.87 | 1.20 | -.44 | A |

Annotation: ***:0.001, **:0.01, *:0.05 significance level. A: Negligible effect, B: Moderate effect, C: Large effect, indicates effect level magnitudes.

As a consequence of the MH test conducted for the students in the two latent country groups at the second level of the MMIRT model, it was concluded that items 8, 9, 11, 12, 13, 14, and 15 displayed DIF at a negligible effect level and items 1, 2, 3, 4, 5, 6, 7, and 10 displayed DIF at a large effect level.

## Discussion

In this study, the differentiation of DIF according to manifest groups, latent classes and latent groups in which multi-levelness is taken into account was examined. For modeling both student-level and country-level data, a MMIRT model is defined. The model developed in the research utilizes the properties of IRT model, an unconstrained latent class model and a multilevel model. The first level of the model enables an opportunity to determine whether there are latent classes that differ in students' strategies for response to items. The second level of information can be used to uncover possible differences between latent classrooms in countries that may be due to curricular or pedagogical differences.

The amount of DIF items detected in both the MMIRT model and the MIRT model is higher than what would be expected from a traditional DIF analysis using manifest classes. This is because the latent group approach maximizes the differences between latent groups, resulting in a larger amount of DIF items and larger differences in item difficulties between latent groups (Samuelsen, 2005). This result is also consistent with previous research based on the use of MIRT models for DIF analysis (Cho & Cohen, 2010; Cohen & Bolt, 2005; Samuelsen, 2005).

It is seen that the amount of items with DIFs and DIF effect sizes obtained as a consequence of the analysis of the MIRT model and the MMIRT model do not exactly overlap. Standard error calculation formulas may not give accurate results in analyses in which single-level models are made within the independence assumption (Kline, 2016). For this reason, a multilevel analysis should be conducted before the analyses, and if there is multilevelness, the analyses should be conducted taking into account the multilevelness. Lee et. al. (2018) concluded that for class-specific ICC conditions, a MMIRT model is recommended instead of a single-level item response model for a clustered dataset with cluster size 20 and cluster amount 50. It was found that the same 5 items (items 2, 3, 4, 5, and 6) displayed DIF in the MIRT model and the MMIRT model. In addition, when compared with the results obtained from the DIF analysis according to the manifest groups, it was seen that only item 3 displayed DIF in all three analyses. As a consequence of the analysis based on gender, it is observed that one item displays DIF, and it becomes clear that making comparisons only according to the manifest groups is not appropriate and will lead to erroneous inferences. Uyar (2015) found that when the data are suitable for the MMIRT model, the power of the MMIRT is higher than determining DIF with manifest groups, and for this reason, when the appropriate model is used, it will be easier for experts to interpret the items displaying B and C level DIF and the reasons for the items to be biased will be determined more objectively. Finch and Finch (2013) stated that even if test takers are matched in terms of the latent trait measured, the school they attend (as a second-level variable) may lead to the presence of DIF.

When the MMIRT model is analyzed at the country level, the countries in latent class 2 generally consist of Asian countries that have achieved successful results in large-scale exams and some European countries that are also successful in these exams. In latent class 1, there are two Asian countries with low achievement levels, five European countries with moderate achievement levels. Singapore, Chinese Taipei, Korea Rep., Hong Kong and Russia, which are in latent class 2, constitute the top five in the countries participating in eTIMSS in terms of mathematics achievement. UAE and Georgia, which are in latent class 1, are in two of the last five places in mathematics achievement in eTIMSS countries. Similar results apply to the science tests. The key benefit of the MMIRT model is based on the hypothesis that the resulting latent classes represent discrete subpopulations and are not statistical artifacts of non-normality that may exist only by chance in the data (Bauer & Curran, 2003).

In this study, no multidimensional analysis was conducted due to the unidimensional structure of the data. Considering that data with multiple dimensions are frequently seen in real life situations, it is recommended to use multilevel and multidimensional MIRT models according to the data structure and it is thought that the results will be enriched. Within the scope of the research, only 2 PL models were analyzed. Analyses can be performed with 3 PL models including the effect of luck success, 4 PL models including the unlucky parameter, or simpler models (1 PL and Rasch model) and the results can be compared. In the study, data from math and science tests consisting of 15 items were used. The effect of increasing or decreasing the number of items and differentiating the selected courses can be examined. It is recommended that researchers who will conduct studies in this field should first meticulously apply preliminary analyses for the data structure, identify latent groups in accordance with the data structure and conduct DIF analysis. In addition, the results of the analysis conducted with mixture models in determining the source of DIF should be preferred even though it requires a more complex analysis because it provides more information than the results of the analysis conducted according to the manifest groups. Since there is no single correct method for determining DIF, it is recommended to apply more than one method in the studies and interpret the outputs accordingly. The duration of the analyses conducted with mixture models can be quite long depending on the dimensionality and level of the data, the selected model and quantity, and the number of items in the data. For this reason, it is recommended that the number of individuals and items should not be increased too much, but should not be set too low so as not to negatively affect the model parameters. If the parameter values are well outside the usual bounds, the analysis can be repeated by increasing the starts values. Increasing the initial values increases the time considerably. In addition, increasing the initial values slightly may not provide the desired improvement in the item statistics and the values may need to be increased further.

## Declarations

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

**Author Contribution:** Ömer DOĞAN: conceptualization, investigation, methodology, data curation, supervision, writing - review & editing. Burcu ATAR: conceptualization, methodology, writing - original draft, formal analysis, visualization.

**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 this study uses data shared with the public.

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

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

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

## References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Journal of Educational Measurement, 29, 67–91.

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

Aydemir, F. (2023). *PISA 2018 matematik ve fen bilimleri alt testlerinde değişen madde fonksiyonunun Rasch Ağacı, Mantel–Haenszel ve Lojistik Regresyon yöntemleriyle incelenmesi*. Unpublished master thesis, Gazi University, Ankara.

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. *Psychological Methods*, *8*(3), 338–363. https://doi.org/10.1037/1082-989X.8.3.338

Bayram, Ö. (2024). *Bir tutum ölçeği üzerinden Mantel–Haenszel ve sıralı lojistik regresyon yöntemlerine göre değişen madde fonksiyonu incelenmesi*. Unpublished master thesis, Kocaeli University, Kocaeli.

Büyüköztürk, Ş. (2018). *Veri analizi el kitabı: istatistik, araştırma deseni, SPSS uygulamaları ve yorum.* Ankara: Pegem Akademi.

Cho, S. J., (2007). *A multilevel mixture irt model for dif analysis*. Unpublished Doctoral Dissertation, University of Georgia.

Cho, S.-J., & Cohen, A. S. (2010). A multilevel mixture irt model with an application to dif. *Journal of educational and behavioral statistics*, *35*(3), 336–370. https://doi.org/10.3102/1076998609353111

Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2013). Markov chain Monte Carlo estimation of a mixture item response theory model. *Journal of Statistical Computation and Simulation*, *83*, 278–306.

Cho, S. J., Suh, Y., & Lee, W. Y. (2015). An NCME instructional module on latent dif analysis using mixture item response models. *educational measurement:* issues and practice.

Choi, Y. J., Alexeev, N. & Cohen, A. S. (2015) Differential item functioning analysis using a mixture 3-parameter logistic model with a covariate on the timss 2007 mathematics test, *International Journal of Testing, 15*(3), 239-253, DOI: 10.1080/15305058.2015.1007241

Cohen, A.S., & Bolt, D.M. (2005). *A mixture model analysis of Differential item functioning. Journal of Educational Measurement*, *42*(2), 133-148.

De Ayala, R. J., Kim, S.-H., Stapleton, L. M., & Dayton, C. M. (2002). *Differential item functioning: A mixture distribution conceptualization. International Journal of Testing, 2,* 243–276.

De Boeck, P., Cho, S.-J., & Wilson, M. (2011). Explanatory secondary dimension modeling of latent differential item functioning. *Applied Psychological Measurement, 35, 583–603.*

*Dras, L. (2023). Multilevel mixture irt modeling for the analysis of differential item functioning.* Unpublished Doctoral dissertation, Brigham Young University.

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

Finch, W. H., & French, B. F. (2012). Parameter estimation with mixture item response theory models: A monte carlo comparison of maximum likelihood and bayesian methods. *Journal of Modern Applied Statistical Methods*, *11*(1), 167-178.

Finch, W. H., & Hernández Finch, M. E. (2013). Investigation of Specific Learning Disability and Testing Accommodations Based Differential Item Functioning Using a Multilevel Multidimensional Mixture Item Response Theory Model. *Educational and Psychological Measurement*, *73*(6), 973–993. https://doi.org/10.1177/0013164413494776

French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement, 47*, 299-317.

Gurkan, G. (2021). *From OLS to multilevel multidimensional mixture IRT: A model refinement approach to investigating patterns of relationships in PISA 2012 data*. Unpublished Doctoral Dissertation, Boston, United States of America.

Holland, P.W. & Thayer, D.T. (1986). *Differential item performance and the Mantel-Haenszel procedure* (Technical Report No. 86–69). Princeton, NJ: Educational Testing Service.

Holland, P.W. & Thayer, D.T. (1988) Differential item performance and the Mantel-Haenszel procedure, in Wainer, H. and Braun, H.I. (Eds.): *Test Validity*, 129–145, Erlbaum, Hillsdale, NJ.

Jiao, H., & Chen, Y.-F. (2014). Differential item and testlet functioning. In A. Kunnan (Ed.), The Companion to Language Assessments (pp.1282-1300). John Wiley & Sons, Inc.

Jiao, H., Kamata, A., Wang, S. & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement, 49*, 82-100.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Kelderman, H., & Macready, G.B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement*, 27, 307-327.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Kristanjansonn, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A Comparison of four methods for detecting differential item functioning in ordered response model. *Educational and Psychological Measurement, 65*(6), 935-953.

Lee, W. Y., Cho, S. J., & Sterba, S. K. (2018). Ignoring a multilevel structure in mixture item response models: impact on parameter recovery and model selection. *Applied psychological measurement*, *42*(2), 136–154. https://doi.org/10.1177/0146621617711999.

Li, F., Cohen, A. S., Kim, S., & Cho, S. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, *33*(5), 353–373. doi: 10.1177/0146621608326422.

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

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2015). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847-862. doi:10.3758/BRM.42.3.847.

Mantel, N. & Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute*, *22*(4), 719–748.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist, 50*(9), 741.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195-215.

Muthén, L.K. & Muthén, B.O. (1998-2017). *Mplus User's Guide*. Eighth Edition. Los Angeles, CA: Muthén & Muthén.

Paek, I., & Cho, S.-J. (2015). A note on parameter estimate comparability: Across latent classes in mixture IRT modeling. *Applied Psychological Measurement, 39*(2), 135–143. https://doi.org/10.1177/0146621614549651

Raju, N.S. (1988). The area between two item characteristic curves, *Psychometrika*, *53*(4), 495–502.

Raju, N.S. (1990) Determining the significance of estimated signed and unsigned areas between two item response functions, *Applied Psychological Measurement*, *14*(2), 197–207.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models. Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks, CA: Sage Publications. Revelle, W. (2023). *Psych: Procedures for psychological, psychometric, and personality research. (Version 2.3.3).* https://cran.r-project.org/web/packages/psych/psych.pdf

Revelle, W. (2023). *Psych: Procedures for psychological, psychometric, and personality research. (Version 2.3.3).* https://cran.r-project.org/web/packages/psych/psych.pdf

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271-282.

Roussos, L. & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*(4), 355–371.

Samuelsen, K. M. (2005). *Examining differential item functioning from a latent class perspective*. Unpublished doctoral dissertation, University of Maryland, College Park.

Sırgancı, G. (2019). *Karma rasch model ile değişen madde fonksiyonunun belirlenmesinde kovaryant (ortak) değişkenin etkisi*. Unpublished doctoral dissertation, Ankara University, Ankara.

Sen, S. (2022). *Mplus ile yapısal eşitlik modellemesi uygulamaları*. Ankara: Nobel.

Sen, S., Cohen, A., & Kim, S.-H. (2020). A short note on obtaining item parameter estimates of IRT models with Bayesian estimation in Mplus. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *11*(3), 266-282. doi: 10.21031/epod.693719

Sen, S., & Toker, T. (2021). An application of multilevel mixture item response theory model. *Journal of Measurement and Evaluation in Education and Psychology*, *12*(3), 226-238. doi: 10.21031/epod.893149

Shealy, R. and Stout, W. (1993) A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF, *Psychometrika*, *58*(2), 159–194.

Thissen, D., Steinberg, L. & Wainer, H. (1988) 'Use of item response theory in the study of group differences in trace lines', in Wainer, H. and Braun, H.I. (Eds.): *Test Validity*, 147–169, Erlbaum, Hillsdale, NJ.

Toker, T. & Green, K. (2021). A comparison of latent class analysis and the mixture rasch model using 8th grade mathematics data in the fourth international mathematics and science study (timss-2011), *International Journal of Assessment Tools in Education 8*(4), 959–974

Unal, F. (2023). *Farklı oranlardaki kayıp verilere farklı atama yöntemleriyle veri atamanın madde tepki kuramına dayalı yöntemlerle değişen madde fonksiyonuna etkisinin incelenmesi*. Unpublished master thesis, Akdeniz University, Antalya.

Uyar, Ş. (2015). Gözlenen gruplara ve örtük sınıflara göre belirlenen değişen madde fonksiyonunun karşılaştırılması. Unpublished doctoral dissertation, Hacettepe University, Ankara.

Yalcin, S. (2018). Determining differential item functioning with the mixture item response theory. *Eurasian Journal of Educational Research 74*, 187-206

Zhang, Y. (2017). *Detection of latent differential item functioing (dif) using mixture 2pl irt model with covariate*. Unpublished doctoral dissertation. University of Pittsburgh. Pittsburgh

ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(2); 138-147

# Robustness of Computer Adaptive Tests to the Presence of Item Preknowledge: A Simulation Study

Hakan KARA* Nuri DOĞAN** Başak ERDEM KARA***

**Abstract**

Item preknowledge describes a scenario where candidates may have access to some of the test items prior to the test administration. This involves sharing test materials and/or answers and it is difficult to identify the individuals with item preknowledge or the shared materials of the test. Nevertheless, it is essential to investigate the 'item preknowledge' problem because it can significantly affect the validity of the test results. It is believed that traditional linear tests are more robust to this type of aberrant response behavior than adaptive tests. In this context, the aim of this study is to examine the effect of item preknowledge on computer adaptive tests and identify the conditions under which adaptive tests are most resistant to the item preknowledge. For this purpose, a Monte Carlo simulation study was performed and 28 different conditions were examined. The results of the study indicated that the EAP estimation method provided better measurement precision than ML over all conditions. When 2PL and 3PL IRT models were compared, it was observed that 2PL had higher precision at most of the conditions. However, when the aberrancy ratio increased and reached 20% for both individuals and items, 3PL outperformed the 2PL model and gave the best results with the EAP combination. The results were discussed in line with the literature on item preknowledge and CAT and implications for practitioners and further research were provided.

*Keywords:* item preknowledge, aberrant responses, computer adaptive tests, test security

## Introduction

Test scores from any assessment tool are used to obtain information about the proficiency level of the examinees. The main assumption in using these scores is that examinees' responses reflect their actual level of proficiency and are not influenced by factors other than the latent trait (Meijer, 1996; Wan & Keller, 2023). However, this assumption is often violated and some other factors such as lucky guessing, cheating, careless responding, creative responding and random responding (Meijer, 1996) are involved in the process. The mentioned undesirable factors may cause responses that are inconsistent with the respondent's ability level, and these unexpected responses are referred to as aberrant responses (Clark, 2010). Aberrant responses occur when the observed patterns of response do not align with the expected ones (Meijer, 1996; Meijer & Sijtsma, 2001) and they are commonly encountered in practical testing situations (Wan & Keller, 2023; Yen et al., 2012).

When aberrant responses are included in the testing process, the test score does not reflect the 'true' level of ability estimate (Magis, 2014). The validity of test scores may suffer from the inclusion of such responses, as they prevent test takers from demonstrating their accurate level of measured latent trait (Rios et al., 2017). Therefore, it is crucial to monitor test results to detect aberrant responses in order to reduce their negative impact on the validity of test scores (Tendeiro & Meijer, 2014; Wan & Keller, 2023).

* PhD Student., Hacettepe University, Faculty of Education, Ankara-Turkey, e-mail: hakankaraodtu@gmail.com, ORCID ID: 0000-0002-2396-3462
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, e-mail: nurid@hacettepe.edu.tr, ORCID ID: 0000-0001-6274-2016
*** Assist. Prof. Dr., Anadolu University, Faculty of Education, Eskişehir-Turkey, e-mail: basakerdem@anadolu.edu.tr, ORCID ID: 0000-0003-3066-2892

To cite this article:
Kara, H., Doğan, N., & Erdem-Kara, B. (2024). Robustness of computer adaptive tests to the presence of item preknowledge: A simulation study. *Journal of Measurement and Evaluation in Education and Psychology, 15*(2), 138-147. https://doi.org/10.21031/epod.1470949

*Received: 19.04.2024*
*Accepted: 6.06.2024*

Aberrant responses may arise from a variety of sources. Yen et al. (2012) classified examinee responses in a selected response test into three groups: (a) responses that reflect the examinee's true ability, (b) correct responses made by lucky guesses, and (c) incorrect responses due to anxiety, carelessness, or distraction. The latter two types of response behavior are aberrant response types because they differ from what is expected and do not reflect the examinee's actual knowledge. Meijer (1996) proposed that there are at least five different factors that can cause aberrant responses: lucky guessing, cheating, careless responding, creative responding and random responding. In this paper the focus is on cheating behavior, specifically the item preknowledge.

## Theoretical Framework

Item preknowledge occurs when examinees may have access to test items prior to taking the exam (Eckerly, 2017). As Tendeiro and Meijer (2014) stated, cheating often enables an examinee to perform better than their actual ability and this could be a result of preknowledge of the items before the test. Belov (2016) stated that item preknowledge describes a scenario where a group of examinees (referred to as aberrant) have access to a subset of items (referred to as compromised items) prior to the administration of the test. Aberrant test takers exhibit improved performance on compromised items compared to non-compromised ones. As a result of item preknowledge, examinees unfairly get the right answers on test items that they would not normally get right. Thus, these items no longer effectively distinguish between examinees (Kim & Moses, 2016). As the percentage of compromised items and individuals with prior knowledge increases, the error in parameter estimates increases because the scores of aberrant examinees are invalid (Belov, 2016; Eckerly, 2017). Given the negative impact of item disclosure on test scores, item preknowledge should be investigated.

Item preknowledge could be defined as a special case of test collusion, which can be defined as the large-scale sharing of test materials or answers to questions. The shared information may come from different sources such as teachers, testing companies, the Internet or communication between examinees (Wollack & Maynes, 2017). It is hard to detect the aberrant examinees or items because there are multiple unknowns, such as the unknown group of cheating examinees accessing the unknown group of compromised items (Belov, 2016). However, it is essential to investigate since it affects the validity of the test results (Eckerly, 2017).

It is generally assumed that adaptive tests might suffer more from aberrant responses compared to traditional linear tests (Kim & Moses, 2016). Traditional linear tests are generally based on classical test theory (CTT), and the effect of aberrant responses on ability estimation in a traditional paper-pencil test might be little if items are equally weighted (Yen et al., 2012). However, item response theory models (IRT) are highly sensitive to these kinds of changes in response patterns (Magis, 2014). IRT models often struggle to accurately calculate true individual response probabilities due to various factors like guessing and cheating, leading to the presence of response disturbances, and IRT can return a strongly biased ability estimation when aberrant responses occur (Jia et al., 2019). As computer adaptive testing (CAT) applications are generally based on IRT models, they become more vulnerable to the biased estimation and measurement errors that aberrant responses may cause (Yen et al., 2012; Zheng & Chang, 2014). CAT is designed to select and administer items in accordance with test takers' proficiency level during the testing process. Ability estimation, whereas IRT models are used, is performed continuously after the administration of each item and the next item is selected based on the estimated ability (Yan et al., 2014; Zheng & Chang, 2014). Therefore, aberrant responses might not only cause the ability estimation error but also affect the item selection (Yen et al., 2012).

In a specific manner, CATs can be administered to small groups of test takers at different times, frequently and consecutively. This approach is referred to as continuous testing and offers flexibility in test scheduling. However, as with other forms of continuous testing, it can raise concerns about test

security. Examinees who have taken the test earlier could share information about the test with those who have taken it later, and many items are at risk of disclosure prior to the test. Students may memorize and share test information with others, and this may artificially inflate the scores of those who have obtained advanced knowledge of the material, therefore posing a threat to its validity (Zhang et al., 2012). CAT applications are, as a consequence, open to the threat of item preknowledge. Similar to its paper-pencil counterpart, CAT may award a higher score to a test taker as a result of his/her prior knowledge of the answers to compromised items. Unlike traditional paper-pencil tests, CATs customize each test for each individual examinee. If certain compromised items are answered correctly due to pre-existing knowledge of the answers, the CAT algorithm can recover the true ability through subsequent item selection based on factors such as the location and number of compromised items (Guo, 2009). CAT applications differ in several aspects, including item bank characteristics, item selection and stopping algorithms, exposure control and IRT model. These aspects influence the way in which compromised items affect test performance. Accordingly, it is important to see the performance of CAT applications under different conditions when item preknowledge exists.

The presence of compromised items is problematic for the reasons mentioned before. Several studies have been conducted on the performance of several detection methods of aberrant responses caused by item preknowledge in CAT environments (e.g., Belov, 2014; Liu, et al., 2019; McLeod et al., 2003; Pan et al., 2022; Qian et al., 2016). However, there is no single best way to detect item preknowledge and it is difficult to detect (Belov, 2014). Therefore, it is also important to understand the conditions that are more or less robust to the presence of item preknowledge. Several studies have been conducted on the impact of compromised items for different purposes. Yi et al. (2008) investigated compromised items under the'item theft' context in the CAT environment. They investigated the potential damage that item theft can cause on CAT under two item selection algorithms (maximum item information and a-stratified methods). The findings suggested that although 'item theft' could result in significant harm to CAT using either item selection approach, the maximum item information method was more susceptible to organized item theft simulation than the a-stratified method. In another study, Guo et al. (2009) investigated the resistance of CAT to small-scale cheating under different item selection methods and compared the results with a traditional paper-pencil test. They indicated that CAT is better at giving resistant results than conventional tests at the presence of small-scale cheating. Lengthier tests and more test forms provided much more secure conditions for conventional tests. In addition, six-item selection methods were compared and '*a*-stratified with *b* blocking' (ASBB) and maximum information (MI) methods had better resistance to small-scale cheating under 30 item test length but they gave similar results to the other four methods in the 60 item test. Lastly, Zhang et al. (2012) investigated the phenomenon of 'item preknowledge' under the name of 'item sharing' context and compared the use of single and multiple item pools under different item selection and exposure control methods. This study suggested that two-pool design provided a higher resistance to item sharing compared to single-pool designs resulting in greater precision in measuring ability using the Maximum Item Information Method with Sympson-Hetter item exposure control method. Although the mentioned studies demonstrated the serious and negative effects of the presence of compromised items, they were limited in some respects. The current study aimed to approach the problem from an expanded perspective by including the ability estimation method, the IRT model, the percentage of aberrant items and the percentage of aberrant individuals. Therefore, it is thought that the results of the study will contribute to the literature related to item preknowledge on CAT.

### The Purpose of the Study

For the reasons discussed above, it is important to understand the possible effects of the presence of item preknowledge, how CAT applications were affected and how the resistance of CAT changes under different conditions. Thus, the present study aims to investigate the performance of CAT in the presence of item preknowledge and to examine the conditions under which it is most resistant to prior knowledge. In this context, the following research question were addressed.

- How does the test performance of the CAT change in case of the presence of item preknowledge under different ability estimation methods (Maximum Likelihood (ML) and Expected a Priori (EAP)),

IRT model (2 PL and 3 PL), percentage of aberrant items (10%, 20% and 30%) and percentage of aberrant individuals (10% and 20%)?

## Methods

Within the scope of the research, it is aimed to investigate the effect of the inclusion of compromised items on the effectiveness of different CAT conditions. The data used in the research were generated by the simulation method using Monte Carlo approach and 28 different conditions were compared in a controlled manner. Simulation data was preferred because it is difficult to meet all the conditions discussed in the study simultaneously in real data.

### Design Overview

To demonstrate the effect of item preknowledge on CAT, normal response (no aberrancy) and item preknowledge conditions were simulated under different conditions (Ability estimation method; MLE and EAP – Percentage of aberrant items; 10%, 20% and 30% - Percentage of aberrant persons; 10% and 20% – IRT model; 2PL and 3PL). The test length was fixed at 30 items for each condition. All manipulated conditions were fully crossed with each other. There were a total of 24 conditions (3 aberrant item ratio x 2 aberrant person ratio x 2 ability estimation method x 2 IRT model) resulting from those manipulated conditions. In addition, response data of no aberrancy were generated for two ability estimation methods and two IRT models, resulting in four extra conditions. For each condition, 20 replications were executed. All procedures were carried out in R Statistical Software (v4.1.2; R Core Team, 2021)

### Data Generation

To see the effect of item preknowledge on CAT performance, we simulated normal responses and responses with item preknowledge for a 30-item test. Two item pools of 300 items were generated using 2PL and 3PL. Item difficulty parameters were generated based on a standard normal distribution N (0, 1) and item discrimination parameters were sampled from log-normal distribution L (0, 0.25). In addition to these, the c parameters were set at .20 (indicating a guessing parameter for a five-option multiple-choice test) for the 3 PL item pool. The ability parameters of 1000 examinees were randomly generated based on standard normal distribution N (0, 1). After the generation of ability parameters and item pools, normal response patterns of 1000 examinees on 300 items were generated as a base condition and CAT simulations were performed on that dataset (catR package; Magis et al., 2022).

For each condition, the ability level for the starting rule was set to '0' and the Maximum Fisher Information (MFI) method was used as the item selection method. MFI is one of the most commonly used methods for item selection in computer adaptive testing and was preferred because it selects the item that provides the maximum information each time tests (Wang, 2017; Weiss & Kingsbury, 1984). In order to avoid the same item being taken by each individual, the randomesque method was used with a five-item group. A value of 0.40 was used for item exposure.

While generating responses for item preknowledge behavior:

1. Firstly, items with the highest item exposure values were taken from the CAT simulation conducted under normal response conditions and these items were coded as compromised items.
2. The dataset was then updated for aberrant responders and items. Examinees with item preknowledge were randomly selected from individuals with low ability levels (th<0), and compromised items were randomly selected from the items identified in the previous stage.
3. Responses of specified individuals on those specified items were simulated based on the Bernoulli random variable with a success probability of .90.
4. The dataset generated in the first step was replaced with the newly generated aberrant dataset for aberrant individuals and items.

## Evaluation Criteria

In order to assess the impact of item preknowledge on CAT applications, RMSE, average bias, mean absolute error (MAE) and correlation values were calculated for each replication. Values were then averaged over 20 replications.

Root mean square error (RMSE) was calculated with the following formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{\theta_i}-\theta_i)^2}{N}} \quad (1)$$

Bias indicates the mean difference between individuals' true and estimated ability level and was calculated by using the following formula:

$$Bias = \frac{\sum_{i=1}^{n}(\widehat{\theta_i}-\theta_i)}{N} \quad (2)$$

Mean absolute error (MAE) represents the mean average difference between individuals' estimated and actual ability level and was calculated with the following formula:

$$MAE = \frac{\sum_{i=1}^{n}|\widehat{\theta_i}-\theta_i|}{N} \quad (3)$$

Lastly, correlation value was obtained by the following formula:

$$\rho_{\widehat{\theta_j},\theta_j} = \frac{cov(\widehat{\theta_j},\theta_j)}{\sigma_{\widehat{\theta_j}}\sigma_{\theta_j}} \quad (4)$$

$\hat{\theta}_j$ represents the estimated ability parameter, $\theta_j$ represents the true ability parameter, and N represents the total number of individuals. Besides, ($\sigma_{\hat{\theta}_j}$) and ($\sigma_{\theta_j}$) stand for the standard error values of the estimated and true ability parameters, respectively.

## Results

In the current study, the performance of CAT is investigated under different ability estimation methods, aberrant item ratio and aberrant person ratio. RMSE, bias, correlation and mean absolute error (MAE) values across all conditions are provided in Table 1. In addition, these values were visualized and presented in Figure 1. Results were interpreted with the help of both Table 1 and Figure 1.

According to Table 1 and Figure 1, the outcomes had the lowest RMSE, bias, MAE and highest correlation at the base (normal) condition regardless of estimation methods for both 2PL and 3PL models. The inclusion of compromised items reduced the measurement precision of the test, as expected, in both MLE and EAP conditions and 2PL and 3PL models. Comparing MLE and EAP estimation methods, EAP demonstrated the highest measurement precision (lowest RMSE, MAE and highest correlation values) across all conditions. In addition, the correlation between true and estimated ability values was high (>.871) across all conditions. However, it was higher for normal conditions compared to aberrant response conditions decreasing with the increasing percentage of aberrant items and persons. As the percentage of individuals with item preknowledge increased, the RMSE, bias, and MAE values increased and correlation decreased for both MLE and EAP estimation methods. Hence, increasing the percentage of aberrant responders resulted in a decline in measurement precision. The same situation held for the increment of aberrant item percentage as well.

**Table 1.**

*RMSE, Bias, Correlation and MAE Values under Different Conditions*

| | Ability Estimation | Aberrant Person | Aberrant Item | RMSE | Bias | Correlation | MAE |
|---|---|---|---|---|---|---|---|
| 2PLM | MLE | | Normal | 0.302 | -0.001 | 0.958 | 0.239 |
| | | 10% person | 10% | 0.329 | -0.032 | 0.949 | 0.260 |
| | | | 20% | 0.360 | -0.059 | 0.938 | 0.274 |
| | | | 30% | 0.427 | -0.090 | 0.914 | 0.306 |
| | | 20% person | 10% | 0.340 | -0.060 | 0.944 | 0.268 |
| | | | 20% | 0.409 | -0.118 | 0.922 | 0.312 |
| | | | 30% | 0.529 | -0.185 | 0.871 | 0.373 |
| | EAP | | Normal | 0.282 | -0.003 | 0.959 | 0.223 |
| | | 10% person | 10% | 0.308 | -0.026 | 0.952 | 0.244 |
| | | | 20% | 0.339 | -0.053 | 0.942 | 0.258 |
| | | | 30% | 0.389 | -0.073 | 0.924 | 0.280 |
| | | 20% person | 10% | 0.329 | -0.053 | 0.952 | 0.259 |
| | | | 20% | 0.390 | -0.101 | 0.927 | 0.296 |
| | | | 30% | 0.480 | -0.153 | 0.890 | 0.343 |
| 3 PLM | MLE | | Normal | 0.364 | -0.024 | 0.938 | 0.288 |
| | | 10% person | 10% | 0.395 | -0.018 | 0.925 | 0.306 |
| | | | 20% | 0.412 | -0.052 | 0.918 | 0.314 |
| | | | 30% | 0.461 | -0.096 | 0.899 | 0.339 |
| | | 20% person | 10% | 0.398 | -0.049 | 0.922 | 0.309 |
| | | | 20% | 0.459 | -0.117 | 0.899 | 0.340 |
| | | | 30% | 0.516 | -0.151 | 0.874 | 0.379 |
| | EAP | | Normal | 0.329 | -0.003 | 0.943 | 0.259 |
| | | 10% person | 10% | 0.355 | -0.004 | 0.934 | 0.283 |
| | | | 20% | 0.362 | -0.0277 | 0.932 | 0.288 |
| | | | 30% | 0.400 | -0.053 | 0.917 | 0.302 |
| | | 20% person | 10% | 0.355 | -0.014 | 0.934 | 0.280 |
| | | | 20% | 0.377 | -0.056 | 0.928 | 0.297 |
| | | | 30% | 0.436 | -0.102 | 0.905 | 0.329 |

Besides, 2PL model had a higher correlation and lower RMSE and MAE values for most of the conditions compared to 3PL model. 2PL with EAP estimation was the best combination at all aberrant items for 10% aberrant person and 2PL with ML combination was better than 3PL mostly. However, 3PL_EAP combination outperformed 2PL_MLE only for the 30% of aberrant item condition, whereas 2PL performed better in all other conditions (Figure 1). The result obtained can be interpreted as that EAP is more resistant to the increase of the percentage of preknowledge items than MLE. For the 20% aberrant person condition, 2PL_EAP had the highest correlation and the lowest RMSE again. But the difference here was that, as the percentage of aberrant items increased (≥.20), the performance of 3PL_EAP became better than 2PL model (Figure 1). This result, again, can be interpreted as the robustness of 3PL model and EAP estimation method to the high percentage of aberrant items and persons.

**Figure 1.**
*Correlation and RMSE Values under Different Conditions*


Overall, the main finding of the study is that the presence of item pre-knowledge impacts CAT test results and the severity of that impact depends on the percentage of the aberrant item and person. However, this impact is not even comparable with the base condition that item preknowledge was not present.

## Discussion

The purpose of the current study was to investigate the robustness of CAT results in the presence of item preknowledge under different conditions. We observed that the presence of compromised items was a threat to the performance of CAT because this presence led to a decline in the measurement precision of the CAT applications. The specific results were stated and discussed in the light of literature in this section.

Firstly, we observed that the increase in the percentage of aberrant items and persons resulted in a decrease in measurement precision as observed in the literature (Belov, 2016). Specifically, base (no aberrance) condition had the highest measurement precision (lowest RMSE, MAE and highest

correlation) and 20% aberrant person condition had the lowest precision, decreasing with the increment of aberrant item percent. It is an expected result since item preknowledge is an important threat for test scores and the number of compromised items and aberrant persons has an impact on the magnitude of this threat. Since CAT is mostly based on IRT models and these models are highly sensitive to the aberrancy of response patterns, ability estimations can be strongly biased (Jia et al., 2019; Kim & Moses, 2016; Magis, 2014).

The examination of the robustness of estimation methods to the presence of item preknowledge indicated that using EAP estimation method provided more measurement precision than MLE through all conditions. To our knowledge, there is no CAT-specific study comparing these estimation methods in the presence of aberrant responses. However, Kim & Moses (2016) compared different ability estimation methods at different aberrancy conditions in a multi-stage testing (MST) context, which is also adaptive. Consistent with the current study, they found that EAP yielded smaller RMSE than did MLE, especially at the highest and lowest ability regions under preknowledge condition.

Another result observed in our study is that 2PL model had higher measurement precision than 3PL at most of the conditions. 2PL with EAP ability estimation was the best combination for 10% person condition at all aberrant item percentages; but, 3PL-EAP combination outperformed the 2PL counterpart when aberrant person was 20% and the percentage of the compromised item was high (≥.20). It means that at a high amount of aberrancy, 3PL and EAP becomes more resistant to the threat of item preknowledge than 2PL. Although 2PL model is operationally used at large-scale testing programs such as GRE, TOEFL and PISA, it should be used carefully because of ignoring the 'guessing effect' (Hambleton, et al., 1991; Kim, et al., 2016). Haberman (2006) stated that the advantage of employing a 3PL model over a 2PL model seemed to be small, considering the much greater computational difficulties associated with 3PL. However, it should be carefully considered when 'guessing effect' is probably present. 'Guessing effect' causes individuals who do not know the answer to the question to answer the question correctly and it is one of the types of aberrant responses. In the current study, results indicated that 'guessing effect' had become more important at higher levels of item preknowledge (both item and person level). That is an expected result since item preknowledge poses an advantage for low-ability individuals and its impact increases with the level of aberrant persons and items. 'Guessing effect' also poses an advantage for the ones who do not have the knowledge to answer the question accurately. 2 PL model was able to compensate for the effect of both item preknowledge and guessing effect up to a certain point, but at some point, 3PL took the lead. Hence, the use of 3PL with EAP estimation method could be suggested especially in situations where item preknowledge is considered to pose a significant threat.

As a limitation of our study, we fixed the test length at 30 and used the Maximum Fischer Information method only as the item selection method. Further research studying different fixed or variable length conditions and different item selection methods can be conducted. Besides, different item exposure control methods can also be addressed in further studies. Additional research should be undertaken to examine different types of aberrant behaviors and under different conditions (such as item pool, ability parameters and number of individuals). Besides, the current study is limited to unidimensional IRT models. However, many educational and psychological tests are multidimensional (Ackerman, et al., 2003) and aberrant response behaviors may affect the statistical biases of the latent traits (Wang, 2015). Therefore, same problem considered in this research can be looked at in MIRT framework in further research.

## Declarations

**Conflict of Interest:** No potential conflict of interest was reported by the authors.
**Author Contribution:** Hakan KARA: conceptualization, investigation, methodology, writing - review & editing. Nuri DOĞAN: Conceptualization, methodology, supervision, writing - review & editing. Başak ERDEM-KARA: Conceptualization, methodology, data analysis, visualization, 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 data has been simulated in this study.

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

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

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

## References

Ackerman T., Gierl M. J., Walker C. M. (2003). Using multidimensional item response theory to evaluate educational psychological tests. Educational Measurement Issues and Practice, 22(3), 37–51. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x

Belov D. I. (2014). Detecting item preknowledge in computerized adaptive testing using information theory and combinatorial optimization. *Journal of Computerized Adaptive Testing*, 2(3), 37-58. https://doi.org/10.7333/jcat.v2i0.36

Belov, D. I. (2016). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement*, *40*(2), 83-97. https://doi.org/10.1177/0146621615603327

Clark, J. M. (2010). Aberrant response patterns as a multidimensional phenomenon: using factor-analytic model comparison to detect cheating. [Unpublished doctoral dissertation, University of Kansas]. ProQuest Dissertations and Theses Global.

Eckerly, C. A. (2017). Detecting item preknowledge and item compromise: Understanding the status quo. In G. J. Cizek & J. A. Wollack (Eds.), Handbook of detecting cheating on tests (pp. 101-123). Routledge.

Guo, F. (2009). Quantifying the impact of compromised items in CAT. In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. www.psych.umn.edu/psylabs/CATCentral/

Guo, J., Tay, L., & Drasgow, F. (2009). Conspiracies and test compromise: An evaluation of the resistance of test systems to small-scale cheating. International Journal of Testing, 9(4), 283–309. https://doi.org/10.1080/15305050903351901

Jia, B., Zhang, X., & Zhu, Z. (2019). A short note on aberrant responses bias in item response theory. Frontiers in Psychology, 10. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00043

Kim, S., & Moses, T. (2016). Investigating robustness of item response theory proficiency estimators to atypical response behaviors under two-stage multistage testing. ETS Research Report Series, 2016(2), 1–23. https://doi.org/10.1002/ets2.12111

Liu, T. , Sun, Y., Li, Z. & Xin, T. (2019) The impact of aberrant response on reliability and validity, measurement. Interdisciplinary Research and Perspectives, 17(3), 133-142, https://doi.org/10.1080/15366367.2019.1584848

Magis, D. (2014). On the asymptotic standard error of a class of robust estimators of ability in dichotomous item response models. British Journal of Mathematical and Statistical Psychology, 67(3), 430–450. https://doi.org/10.1111/bmsp.12027

Magis, D. , Raiche, G. & Barrada, J. R. (2022). catR: Generation of IRT response patterns under computerized adaptive testing (package version 3.17). https://cran.r-project.org/web/packages/catR

McLeod, L., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. Applied Psychological Measurement, 27(2), 121-137. https://doi.org/10.1177/0146621602250534

Meijer, R. R. (1996). Person-Fit research: An introduction. Applied Measurement in Education, 9, 3–8. https://doi.org/10.1207/s15324818ame0901_2

Meijer, R. & Sijtsma K. (2001). Methodology review: Evaluating person fit. Applied Psychological Measurement, 25(2), 107–135. https://doi.org/10.1177/01466210122031957

Pan, Y., Sinharay, S., Livne, O., & Wollack, J. A. (2022). A machine learning approach for detecting item compromise and preknowledge in computerized adaptive testing. Psychological Test and Assessment Modeling, 64(4), 385-424. https://doi.org/10.31234/osf.io/hk35a

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

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

Rios J. A., Guo H., Mao L., Liu O. L. (2017). Evaluating the impact of noneffortful responses on aggregated scores: To filter unmotivated examinees or not? International Journal of Testing, 17(1), 74–104. https://doi.org/10.1080/15305058.2016.1231193

Tendeiro, J. N., & Meijer, R. R. (2014). Detection of invalid test scores: The usefulness of simple nonparametric statistics. Journal of Educational Measurement, 51(3), 239–259. https://doi.org/10.1111/jedm.12046

Wan, S., & Keller, L. A. (2023). Using cumulative sum control chart to detect aberrant responses in educational assessments. Practical Assessment, Research and Evaluation, 28(2). https://doi.org/10.7275/pare.1257

Wang, K. (2017). A fair comparison of the performance of computerized adaptive testing and multistage adaptive testing (Publication No. 10273809). [Doctoral Dissertation, Michigan State University]. ProQuest Dissertations & Theses.

Weiss, D.J., & Kingsbury, G.G. (1984). Application of computer adaptive testing to educational problems. Journal of Educational Measurement, 21 (4), 361-375. https://doi.org/10.111 1/j.1745-3984.1984.tb01040.x

Wollack, J. A., & Maynes, D. D. (2017). Detection of test collusion using cluster analysis. In Handbook of Quantitative Methods for Detecting Cheating on Tests (pp. 124-150). Routledge.

Yan, D., von Davier, A. A., & Lewis, C. (2014). Computerized multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), Computerized Multistage Testing. CRC Press.

Yen, Y. C., Ho, R. G., Laio, W.W., Chen, L.J., & Kuo, C. C. (2012). An empirical evaluation of the slip correction in the four parameter logistic models with computerized adaptive testing. Applied Psychological Measurement, 36(2), 75–87. https://doi.org/10.1177/0146621611432862

Yi, Q., Zhang, J., & Chang, H. H. (2008). Severity of organized item theft in computerized adaptive testing: A simulation study. Applied Psychological Measurement, 32(7), 543-558. https://doi.org/10.1177/0146621607311336

Zhang, J., Chang, HH. & Yi, Q. (2012). Comparing single-pool and multiple-pool designs regarding test security in computerized testing. Behavior Research Methods, 44, 742–752. https://doi.org/10.3758/s13428-011-0178-5

Zheng, Y., & Chang, H.H. (2014). 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; 2024; 15(2); 148-165

# Investigating The Performance of Item Selection Algorithms in Cognitive Diagnosis Computerized Adaptive Testing

Semih AŞİRET *  Seçil ÖMÜR SÜNBÜL**

**Abstract**

This study aimed to examine the performances of item selection algorithms in terms of measurement accuracy and computational time, using factors such as test length, number of attributes, and item quality in fixed-length CD-CAT and average test lengths and computational time, using factors such as number of attributes and item quality in variable-length CD-CAT. In the research, two different simulation studies were conducted for the fixed and variable-length tests. Item responses were generated according to the DINA model. Two item banks, which consisted of 480 items for 5 and 6 attributes, were generated, and the item banks were used for both the fixed and variable-length tests. Q-matrix was generated item by item and attribute by attribute. In the study, 3000 examinees were generated in such a way that each examinee had a 50% chance of achieving each attribute. The cognitive patterns of the examinees were estimated by using MAP. In the variable-length CD-CAT, the first-highest posterior probability threshold is 0.80, and the second-highest posterior probability threshold is 0.10. The CD-CAT administration and other analyses were conducted using R 3.6.1.At the end of the study in which the fixed-length CD-CAT was used, it was concluded that an increase in the number of attributes resulted in a decrease in the pattern recovery rates of item selection algorithms. Conversely, these rates improved with higher item quality and longer test lengths. The highest values in terms of pattern recovery rate were obtained from JSD and MPWKL algorithms. In the variable-length CD-CAT, it was concluded that the average test length increased with the number of attributes and decreased with higher item quality. Across all conditions, the JSD algorithm yielded the shortest average test length. Additionally, It has been determined that GDI algorithm had the shortest computation time in all scenarios, whereas the MPWKL algorithm exhibited the longest computation time.

*Keywords: computerized adaptive testing, cognitive diagnosis models, item selection algorithms*

## Introduction

Monitoring students' learning situation and understanding their progress has a critical importance in educational sciences. Assessment is not only limited to measuring students' existing knowledge and skills; it also plays a vital role in guiding their learning processes and increasing their motivation. In this context, assessment should be recognized as an integral part of the educational processes. Stiggins (2002) also supports this perspective and emphasizes that assessment should not only reveal the current state of learning but also be used to improve learning. Assessment should present interpretative, diagnostic, highly informative, and predictive information (Pellegrino et al., 1999). However, in many studies (Bennett, 2011; Black & William, 2018; Heritage, 2010; William, 2011), it is reported that only the learning situation is supervised and the information that will facilitate the learning of examinees is not provided.

* PhD., Mersin University, Faculty of Education, Mersin-Turkey, semihasiret@gmail.com, ORCID ID: 0000-0002-0577-2603
** Assoc. Prof., Mersin University, Faculty of Education, Mersin-Turkey, secilomur@gmail.com, ORCID ID: 0000-0001-9442-1516

To cite this article:
Aşiret, S., & Ömür-Sübül, S. (2024). Investigating the performance of item selection algorithms in cognitive diagnosis computerized adaptive testing. *Journal of Measurement and Evaluation in Education and Psychology, 15*(2), 148-165. https://doi.org/10.21031/epod.1456094

*Received: 25.03.2024*
*Accepted: 21.06.2024*

Assessments based on Cognitive Diagnostic Models (CDM), which evaluate whether examinees have certain attributes, aim to provide descriptive feedback for each examinee rather than giving examinees subscale scores or summative scores. Therefore, diagnostic assessments provide detailed and useful information about each examinee's learning strengths and weaknesses and assess student achievement. This makes assessments based on cognitive diagnosis powerful and interesting, especially in areas where formative assessment is aimed and classroom assessments will be used.

CDMs are discrete latent variable models that enable the diagnosis of the operations required to solve a problem in a test or the presence or absence of many minor skills (de la Torre, 2009). Diagnoses obtained as a result of analysis with CDM can provide more details for a particular area and allow interventions that will bring solutions. With this model, a cognitive pattern can be produced for each examinee or group about whether the necessary skills or process steps are sufficient for a situation (Rupp & Templin, 2008).

CDMs are discrete latent variable models designed to diagnose the specific operations required to solve a problem in a test or to determine the presence or absence of various minor skills (de la Torre, 2009). Analyses conducted using CDMs can yield detailed diagnostic information for a particular domain and facilitate targeted interventions to address identified issues. This model allows for the generation of a cognitive profile for each examinee or group, indicating whether the necessary skills or process steps are sufficient for a given situation (Rupp & Templin, 2008).

The rise of computer technologies and their increasing accessibility for examinees has paved the way for the emergence of computerized adaptive testing (CAT) as a significant and popular research and application topic within psychometrics and education (Magis et al., 2017). Many CAT implementations position examinees along a latent continuum, but the need for individualized diagnostic feedback to examinees remains a challenge in this approach. Cognitive Diagnosis Computerized Adaptive Tests (CD-CAT) integrate Cognitive Diagnostic Models (CDM) with CAT methodologies. CD-CAT aims to classify examinees according to their latent status and apply latent class models to these latent classes (Cheng, 2009). More broadly, the primary aim of CD-CAT is to deliver individualized diagnostic feedback to examinees. Similar to CAT applications, CD-CAT encompasses the creation of an item bank, selecting the initial item to begin the test, estimating cognitive pattern, selecting subsequent items, terminating rules, estimating the final cognitive pattern, and reporting. However, item selection algorithms used in CAT are not suitable for CD-CAT. This is because CDMs operate with discrete latent variables, and algorithms such as Maximum Fisher Information (MFI) fail to make accurate predictions when the number of items is low. They are susceptible to the effects of chance success.

In recent years, numerous theories and algorithms related to CD-CAT applications have been developed (Cheng, 2009; Kaplan et al., 2015; McGlohen & Chang, 2008; Wang, 2013; Tatsuoka, 2002; Tatsuoka & Ferguson, 2003; Xu et al., 2003; Zheng & Chang, 2016). The item selection algorithms in the CD-CAT studies are primarily based on the Shannon Entropy (SHE) algorithm developed by Tatsuoka (2002) and Tatsuoka and Ferguson (2003), as well as the Kullback-Leibler (KL) information developed by Xu et al. (2003). However, Cheng (2009) utilized the Post-Weighted Kullback-Leibler information (PWKL) and Hybrid Kullback-Leibler information (HKL), while Wang (2013) employed Mutual information (MI) and Kaplan et al. (2015) used Modified Post-Weighted Kullback-Leibler information (MPWKL) and GDINA discrimination index (GDI). Additionally, Zheng and Chang (2016) developed the Post-Weighted Cognitive Discrimination Index (PWCDI) and the Post-Weighted Attribute-level Cognitive Discrimination Index (PWACDI), and Minchen and de la Torre (2016) introduced the Jensen-Shannon divergence (JSD) index.

The critical aspect of CD-CAT is the item selection algorithms (Cheng, 2009; Zheng, 2015; Zheng & Chang, 2016). Various item selection algorithms have been developed in CAT applications to cater to different needs. These algorithms are firmly established in CAT studies based on IRT. However, limited studies discuss item selection algorithms in the context of CD-CAT, as it is a relatively new field. Numerous factors, such as the number of attributes, structure of the Q matrix, item quality, termination rule, and estimation method, can influence the accuracy of results in these studies. Zheng (2015)

emphasizes that the primary goal of item selection algorithms is to achieve high measurement accuracy, which is also true for CD-CAT.

Many item selection algorithms have been developed for CD-CAT applications in recent years. However, most of these algorithms have not been evaluated under the same conditions. This study aims to examine various item selection algorithms in CD-CAT and compare them based on test length, number of attributes, item quality, and termination rule. By manipulating these factors at different levels in CD-CAT, the goal is to determine the item selection algorithms that provide the most accurate and maximum pattern recovery rates, computation time, and average test length.

## Method

### *Factors that Manipulated in the Research*

***Number of attributes*:** The number of attributes is one of the important factors affecting the accuracy of estimations in CD-CAT. Rupp and Templin (2008) stated that the number of attributes between 4 and 6 is moderate. In this study, since the number of attributes (K) was aimed at a medium level, K was manipulated to 5 and 6.

***Test length:*** There are two ways to handle test length: fixed-length and variable-length tests. DiBello and Stout (2007), as well as Wang (2013), argue that tests should be short to avoid wasting class time, especially since CD-CAT is mostly used in low-stakes tests and classroom assessments. In addition, tests in classroom assessments should be answered during the course time after each item is administered. Therefore, test lengths were manipulated to 5, 10, 15, and 20 in this study for the fixed-length tests.

***Item Selection Algorithms:*** Item selection algorithms play a crucial role in CD-CAT studies, according to Cheng (2009). Various item selection algorithms have been developed for CD-CAT studies. The fixed-length test for CD-CAT used item selection algorithms including KL (Xu et al., 2003), SHE (Tatsuoka, 2002), PWKL, and HKL (Cheng, 2009), MI (Wang, 2013), GDI, and MPWKL (Kaplan et al., 2015), PWCDI, PWACDI (Zheng & Chang, 2016), and JSD (Minchen & de la Torre, 2016). Meanwhile, the CD-CAT based on variable-length test used PWKL and HKL (Cheng, 2009), MI (Wang, 2013), GDI, and MPWKL (Kaplan et al., 2015), PWCDI, PWACDI (Zheng & Chang, 2016), and JSD (Minchen & de la Torre, 2016) item selection algorithms. Additionally, random selection was used as the base algorithm for all conditions to facilitate comparisons of other algorithms' performances.

***Item quality:*** The quality of the items was determined according to the discrimination index. In this study, item parameter distributions by Kaplan et al. (2015) were used. Therefore, the item parameters were generated from a uniform distribution. These item quality parameters are given in Table 1.

**Table 1.**

*Item parameters*

| **Item quality** | $\left[\mathbf{1}-\boldsymbol{P_j}(\mathbf{1})\right]$ & $\boldsymbol{P_j}(\mathbf{0})$ |
|---|---|
| LD-LV | U (0.15, 0.25) |
| LD-HV | U (0.10, 0.30) |
| HD-LV | U (0.05, 0.15) |
| HD-HV | U (0.00, 0.20) |

Note: LD= low discrimination, HD=high discrimination, LV= low variance, HV= high variance

***Termination Rule:*** CD-CAT studies use fixed and variable test lengths as termination rules. In this study, a two-criterion termination rule, suggested by Hsu et al. (2013), was used for variable-length CD-CAT. The first highest posterior probability threshold value was set at 0.80, and the second highest posterior probability threshold was set at 0.10. As the number of attributes increased, the number of cognitive patterns required would also increase exponentially to ensure that all items in the item bank were used. For this reason, the maximum test length was set to 40.

## Data Generation and Analysis

Within the scope of this study, R. 3.6.1 (R Core Team, 2020) carried out manipulating factors, data generation, and data analysis according to the levels of these factors.

***Generating the item bank and examinees:*** Creating the item bank involves generating the Q matrix and the parameters of the DINA model. In this study, the longest test includes 20 items in the fixed-length CD-CAT. The maximum test length was set to 40 items in the study using the variable test length termination rule. Stocking (1994) suggested that the item bank should be at least 12 times the test length (Cheng, 2009). Therefore, two separate item banks with a total of 480 items, consisting of 5 and 6 attributes, which were used for both fixed and variable-length CD-CAT were created. The Q matrix was developed item-by-item and attribute-by-attribute. To ensure equal representation of each attribute in the item bank and to make it applicable to real-world scenarios, data were generated so that each item had a 30% chance of measuring each attribute, and each item was required to measure at least one attribute. The data were generated so that there was no correlation between the attributes. The Q matrix contains $2^K - 1$ cognitive patterns. 3000 examinees were generated, each with a 50% chance of mastering each attribute, and common examinees were used for both studies. Based on the estimated item parameters and the Q matrix, the item responses of 3000 examinees and the probability of each examinee answering each item correctly according to the DINA model were computed.

Table 2 shows the distribution of the number of items measuring each attribute and the number of examinees with each attribute according to the number of manipulated attributes.

**Table 2.**

*Number of Items Measuring Each Attribute and the Number of Examinees with Each Attribute*

| | Attributes | | | | | |
|---|---|---|---|---|---|---|
| K=5 | 1 | 2 | 3 | 4 | 5 | |
| Number of items (J=480) | 174 | 184 | 170 | 175 | 169 | |
| Number of examinees | 1484 | 1462 | 1494 | 1454 | 1517 | |
| K=6 | 1 | 2 | 3 | 4 | 5 | 6 |
| Number of items (J=480) | 171 | 167 | 160 | 164 | 161 | 162 |
| Number of examinees | 1494 | 1476 | 1535 | 1500 | 1495 | 1510 |

In Table 3, the number of items measuring the possible number of attributes for 5 and 6 attributes in the item bank consisting of 480 items and the number of examinees with each attribute are given. In producing the Q matrix, each item was created to measure 30% of the attributes on average to be close to the real situation.

**Table 3.**

*The Number of Items Measuring the Possible Number of Attributes and the Number of the Examinees with the Attribute*

| Number of Attributes (K=5) | 0 | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|---|
| Number of Items (J=480) | 0 | 192 | 199 | 69 | 16 | 1 | |
| Number of Examinees | 108 | 476 | 948 | 923 | 455 | 90 | |
| Number of Attributes (K=6) | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| Number of Items (J=480) | 0 | 162 | 176 | 102 | 32 | 7 | 1 |
| Number of Examinees | 45 | 282 | 707 | 947 | 685 | 271 | 63 |

***Analysis Model:*** The DINA model is frequently preferred in simulation studies based on CDM and in low-stake tests due to the ease of parameter estimation and interpretation (Cheng, 2009; de la Torre, 2011; DeCarlo, 2011). Therefore, in this study, the DINA model was used.

***First item selection:*** CD-CAT starts with the first item selection. Within the scope of this study, the first item selection was made randomly and kept constant in other algorithms.

***Estimating the cognitive pattern:*** The Maximum Likelihood Estimation (MLE) method cannot estimate when examinees answer all items correctly or incorrectly (de Ayala, 2010). A similar situation applies to CDM studies. Test lengths can be short (e.g., five items), as CD-CAT studies are frequently conducted for classroom assessment. In such short-length tests, examinees' item response patterns are highly likely to be either all 0s or all 1s. Therefore, in this study, the cognitive patterns of examinees were estimated using the Maximum a posteriori (MAP) estimation method.

***Evaluation criteria:*** Within the scope of this study, the Pattern Recovery Rates (PRR) and computation time were used to evaluate the item selection algorithms for the fixed-length CD-CAT. For the variable test-length CD-CAT, PRR, computation time, and average test length were used to evaluate the performance of item selection algorithms.

For fixed-length CD-CAT, PRR is the rate of all correctly defined attribute patterns (Zheng & Chang, 2016). It refers to the proportion of examinees within the sample whose estimated cognitive pattern, $\hat{\alpha}_i$, is identical to their true cognitive pattern, $\alpha_i$, across all attributes. The higher PRR indicates greater classification accuracy. PRR is calculated by Equation 1.

$$\mathrm{PRR_k} = \frac{\sum_{i=1}^{N} R_i}{N} = \frac{\sum_{i=1}^{N}\left(I_{\hat{\alpha}_i,\alpha_i}\right)}{N}, \qquad (k=1, 2, \ldots., K) \qquad (1)$$

The computation times for the item selection algorithms were measured in seconds from the start of the process to estimate the first examinee's cognitive pattern until all examinees' cognitive patterns were estimated. The "*tictoc*" package (Izrailev, 2021) was used for this purpose. After calculating the time taken by all examinees, the total time was divided by the total number of examinees (in seconds) to get the average computation time for each examinee. This value was multiplied by 1000 for easier interpretation and reported as milliseconds per examinee. To calculate the relative average computation time of the item selection algorithms, it was divided by the computation time of the algorithm with the lowest average computation time by the computation time of the other algorithms.

For variable-length CD-CAT, the posterior probability of the cognitive pattern was used as the termination criterion instead of the fixed test length. After each selected item was administered to each examinee, the posterior probabilities of the cognitive patterns were estimated. In addition to the criterion that the highest posterior probability value is greater than 0.80 and the second highest posterior probability is less than 0.10 when the maximum number of items administered is 40, the test was terminated even if the posterior probability estimated for the examinee could not exceed 0.80. Therefore, these examinees were retained as examinees who did not complete the test. The estimated cognitive patterns of the examinees and the items used were recorded in the loop. After the loop was completed, minimum, maximum, and average statistics of the number of items used for each examinee were recorded for each item selection algorithm. In addition, the total number of examinees who could not complete the test was calculated. After these processes, the item selection algorithms' attribute and pattern recovery rates and average computation times were calculated. Finally, tables and graphs were produced using R 3.6.1. The "*ggplot2*" package (Wickham, 2016) was used to produce and edit the graphics.

## Findings

### *Fixed-length CD-CAT*

***Pattern recovery rates of item selection algorithms:*** The results of the pattern recovery rates of the item selection algorithms for fixed-length CD-CAT across various test lengths, item qualities, and number of attributes are presented in Table 4. These results are also graphically represented in Figure 1.

ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(2); 148-165

**Table 4**.
*Pattern Recovery Rates of Item Selection Algorithms in Fixed-Length CD-CAT*

| | | | Item Selection Algorithms | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| K | TL | IQ | Random | GDI | HKL | JSD | KL | MPWKL | MI | PWACDI | PWCDI | PWKL | SHE |
| 5 | 5 | LD-LV | 0,142 | 0,323 | 0,274 | 0,403 | 0,179 | 0,402 | 0,328 | 0,343 | 0,370 | 0,273 | 0,282 |
| | | LD-HV | 0,158 | 0,385 | 0,318 | 0,533 | 0,183 | 0,533 | 0,385 | 0,407 | 0,430 | 0,319 | 0,374 |
| | | HD-LV | 0,192 | 0,544 | 0,423 | 0,730 | 0,263 | 0,730 | 0,549 | 0,595 | 0,672 | 0,423 | 0,544 |
| | | HD-HV | 0,270 | 0,662 | 0,458 | 0,916 | 0,321 | 0,860 | 0,661 | 0,656 | 0,721 | 0,450 | 0,655 |
| | 10 | LD-LV | 0,242 | 0,621 | 0,599 | 0,641 | 0,300 | 0,640 | 0,619 | 0,595 | 0,612 | 0,584 | 0,608 |
| | | LD-HV | 0,322 | 0,727 | 0,713 | 0,759 | 0,307 | 0,758 | 0,733 | 0,690 | 0,724 | 0,711 | 0,729 |
| | | HD-LV | 0,405 | 0,902 | 0,881 | 0,920 | 0,480 | 0,921 | 0,904 | 0,888 | 0,906 | 0,881 | 0,908 |
| | | HD-HV | 0,500 | 0,986 | 0,956 | 0,990 | 0,688 | 0,991 | 0,985 | 0,974 | 0,983 | 0,959 | 0,988 |
| | 15 | LD-LV | 0,344 | 0,798 | 0,772 | 0,820 | 0,419 | 0,812 | 0,802 | 0,776 | 0,798 | 0,772 | 0,805 |
| | | LD-HV | 0,429 | 0,895 | 0,874 | 0,911 | 0,493 | 0,899 | 0,900 | 0,859 | 0,888 | 0,875 | 0,889 |
| | | HD-LV | 0,515 | 0,980 | 0,975 | 0,985 | 0,645 | 0,984 | 0,984 | 0,970 | 0,983 | 0,975 | 0,982 |
| | | HD-HV | 0,665 | 0,999 | 0,998 | 0,999 | 0,793 | 0,999 | 0,998 | 0,998 | 1,000 | 0,997 | 0,999 |
| | 20 | LD-LV | 0,462 | 0,899 | 0,881 | 0,912 | 0,511 | 0,909 | 0,906 | 0,867 | 0,897 | 0,887 | 0,895 |
| | | LD-HV | 0,529 | 0,957 | 0,949 | 0,956 | 0,627 | 0,961 | 0,953 | 0,938 | 0,954 | 0,948 | 0,959 |
| | | HD-LV | 0,568 | 0,996 | 0,997 | 0,998 | 0,756 | 0,997 | 0,998 | 0,993 | 0,996 | 0,997 | 0,997 |
| | | HD-HV | 0,775 | 1,00 | 1,00 | 1,00 | 0,885 | 1,00 | 1,00 | 0,999 | 1,00 | 1,000 | 1,00 |
| 6 | 5 | LD-LV | 0,088 | 0,189 | 0,164 | 0,204 | 0,108 | 0,206 | 0,189 | 0,201 | 0,184 | 0,161 | 0,188 |
| | | LD-HV | 0,086 | 0,222 | 0,210 | 0,261 | 0,116 | 0,254 | 0,222 | 0,225 | 0,235 | 0,199 | 0,210 |
| | | HD-LV | 0,131 | 0,341 | 0,237 | 0,360 | 0,140 | 0,362 | 0,339 | 0,321 | 0,353 | 0,237 | 0,34 |
| | | HD-HV | 0,134 | 0,398 | 0,259 | 0,435 | 0,269 | 0,426 | 0,413 | 0,376 | 0,424 | 0,391 | 0,388 |
| | 10 | LD-LV | 0,160 | 0,482 | 0,463 | 0,493 | 0,180 | 0,499 | 0,483 | 0,457 | 0,483 | 0,464 | 0,475 |
| | | LD-HV | 0,183 | 0,610 | 0,583 | 0,631 | 0,212 | 0,628 | 0,604 | 0,565 | 0,605 | 0,579 | 0,593 |
| | | HD-LV | 0,233 | 0,843 | 0,79 | 0,847 | 0,321 | 0,851 | 0,841 | 0,777 | 0,821 | 0,799 | 0,839 |
| | | HD-HV | 0,335 | 0,956 | 0,907 | 0,969 | 0,413 | 0,963 | 0,965 | 0,916 | 0,933 | 0,938 | 0,964 |
| | 15 | LD-LV | 0,255 | 0,687 | 0,656 | 0,688 | 0,276 | 0,691 | 0,686 | 0,645 | 0,669 | 0,658 | 0,674 |
| | | LD-HV | 0,292 | 0,806 | 0,778 | 0,821 | 0,346 | 0,807 | 0,802 | 0,744 | 0,788 | 0,777 | 0,786 |
| | | HD-LV | 0,389 | 0,959 | 0,942 | 0,960 | 0,478 | 0,962 | 0,955 | 0,923 | 0,951 | 0,943 | 0,962 |
| | | HD-HV | 0,495 | 0,997 | 0,992 | 0,996 | 0,673 | 0,997 | 0,997 | 0,992 | 0,994 | 0,994 | 0,996 |
| | 20 | LD-LV | 0,335 | 0,817 | 0,795 | 0,825 | 0,376 | 0,829 | 0,817 | 0,765 | 0,809 | 0,802 | 0,811 |
| | | LD-HV | 0,392 | 0,902 | 0,888 | 0,910 | 0,458 | 0,914 | 0,901 | 0,859 | 0,893 | 0,893 | 0,894 |
| | | HD-LV | 0,496 | 0,989 | 0,988 | 0,989 | 0,617 | 0,990 | 0,987 | 0,98 | 0,988 | 0,987 | 0,991 |
| | | HD-HV | 0,638 | 1,00 | 0,998 | 1,00 | 0,771 | 1,00 | 0,999 | 0,999 | 1,00 | 0,999 | 0,998 |

* PhD., Mersin University, Faculty of Education, Mersin-Turkey, semihasiret@gmail.com, ORCID ID: 0000-0002-0577-2603
** Assoc. Prof., Mersin University, Faculty of Education, Mersin-Turkey, secilomur@gmail.com, ORCID ID: 0000-0001-9442-1516

To cite this article:
Aşiret, S., Ömür Sübül, S. (2024). Investigating the performance of item selection algorithms in cognitive diagnosis computerized adaptive testing. Journal of Measurement and Evaluation in Education and Psychology, *15*(2), 148-165. https://doi.org/10.21031/epod.1456094

*Received: 25.03.2024*
*Accepted: 21.06.2024*

ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(2); 148-165

**Figure 1.**

*PRR of Item Selection Algorithms in Fixed-Length CD-CAT*

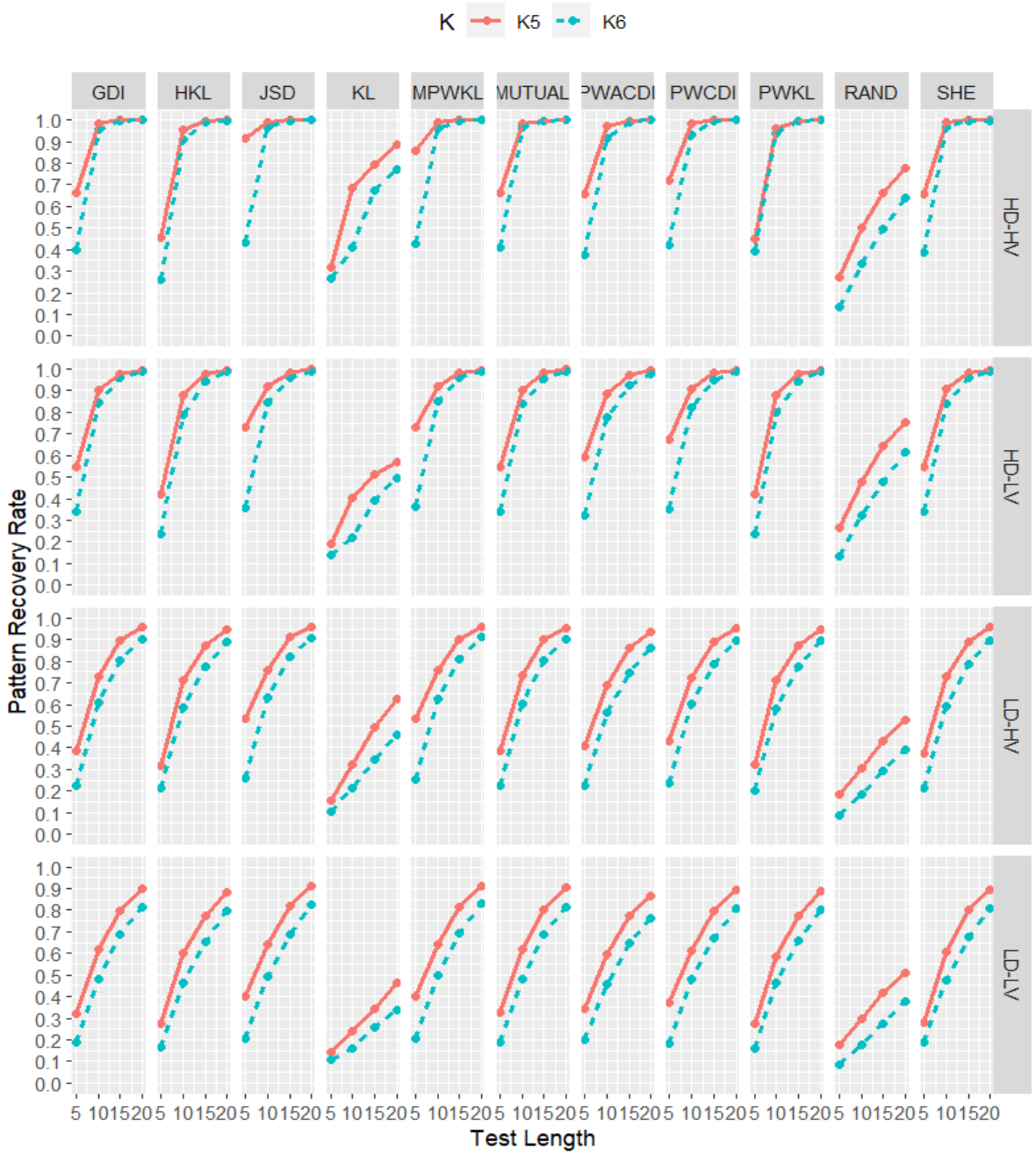



Table 4 and Figure 1 show that PRR for item selection algorithms increases significantly with increasing test length and item discrimination-item variance. Meanwhile, they decrease with an increasing number of attributes. Analysis of Figure 1 indicates that the increase in PRR is most pronounced when the test length is increased from 5 to 10, compared to other test lengths. At test lengths of 15 and 20, the rates for high item quality (HD-LV and HD-HV) are very close between 5 and 6 attributes, whereas, for low

* PhD., Mersin University, Faculty of Education, Mersin-Turkey, semihasiret@gmail.com, ORCID ID: 0000-0002-0577-2603
** Assoc. Prof., Mersin University, Faculty of Education, Mersin-Turkey, secilomur@gmail.com, ORCID ID: 0000-0001-9442-1516

To cite this article:
Aşiret, S., Ömür Sübül, S. (2024). Investigating the performance of item selection algorithms in cognitive diagnosis computerized adaptive testing. Journal of Measurement and Evaluation in Education and Psychology, *15*(2), 148-165. https://doi.org/10.21031/epod.1456094

*Received: 25.03.2024*
*Accepted: 21.06.2024*

item quality (LD-LV and LD-HV), the rates for 5 attributes are higher than those for 6 attributes. Additionally, at 20 test lengths with high item discrimination (HD-HV and HD-LV), the PRR of item selection algorithms is very close to 1 (0.99-1.00).

The lowest PRR was obtained by random selection, followed by the KL algorithm. Figure 1 further supports that these two algorithms performed worse than others. The highest PRR among item selection algorithms, under conditions of low item quality, 6 attributes, and a test length of 5, is 0.206 (MPWKL). For test length 5, the highest PRR was obtained with the JSD and MPWKL algorithms, with minimal differences. Generally, the highest PRR was achieved with the JSD and MPWKL across most conditions. Specifically, these algorithms outperformed others in short tests (5) with 5 attributes. For the HD-HV item quality level, the JSD algorithm's PRR is 0.916 for 5 test lengths and 5 attributes. The PWCDI algorithm follows JSD and MPWKL in the PRR for 5 test lengths and 5 attributes. However, the performance of PWCDI decreased with increasing test length, except for the HD-HV item quality.

The PWACDI algorithm consistently had a lower PRR after KL and random selection, except for test length 5. Similar results were observed for the GDI, SHE, and MI algorithms. In short tests with 5 attributes, MI performed better than SHE, whereas both gave similar results in longer tests. For 6 attributes, MI and GDI outperformed SHE. The PWKL and HKL generally provided similar results across different conditions, but their PRR was lower than those of MPWKL, JSD, GDI, MI, SHE, and PWCDI algorithms in most conditions.

***Average Computation Times of Item Selection Algorithms:*** The computation times of various item selection algorithms, considering different item qualities and numbers of attributes for 10 test lengths, were measured separately for each algorithm. These calculations were performed in milliseconds for a single examinee on a computer with an i7-7700HQ processor. The average computation times are presented in Table 5. Furthermore, Figures 2 and 3 show the relative average computation times of the item selection algorithms compared to the GDI algorithm for five and six attributes, respectively. The other algorithms' relative average computation times were calculated compared to the GDI because, after random selection, it consistently had the lowest average computation time under all conditions. Given the substantially lower PRR values of the random selection compared to other algorithms, it was excluded from consideration as a reference algorithm.

**Table 5.**

*Average Computation Time of Item Selection Algorithms for an Examinee at Fixed-Length CD-CAT (10 items, milliseconds)*

| K | Item Quality | Item Selection Algorithms | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Random | GDI | HKL | JSD | KL | MPWKL | MI | PWACDI | PWCDI | PWKL | SHE |
| 5 | LD-LV | 2,49 | 19,4 | 53,81 | 524,34 | 43,83 | 980,06 | 55,04 | 75,28 | 77,76 | 46,51 | 60,27 |
| | LD-HV | 2,57 | 20,92 | 57 | 514,09 | 46,73 | 984,6 | 60,12 | 84,39 | 84,32 | 49,49 | 63,54 |
| | HD-LV | 2,54 | 20,91 | 56,79 | 510,72 | 46,39 | 979,08 | 60,14 | 85,38 | 84,63 | 49,15 | 63,57 |
| | HD-HV | 2,72 | 20,81 | 56,49 | 518,86 | 46,25 | 980,41 | 60 | 84,41 | 84,2 | 49,39 | 62,99 |
| 6 | LD-LV | 4,03 | 26,56 | 75,85 | 901,22 | 63,49 | 1673,60 | 81,4 | 238,62 | 235,13 | 65,46 | 85,24 |
| | LD-HV | 4,21 | 26,68 | 76,52 | 894,85 | 64 | 1692,10 | 83,1 | 249,64 | 249,23 | 66,86 | 85,31 |
| | HD-LV | 3,98 | 26,58 | 75,75 | 892,6 | 63,46 | 1689,31 | 82,62 | 248,14 | 248,15 | 66,28 | 85,2 |
| | HD-HV | 4,32 | 27,29 | 77,37 | 899,6 | 64,78 | 1692,96 | 84,34 | 256,54 | 259,03 | 68,38 | 86,71 |

Table 5 shows that the GDI algorithm (19.4-27.29 ms) has a shorter average computation time than other algorithms, except for random selection. Figures 2 and 3 demonstrate that, at both attribute levels, the average computation time of the MPWKL and JSD algorithms is significantly higher than that of the other algorithms. The algorithm with the highest average computation time is MPWKL (980.06-1692.96 ms). When Figure 2 and Figure 3 are examined, the algorithms with the lowest relative average computation times at both quality levels are GDI, KL, PWKL, HKL, MI, SHE, PWACDI, JSD, and MPWKL, respectively.

**Figure 2.**

*Rates of Average Computation Times of Item Selection Algorithms to Average Computation Time of GDI for an Examinee with K= 5 for Fixed-Length CD-CAT*

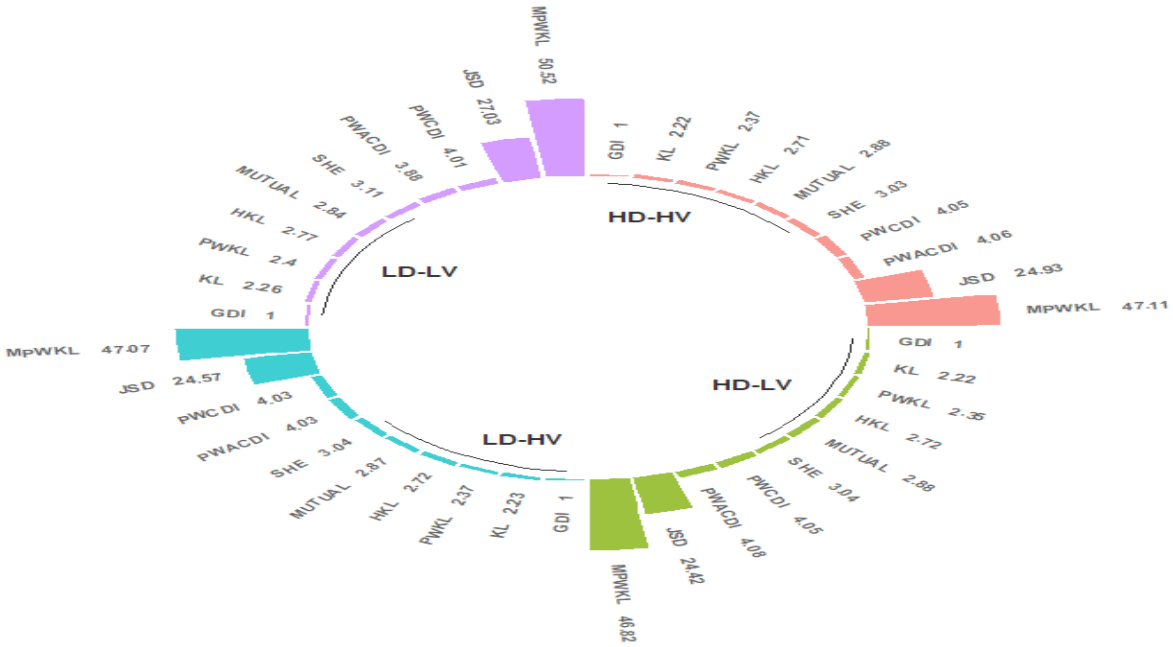
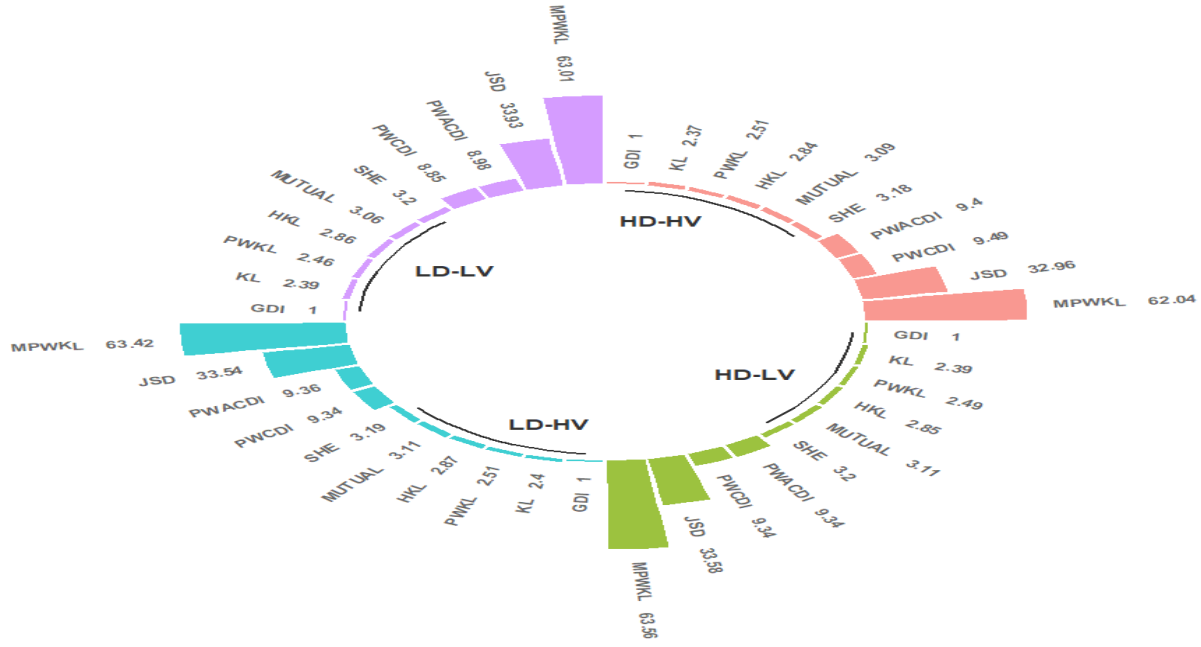



**Figure 3.**

*Rates of Average Computation Times of Item Selection Algorithms to Average Computation Time of GDI for an Examinee with K=6 for Fixed-Length CD-CAT*



According to Table 4, the GDI has the smallest proportional increase in computation time as the number of attributes increases. The relative average computation times for PWKL, HKL, KL, MI, and SHE compared to GDI show very small increases as the number of attributes rises. Specifically, the relative average computation times for JSD and MPWKL increased by approximately 1.35 times with the number of attributes, while PWCDI and PWACDI showed an increase of approximately 2.30 times. It can be said that PWCDI and PWACDI are more significantly affected by the increase in the number of attributes compared to other algorithms.

***Variable-Length CD-CAT***

***Average test length of item selection algorithms:*** Descriptive statistics for item selection algorithms at various item quality levels are given in Table 6, and average test lengths are graphically represented in Figure 4. In addition, the number of examinees who could not complete the test at different item quality levels is shown in Table 7.

**Table 6.**

*Descriptive Statistics of Item Selection Algorithms in the Variable-Length CD-CAT ($p_1$=0.80; $p_2$=0.10)*

| K | Item Quality | | Descriptive Statistics / Test Length | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | GDI | HKL | JSD | MPWKL | MI | PWACDI | PWCDI | PWKL |
| 5 | LD-LV | Min. | 6 | 4 | 6 | 6 | 6 | 6 | 6 | 4 |
| | | Max. | 40 | 40 | 40 | 40 | 40 | 40 | 40 | 40 |
| | | Average | 13,27 | 13,81 | 12,83 | 13,19 | 13,18 | 14,23 | 13,39 | 13,87 |
| | LD-HV | Min. | 6 | 4 | 6 | 6 | 6 | 6 | 6 | 4 |
| | | Max. | 34 | 35 | 31 | 34 | 33 | 38 | 30 | 36 |
| | | Average | 11,65 | 12,06 | 11,24 | 11,67 | 11,54 | 11,94 | 11,58 | 12,14 |
| | HD-LV | Min. | 5 | 3 | 5 | 4 | 5 | 5 | 5 | 3 |
| | | Max. | 20 | 17 | 21 | 23 | 23 | 32 | 21 | 18 |
| | | Average | 7,25 | 7,58 | 6,74 | 7,12 | 7,16 | 7,60 | 7,11 | 7,61 |
| | HD-HV | Min. | 4 | 2 | 5 | 4 | 4 | 5 | 5 | 2 |
| | | Max. | 12 | 19 | 5 | 12 | 9 | 15 | 11 | 20 |
| | | Average | 5,49 | 7,18 | 5,00 | 5,97 | 5,49 | 6,34 | 6,21 | 6,99 |
| 6 | LD-LV | Min. | 7 | 6 | 7 | 7 | 7 | 6 | 6 | 7 |
| | | Max. | 40 | 40 | 40 | 40 | 40 | 40 | 40 | 40 |
| | | Average | 16,70 | 17,22 | 16,40 | 16,69 | 16,73 | 17,88 | 16,78 | 17,24 |
| | LD-HV | Min. | 7 | 5 | 6 | 7 | 7 | 5 | 6 | 6 |
| | | Max. | 40 | 40 | 40 | 38 | 39 | 40 | 40 | 39 |
| | | Average | 13,87 | 14,01 | 13,51 | 13,87 | 13,84 | 14,65 | 13,85 | 14,19 |
| | HD-LV | Min. | 6 | 4 | 6 | 6 | 6 | 4 | 5 | 4 |
| | | Max. | 22 | 24 | 22 | 23 | 25 | 28 | 24 | 25 |
| | | Average | 8,62 | 9,30 | 8,31 | 8,60 | 8,55 | 9,29 | 8,71 | 9,38 |
| | HD-HV | Min. | 6 | 3 | 6 | 6 | 6 | 4 | 5 | 5 |
| | | Max. | 16 | 24 | 15 | 16 | 15 | 21 | 15 | 18 |
| | | Average | 6,82 | 8,08 | 6,47 | 7,06 | 6,80 | 7,56 | 7,33 | 7,89 |

**Figure 4.**

*Average Test Length of Item Selection Algorithms for Variable-length CD-CAT*

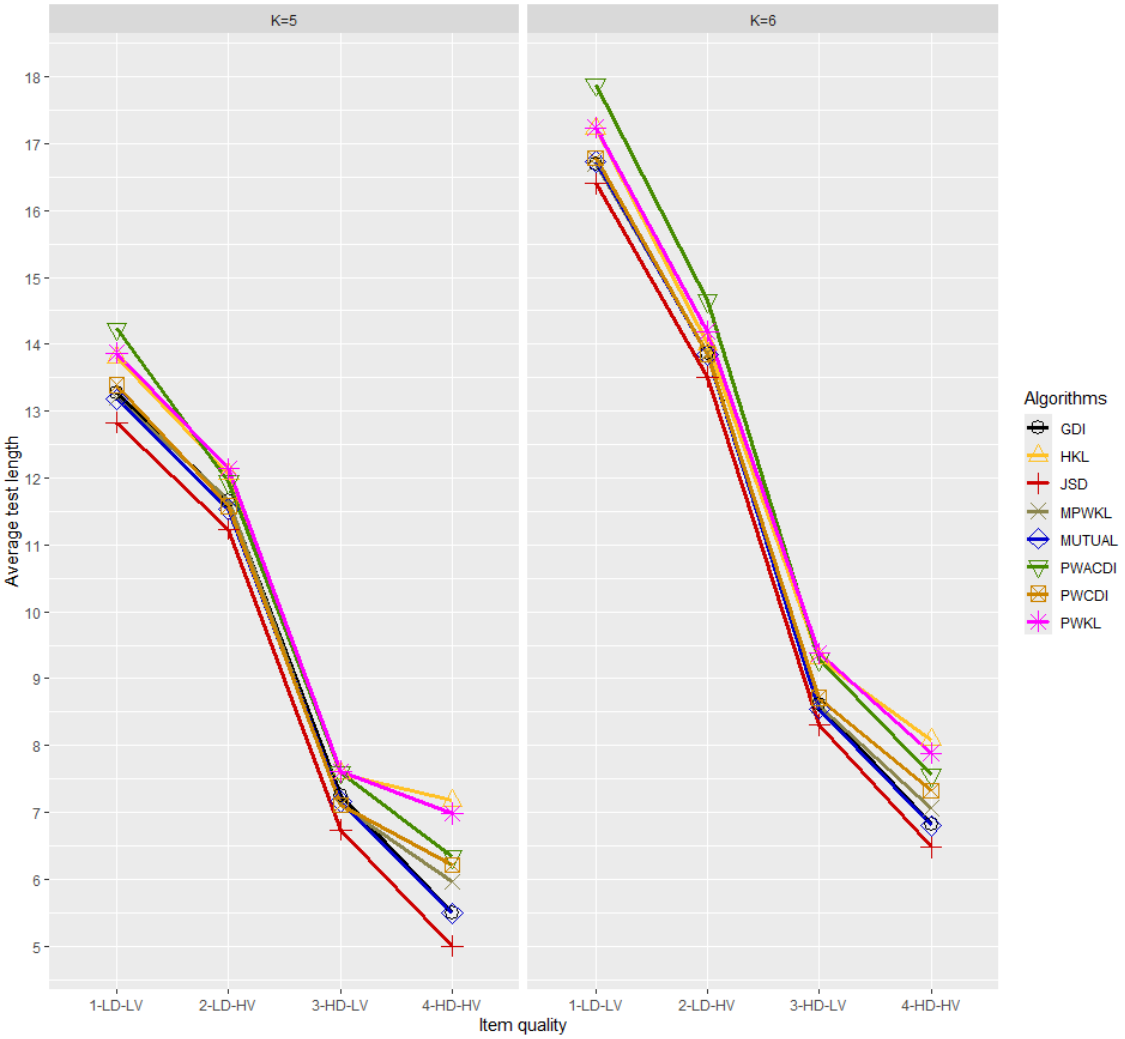



**Table 7.**

*Number of Examinees Who Could not Complete the Test in the Variable-Length CD-CAT (N=3000, $p_1$=0.80; $p_2$=0.10, Maximum Test Length=40)*

| K | Item Quality | GDI | HKL | JSD | MPWKL | MI | PWACDI | PWCDI | PWKL |
|---|---|---|---|---|---|---|---|---|---|
| 5 | LD-LV | 1 | 2 | 2 | 2 | 2 | 15 | 2 | 7 |
| | LD-HV | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | HD-LV | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | HD-HV | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | LD-LV | 13 | 12 | 8 | 11 | 7 | 51 | 21 | 13 |
| | LD-HV | 2 | 1 | 1 | 0 | 0 | 2 | 1 | 0 |
| | HD-LV | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | HD-HV | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

When analyzing Table 6 and Figure 4, it is seen that the average test lengths of the algorithms varied between 12.83 and 14.23 for the LD-LV item quality level at K=5 and from 16.40 to 17.88 at K=6. For

the LD-HV item quality level at K=5, the average test lengths ranged from 11.24 to 12.06 and from 13.51 to 14.65 at K=6. The average test lengths for HD-LV ranged from 6.74 to 7.61 at K=5 and 8.31 to 9.38 at K=6. For the HD-HV item quality level, the average test lengths varied between 5.00 and 7.18 at K=5 and 6.47 and 8.08 at K=6. Table 6 shows that the average test lengths of the algorithms increased with the number of attributes. Moreover, as the variance in item quality increased, the average test lengths of all algorithms also increased. However, the increase in item discrimination had a more significant impact on the average test lengths than the increase in variance in item quality. The PWACDI algorithm yielded the maximum average test length for items with low discrimination. When item discrimination increased, the HKL algorithm showed a higher average test length than PWACDI at K=5 and higher than HKL and PWKL at K=6. Particularly at the HD-HV level, the difference in average test lengths between these algorithms and others is more pronounced. The JSD algorithm produced the lowest average test length across all item quality levels. The average test lengths of MPWKL, GDI, and MI were similar for the LD-LV, LD-HV, and HD-LV item quality levels. However, at the HD-HV level, and for both K=5 and K=6, the average test length of MPWKL, GDI, and MI algorithms was longer. At K=6, the average test length of the PWCDI algorithm was close to that of MPWKL, GDI, and MI, except at the HD-HV level, where it was higher. At K=5, the average test length of PWCDI was higher than MPWKL, GDI, and MI at most item quality levels, except for HD-LV, which was very close to the average test lengths of these algorithms.

In Table 7, it is shown that some examinees could not complete the test at K=5 for the LD-LV item quality level, while all examinees completed the test for the other item quality levels according to the termination rule. Specifically, for the LD-LV item quality level, fifteen examinees in the PWACDI could not complete the test. Similarly, seven examinees in PWKL, one examinee in the GDI, and two examinees in other algorithms could not complete the test. At K=6, some examinees could not complete the test in any algorithm for the LD-LV item quality level. For the LD-LV, all examinees completed the test for the MPWKL, MI, and PWKL algorithms, while two examinees in the GDI and PWACDI algorithms and one examinee in other algorithms could not complete the test.

***Average Computation Times of Item Selection Algorithms:*** For the variable-length CD-CAT, the average computation times of the item selection algorithms were calculated for various item quality levels and numbers of attributes, similar to the fixed-length CD-CAT. Additionally, the ratio of the average computation time of each algorithm to that of the GDI algorithm is given in Table 8. The relative computation times are graphically represented in Figure 5 for K=5 and Figure 6 for K=6.

**Table 8.**
*Average Computation Time of Item Selection Algorithms for an Examinee at Variable-Length CD-CAT (10 items, milliseconds)*

| K | Item Quality | GDI | HKL | JSD | MPWKL | MI | PWACDI | PWCDI | PWKL |
|---|---|---|---|---|---|---|---|---|---|
| 5 | LD-LV | 102 | 294 | 675 | 1931 | 290 | 467 | 396 | 254 |
| | LD-HV | 95 | 226 | 584 | 1806 | 264 | 362 | 339 | 188 |
| | HD-LV | 51 | 146 | 360 | 1221 | 158 | 254 | 234 | 128 |
| | HD-HV | 38 | 211 | 335 | 1080 | 119 | 181 | 177 | 121 |
| 6 | LD-LV | 177 | 504 | 1446 | 4133 | 513 | 1370 | 1292 | 438 |
| | LD-HV | 133 | 360 | 1249 | 3723 | 422 | 1164 | 1016 | 323 |
| | HD-LV | 93 | 251 | 797 | 2448 | 280 | 772 | 749 | 222 |
| | HD-HV | 63 | 214 | 655 | 1941 | 199 | 588 | 581 | 184 |

In Table 8, we can see that the average computation times of the algorithms are similar when using a fixed-length CD-CAT. GDI is the fastest, while MPWKL is the slowest algorithm. The JSD algorithm has the second slowest average computation time, following MPWKL. The average computation time decreases as item quality increases, but it increases significantly with more attributes.

**Figure 5.**

*Rates of Average Computation Times of Item Selection Algorithms to Average Computation Time of GDI for an Examinee with K=5 for Variable-Length CD-CAT*

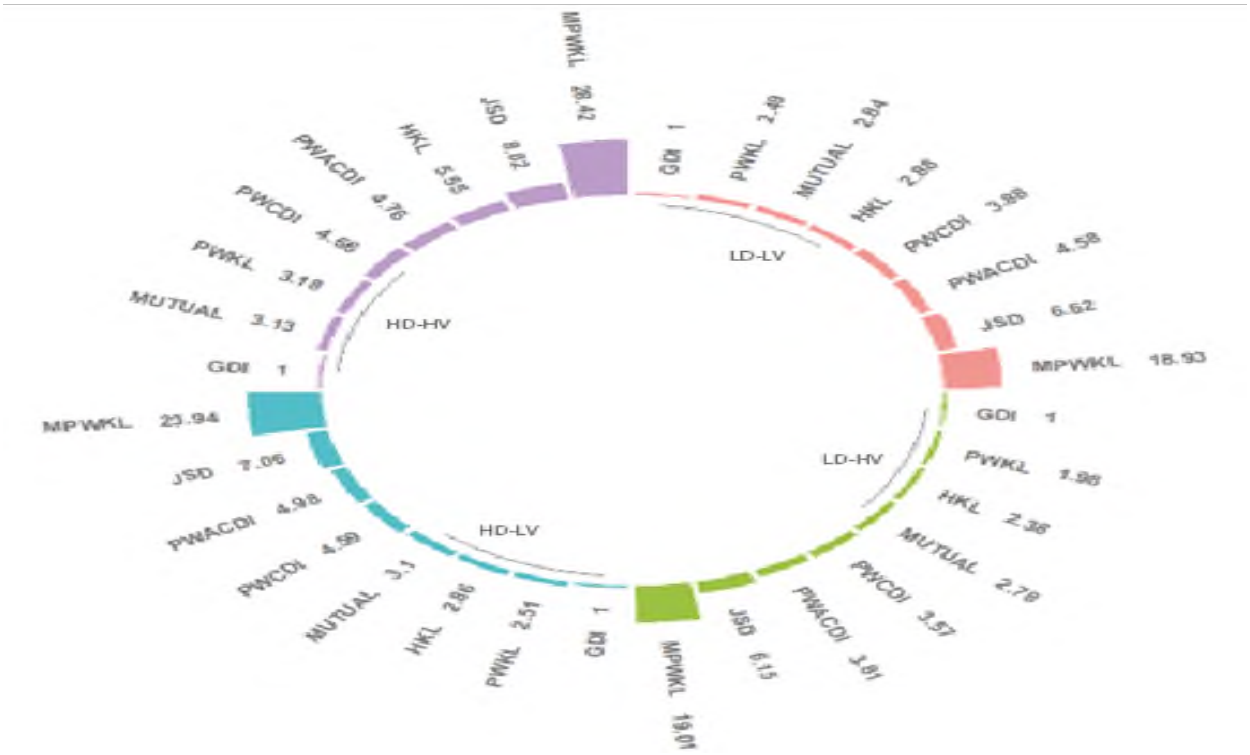
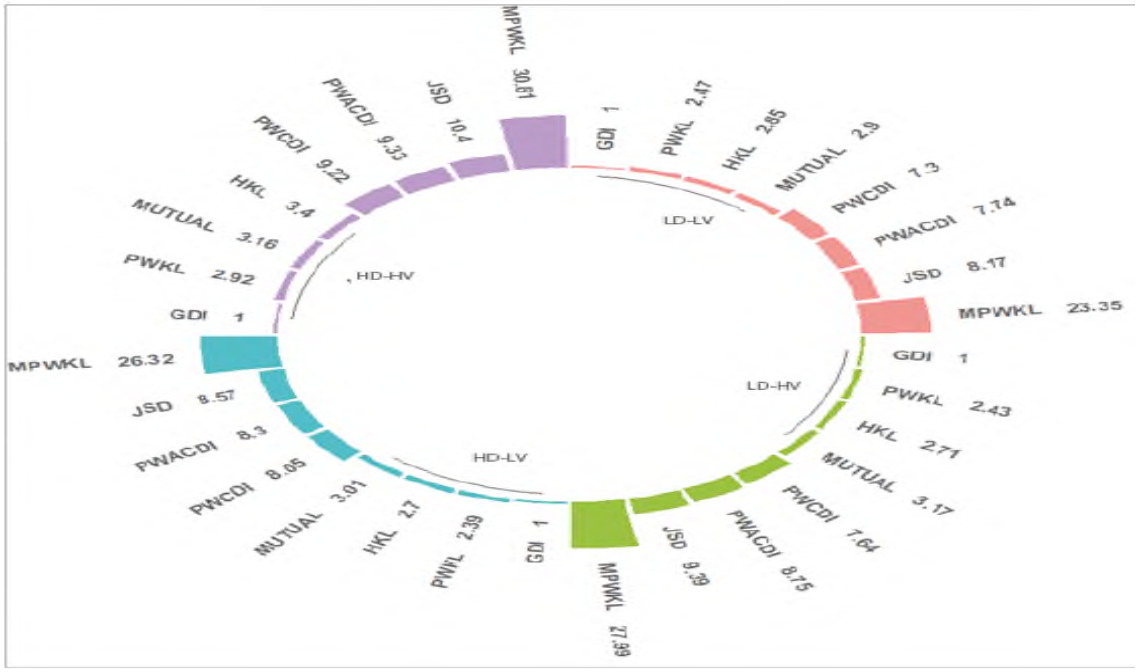



**Figure 6.**

*Rates of Average Computation Times of Item Selection Algorithms to Average Computation Time of GDI for an Examinee with K=6 for Variable-Length CD-CAT*




Figures 5 and 6 demonstrate that the GDI algorithm consistently has the lowest average computation time across all conditions. The increase in the relative average computation time of the HKL algorithm at the HD-HV level is more pronounced for 5 attributes. Additionally, the relative average computation

time of the HKL shows minimal variation with an increase in the number of attributes. When the number of attributes is 5, the JSD is approximately 7-9 times slower than the GDI. As the number of attributes increases, this ratio escalates to 8-10 times. The MPWKL is 19-24 times slower than the GDI algorithm at the 5 attribute level and 23-31 times slower at the 6 attribute level. Furthermore, as item discrimination, item variance, and the number of attributes increase, the relative average computation time of the MPWKL increases significantly. The MI computes 2.81-4.98 times slower than the GDI at 5 attributes, with only a slight increase in these ratios when the number of attributes rises to 6. The PWACDI algorithm is 4.42-4.70 times slower than the GDI algorithm at the 5 attribute level, with this ratio increasing to 7.74-9.33 when the number of attributes is 6. Similarly, for the 5 attribute condition, the relative average computation time of the PWCDI algorithm ranges from 3.88 to 4.66, while for the 6 attribute condition, these rates vary between 7.30 and 9.22. The PWKL algorithm, on the other hand, has a relative computation ratio ranging from 2 to 3 times at both the 5 and 6 attribute levels.

## Discussion, Conclusion and Recommendations

Nowadays, most psychometric studies focus on tests that measure one-dimensional latent attributes. These tests are often used in outcome-based assessments such as selection and placement. DiBello and Stout (2007) expressed the demand for measurement tools for formative assessments by teachers and education administrators in recent years. Giving fast and accurate feedback plays an important role in process evaluation to increase teaching effectiveness in classroom environments. In order to give effective and accurate feedback to the examinee, the strengths and weaknesses of the examinees must be determined accurately and properly. CDM can be useful in this context. However, giving quick feedback to examinees with a measurement tool developed based on CDM can be difficult due to time limitations in classroom environments. In this respect, the CD-CAT application provides convenience by giving quick feedback to the examinee as soon as possible.

In this study, two different simulation studies were carried out to examine the performance of item selection algorithms under various conditions. In the first simulation study, fixed-length CD-CAT, item selection algorithms through different item quality levels and attributes were evaluated regarding pattern recovery rates and average computation time. In the second simulation study, variable-length CD-CAT, the performance of item selection algorithms through various item quality and number of attribute levels was evaluated according to the average test length and computation times criterion.

In this study, the PRR of item selection algorithms decreased as the number of attributes increased. This is primarily due to the increase in the number of possible cognitive patterns as the number of attributes increases. For instance, with 5 attributes, there are 32 possible cognitive patterns, whereas this number increases to 64 with 6 attributes. Additionally, as item quality and test length increased, the PRR of the algorithms converged for both 5 and 6 attribute conditions. These findings are consistent with those reported by Wang (2013), Lin and Chang (2018), and Huang (2018).

The fixed-length CD-CAT study concluded that random selection is unsuitable for use since the pattern recovery rates of random selection are the lowest in all conditions. This finding is consistent with those reported in previous studies by Cheng (2009), Kaplan et al. (2015), Xu et al. (2003), Wang (2013), and Yigit et al. (2019). The primary reason for the consistently low PRR of the random selection across all conditions is that it does not consider the items' characteristics or the examinee's previous responses during item selection. Besides, it was also found that the attribute and pattern recovery rates of the KL are lower than those of other algorithms. Xu et al. (2003), Cheng (2009), and Zheng and Chang (2016) reported that the PRR of the KL is lower than that of other algorithms, except for the random selection. This study corroborates these results, confirming that the KL algorithm has the lowest PRR following random selection. The main reason is that the KL algorithm treats the probability of each cognitive pattern being the actual cognitive pattern as equal during the estimation process. In contrast, other algorithms adjust the weights of each cognitive pattern based on posterior probabilities after each item is administered, thereby providing more accurate estimations of the true cognitive pattern.

This study found that the PRR of the MI was higher than those of the SHE algorithm for tests with 5 attributes and shorter lengths. However, when the test length increased, and the number of attributes was 6, the PRR of these algorithms converged. These findings are consistent with those reported by Wang (2013). While the HKL and PWKL yielded similar PRR, the HKL generally exhibited a slightly higher PRR than the PWKL. Additionally, the PRR of the SHE and PWKL are very close to each other. These findings align with those reported by Cheng (2009).

In short tests, when the discrimination and variance values of the items were low, the PRR of the JSD and MPWKL were higher than the other algorithms. Kaplan et al. (2015) compared the correct classification rates of the MPWKL, GDI, and PWKL item selection algorithms at 10, 20, and 40 test lengths and across varying item quality levels. The results of that study indicated that the MPWKL and GDI achieved similar classification rates, whereas the PWKL demonstrated a lower classification rate than the MPWKL and GDI. Zheng and Chang (2016) analyzed the PRR of the MI, MPWKL, PWKL, KL, CDI, ACDI, PWCDI, and PWACDI algorithms for test lengths of 5 and 10. They found that the MPWKL, PWCDI, and PWACDI algorithms had the highest PRR, followed by the MI, CDI, ACDI, and KL algorithms. Yigit et al. (2019) compared the classification accuracy of JSD, GDI, and random selection algorithms under the MC-DINA model at test lengths of 5, 10, and 20, and under conditions of low and high discrimination-variance. They reported that the correct classification rates of the JSD were higher than those of the GDI and random selection in most conditions. The findings of this study are consistent with those reported by Cheng (2009), Kaplan et al. (2015), Wang (2013), Yigit et al. (2019) and Zheng and Chang (2016). However, in terms of measurement accuracy, it was observed that JSD and MPWKL could not measure with sufficient accuracy except for 5 test lengths and HD-HV level. In this respect, while using JSD and MPWKL algorithms with 5 test lengths and HD-HV levels can be recommended, longer tests are recommended for these algorithms in different item quality conditions.

In the fixed-length and variable-length CD-CAT studies, GDI had the lowest average computation time, while MPWKL had the highest. This is because, unlike other item selection algorithms, the computational complexity of GDI does not increase exponentially with the number of attributes. Zheng and Chang (2016) found that MPWKL and PWCDI had the longest computation times. The current study's average computation times for MPWKL, JSD, and PWCDI were higher than those for other algorithms. However, the findings related to the amount of time differ from those of Kaplan et al. (2015) and Zheng and Chang (2016). One possible reason could be that Kaplan et al. (2015) worked with a limited number of cognitive patterns, whereas this study used all possible cognitive patterns. Another reason could be that the cognitive pattern estimation method was used. Zheng and Chang (2016) used the MLE estimation method, while this study used the MAP method, which adds values for each cognitive pattern by multiplying the likelihood value with the prior probability value after each item is administered. Additionally, EAP estimation was performed within the CD-CAT process, and items administered and estimated cognitive patterns were recorded in a matrix after each item was administered, potentially affecting computation time. In this study, R 3.6.1 was used for statistical calculations. It is believed that software differences may influence the average computation time of the item selection algorithms.

However, considering measurement accuracy and the average computation times of the JSD and MPWKL, the JSD can be preferred primarily because it performs faster computation. Since item selection algorithms give more accurate results on 10 tests or more, it can be said that 10 test lengths are sufficient for classroom assessments for item banks consisting of items with high discrimination in practical studies. As item quality and test length increase, the classification accuracies of item selection algorithms are close to each other and approach 1. In this respect, when the measurement accuracy and computation time of the item selection algorithms are evaluated together, although the measurement accuracy of the GDI algorithm is slightly smaller than the JSD and MPWKL algorithms, it is recommended to be used in long tests and for item banks with high item discrimination, since the average computation time is faster. MI, SHE, PWKL, HKL, and PWCDI can also be used in long tests (20), and banks consist of items with high discrimination. Due to the decrease in measurement accuracy as the number of attributes increases, in practical applications, it is recommended to avoid very long attribute

numbers or to use longer tests and items with high discrimination in cases where the number of attributes is high.

In a comprehensive review of relevant literature, Kaplan et al. (2015) found that the average test lengths for the MPWKL and GDI were similar across all item quality levels in the variable-length CD-CAT study. However, the PWKL exhibited longer average test lengths than these two algorithms. In another study, Kaplan (2016) reported that the GDI algorithm had a lower average test length than the PWKL, with this difference becoming more pronounced as the number of attributes increased. Additionally, Zheng and Chang (2016) determined that under low item quality conditions, the PWCDI had the shortest average test length, followed by the PWACDI, MI, and PWKL, with MI and PWKL showing similar average test lengths. The shortest average test lengths were observed for the PWCDI and MI in high-item quality conditions, followed by the PWACDI and PWKL. Finally, Yiğit et al. (2019) reported that the JSD had a shorter average test length than the GDI under all conditions. In this study, the JSD consistently had the shortest average test length across all conditions. The average test lengths for the MPWKL, GDI, MI, and PWCDI were similar and slightly longer than those for the JSD. The PWACDI, HKL, and PWKL had longer average test lengths than the other algorithms. These findings are consistent with those reported in other studies within the related literature (Kaplan et al., 2015; Kaplan, 2016; Yiğit et al., 2019; Zheng & Chang, 2016).

In the variable-length CD-CAT study, it was concluded that an increase in item discrimination and variance in item quality results in a decrease in the average test length. Conversely, increasing the number of attributes leads to longer average test lengths. Due to the increase in average test length with a higher number of attributes, it is recommended to avoid an excessive number of attributes or to limit the maximum number of attributes measured by each item. At the HD-HV item quality level, the average test lengths of the algorithms range from 5 to 7 for K=5 and from 6 to 8 for K=6. Consequently, it is posited that classroom assessments with high-quality item banks will facilitate the effective utilization of CD-CAT. Although the JSD algorithm demonstrates the shortest average test length under all conditions, its average computation time exceeds that of other algorithms, except for MPWKL for low item quality levels. Therefore, it is recommended to utilize the JSD algorithm when item quality is high for short tests. When item quality is low, considering computation time, it is advisable to use the GDI and MI algorithms in addition to the JSD algorithm.

In this study, two criterion rules (Hsu et al., 2013) were used in variable-length CD-CAT. In this rule, the highest posterior probability value of the cognitive pattern was 0.80, and the second highest posterior probability value was 0.10. Hsu et al. (2013) suggested that these values should be considered as 0.90 and 0.05, respectively, in high-stake tests. Similar work can be performed using different posterior probability values. Moreover, the maximum test length limitation (40) was determined, as well as the posterior probability value. The performance of item selection algorithms can be examined by changing this value.

In this study, the DINA model was only utilized among the various cognitive diagnostic models. Similar studies can be performed again for different CDMs. In addition, since only the DINA model was used in the study, the Q matrix was developed only under this model. In practice, however, some datasets may fit different CDMs. For this reason, similar studies can be carried out for Q matrices consisting of mixed models.

The results of this study hold significant practical implications. The proposed algorithms are expected to guide future research and practical applications by facilitating the use of shorter tests and reducing the overall testing duration.

## Declarations

**Conflict of Interest:** The authors declare that they have no conflict of interest.

**Ethical Approval:** This study did not necessitate ethical approval as it utilized simulated data

## References


Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice*, 18(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, 25(6), 551–575. https://doi.org/10.1080/0969594X.2018.1441807

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74(4)*, 619–632. https://dx.doi.org/10.1007/S11336-009-9123-2

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

de la Torre, J. (2009). A cognitive diagnosis model for cognitively-based multiple-choice options. *Applied Psychological Measurement, 33(3),* 163–183. https://doi.org/10.1177/0146621608320523

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76, 179-199. https://doi.org/10.1007/S11336-011-9207-7

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement, 35(1),* 8–26. https://doi.org/10.1177/0146621610377081

DiBello, L., Roussos, L. A., & Stout, W. F. (2007). Handbook of Statistics. C. R. Rao ve S. Sinharay (Ed). *Review of Cognitively Diagnostic Assessment and a Summary of Psychometric Models.* 26, 979-1030. https://doi.org/10.1016/S0169-7161(06)26031-0

Heritage, M. (2010). *Formative assessment: Making it happen in the classroom*. Corwin Press, https://doi.org/10.4135/9781452219493

Hsu, C. L., Wang, W. H., & Chen, S. Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement, 37(7),* 563-582. https://doi.org/10.1177/0146621613488642

Huang, H. (2018). Effects of Item Calibration Errors on Computerized Adaptive Testing under Cognitive Diagnosis Models. *Journal of Classification*, 35:437-465. https://doi.org/10.1007/s00357-018-9265-y

Izrailev, S. (2020). *tictoc: Functions for Timing R Scripts, as well as Implementations of "Stack" and "StackList" Structures_*. R package
version 1.2.1, <https://CRAN.R-project.org/package=tictoc>

Kaplan, M. (2016). Nitelik Sayısının Madde Seçme Algoritmalarının Performansı Üzerindeki Etkisi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 7(2), 285-295. https://doi.org/10.21031/epod.268486

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

Lin, C.-J., & Chang, H.-H. (2019). Item Selection Criteria with Practical Constraints in Cognitive Diagnostic Computerized Adaptive Testing. *Educational and Psychological Measurement*, 79(2), 335–357. https://doi.org/10.1177/0013164418790634.

Liu, H., You, X., Wang, W., Ding, S., & Chang, H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of Classification*, 30, 152-172. https://doi.org/10.1007/s00357-013-9128-5

Magis, D., Yan, D., & von Davier, A. A. (2017). *Computerized adaptive and multistage testing with R: Using packages catR and mstR*. Cham, Switzerland: Springer International Publishing.

McGlohen, M.K., & Chang, H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavioral Research Methods, 40,* 808–821. https://doi.org/10.3758/BRM.40.3.808

Minchen, N. D., & de la Torre, J. (2016, July). *The continuous G-DINA model and the Jensen-Shannon divergence*. Paper presented at the International Meeting of the Psychometric Society, Asheville, NC.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "Two Disciplines" Problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education, 24(1),* 307–353. https://doi.org/10.3102/0091732X024001307

R Core Team (2021*). R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, 6(4), 219–262. https://doi.org/10.1080/15366360802490866

Stiggins, R. J. (2002). Assessment Crisis: The Absence of Assessment for Learning. *Phi Delta Kappan*, *83*(10), 758–765. https://doi.org/10.1177/003172170208301010

Stocking, M.L. (1994). *Three practical issues for modern adaptive testing item pools*. Ets research report series, 34. https://doi.org/10.1002/j.2333-8504.1994.tb01578.x

Tatsuoka, C., & Ferguson, T. (2003). Sequential classification on partially ordered sets. *Journal of Royal Statistics, 65*, 143–157.

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society, 51*, 337–350.

Thissen, D., & Mislevy, R. J. (2000). Computerized Adaptive Testing: A primer. H. Wainer, (Ed). *Testing algorithms*, Mahwah, NH: Lawrence Erlbaum Associates, Inc, p. 101-133.

Xu, X., Chang, H.-H., & Douglas, J. (2003, April). *Computerized adaptive testing strategies for cognitive diagnosis.* Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Quebec, Canada.

van der Linden, W.J., & Glas, G.A.W. (2002). *Computerized Adaptive Testing: Theory and Practice*. Kluwer Academic Publishers.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report 05-16). Princeton, NJ: Educational Testing Service.

von Davier, M., & Cheng, Y. (2014). Multistage testing using diagnostic models. İn D. L. Yan, A. A. von Davier & C. Lewis (eds.), *Computerized multistage testing: Theory and applications* (p. 219-227). New York, NY: CRC Press.

Wang, C. (2013). Mutual Information Item Selection Method in Cognitive Diagnostic Computerized Adaptive Testing with Short Test Length. *Educational and Psychological Measurement*, *73*(6), 1017–1035.

Wiliam, D. (2011). What Is Assessment for Learning? *Studies in Educational Evaluation*, 37, 3-14. https://doi.org/10.1016/j.stueduc.2011.03.001

Wickham, H., (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York,.

Yigit, H. D., Sorrel, M. A., & de la Torre, J. (2019). Computerized Adaptive Testing for Cognitively Based Multiple-Choice Data. *Applied Psychological Measurement*, *43*(5), 388–401.

Zheng, C. (2015). *Some practical item selection algorithms in cognitive diagnostic computerized adaptive testing—Smart diagnosis for smart learning.* Unpublished Doctoral Dissertation. University of Illinois at Urbana–Champaign.

Zheng, C., & Chang, H.-H. (2016). High-efficiency response distribution–based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement, 40*, 608-6

ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(2); 166-182

# The Effects of Missing Data Handling Methods on Reliability Coefficients: A Monte Carlo Simulation Study

Tugay KAÇAK* Abdullah Faruk KILIÇ**

**Abstract**

This study holds significant implications as it examines the impact of different missing data handling methods on the internal consistency coefficients. Using Monte Carlo simulations, we manipulated the number of items, true reliability, sample size, missing data ratio, and mechanisms to compare the relative bias of reliability coefficients. The reliability coefficients under scrutiny in this study encompass Cronbach's Alpha, Heise & Bohrnsted's Omega, Hancock & Mueller's H, Gölbaşı-Şimşek & Noyan's Theta G, Armor's Theta, and Gilmer-Feldt coefficients. Our arsenal of techniques includes single imputation methods like zero, mean, median, and regression imputation, as well as multiple imputation approaches like expectation maximization and random forest. We also employ the classic deletion method known as listwise deletion. The findings suggest that, for missing completely at random (MCAR) or missing at random (MAR) data, single imputation approaches (excluding zero imputation) may still be preferable to expectation maximization and random forest imputation, thereby underscoring the importance of our research.

*Keywords: missing data, reliability coefficients, missing data handling methods*

## Introduction

A common challenge in surveys and data collection processes is missing data, which can occur when respondents forget, skip, or choose not to answer one or more questions in a questionnaire. Especially for achievement tests or personality tests, sometimes test administrators suggest skipping the item if it is complex or confusing to respondents. Therefore, there are several reasons for missingness. Even the researchers do not care about the reason for missing data after the data collection process. However, the mechanism of missing data may affect the results of the analysis. There are three main mechanisms for missing data: i) Missing Completely at Random (MCAR), ii) Missing at Random (MAR), and iii) Missing Not at Random (MNAR) (Enders, 2010). So, mechanisms of missing data are the primary consideration when handling the missing data. In this study, we focused on MCAR and MAR because these mechanisms are more commonly encountered.

MCAR mainly points to randomness about missing data (Enders, 2010). As a formal definition, the probability of missing data about a variable is unrelated to other variables and itself. However, in MAR, the probability of missing data about a variable is related to other variables but unrelated to itself (Baraldi & Enders, 2010; Enders, 2010; Graham, 2012; Graham et al., 2013; Howell, 2007). The mechanisms of missing data may reduce the power of analysis or alter the distribution of the data set, which is another detail that can affect the results of statistical inferences. Even after adjusting for other factors, the likelihood of missing data on a variable is correlated with the values of that variable, which results in missing, not at random (MNAR). The degree of a patient's illness, for instance, may determine how likely they are to drop out of a clinical experiment, therefore affecting the missing data (Enders, 2010; Howell, 2007). MNAR presents substantial difficulties because it relies on unobserved values. Selection models and pattern mixture models provide methods for addressing missing not-at-random (MNAR)

*Research Assistant, Trakya University, Faculty of Education, Edirne-Türkiye, kacaktugay@gmail.com, ORCID ID: 0000-0002-5319-7148
**Associate Professor, Trakya University, Faculty of Education, Edirne-Türkiye, abdullahfarukkilic@gmail.com, ORCID ID: 0000-0003-3129-1763

To cite this article:
Kaçak, T., & Kılıç, A. F. (2024). The effects of missing data handling methods on reliability coefficients: A Monte Carlo simulation study. *Journal of Measurement and Evaluation in Education and Psychology, 15*(2), 166-182. https://doi.org/10.21031/epod.1485482

*Received: 17.05.2024*
*Accepted: 21.06.2024*

data, but they necessitate stringent assumptions. Conducting sensitivity analysis and actively collecting data are essential for dealing with complications arising from missing not-at-random (MNAR) data. So, we excluded the MNAR mechanism from the simulation study.

Many statistical programs offer limited methods to handle missing data. The most popular one is listwise deletion (LD). LD is an essential way to get a complete case by excluding the observations with empty cells. Complete case analysis is the primary choice. There are two common imputation approaches to obtain complete case: single imputation methods and multiple imputation methods. Mean, median, and regression imputation may be given as examples for single imputation (SI) methods. Random Forest imputation, classification and random trees, expectation-maximization may be given as examples for multiple imputation (MI) methods. It is known that imputation methods affect SEM fits (Fan & Wu, 2022; Li & Lomax, 2017), model parameters (T. Dai et al., 2024), growth curve parameters (D. Y. Lee et al., 2019), performances of factor retention methods (Goretzko, 2021; Goretzko et al., 2020), MANOVA results (Finch, 2016), the Rasch model statistics, Mokken's scalability coefficient H, Cronbach's Alpha (Roth et al., 1999; Sijtsma & Van Der Ark, 2003; Van Ginkel et al., 2007), and item parameters in IRT (S. Dai, 2021). However, there are limited studies focused on reliability coefficients except for Cronbach's Alpha. Therefore, in this study, we aim to compare the performance of reliability coefficients like Heise & Bohrnstedt's Omega, Hancock & Mueller's H, Gölbaşı-Şimşek & Noyan's Theta G and Armor's Theta, and also Cronbach's Alpha in terms of handling missing data methods. The present research was expected to elucidate the impact of handling missing data approaches on reliability coefficients.

**How to Handle Missing Data?**

Deletion methods can be performed either at the row or column level. Listwise deletion involves excluding individuals with missing data from the analysis, while column deletion entails not including variables containing missing data in the analyses (Enders, 2010; Little & Rubin, 2019; Schafer & Graham, 2002; Scheffer, 2002). While these approaches offer practical solutions to researchers, listwise deletion may introduce bias in analysis results due to sample reduction, and column-wise deletion may have a diminishing effect on content and construct validity. Therefore, contemporary approaches are progressing towards preserving both rows/lists (individuals) and columns (variables), utilizing techniques such as regression imputation between cells, expectation maximization (EM) for imputation, and considering variables' interrelationships, similarities among individuals in terms of measured characteristics, and various other statistical methods (Scheffer, 2002). All this literature raises this question: *"Which method for handling missing data provides more unbiased results for my survey?"*

In this study, we compared a deletion method (Listwise Deletion), single imputation methods ($Mean_{Person}$, $Median_{Item}$, Zero, Regression Imputation), and multiple imputation methods (Expectation Maximization and Random Forest Imputation). Deletion and single imputation methods were chosen because they are the default methods in most statistical software. Multiple imputation methods were also chosen because they are reported to give unbiased results in the current literature (Leite & Beretvas, 2010).

***Listwise Deletion***

The extent to which listwise deletion will cause problems in the data analysis process depends on the missing data mechanism. However, it is known to cause power loss of analysis in studies (Myers, 2011; Newman, 2014).

***Mean Imputation***

The mean imputation method (ME) involves replacing missing data in a row with the arithmetic mean of the non-missing values in that row or cell (Enders, 2010). The mean can be imputed by row or by column. In the case of categorical data, the row mean may be applicable in a continuous structure. In such cases, categorical observed variables should be treated as continuous, and analyses should be conducted accordingly. In this study, we used listwise mean imputation, which is actually named $Mean_{Person}$.

***Median Imputation***

Median imputation (MD) involves replacing missing data in a cell with the median of the values in the row or column where the missing data occurs (Zhang, 2016). In the case of categorical data, if the number of rows or columns is even, the imputed data may be treated as continuous. Therefore, observed variables should be considered continuous, and analyses should be conducted accordingly. In this study, we used column-wise median imputation which names as $Median_{Item}$.

***Zero Imputation***

Zero imputation involves imputing value of "0" to the cell with missing data (Wei et al., 2018). Clustering responses to variables to a single value can affect the variance, skewness, and kurtosis of the variable.

***Regression Imputation***

Regression imputation (RI) involves assuming linear relationships between variables and creating a regression equation. Using this equation, it predicts and imputes a value for missing cells (Enders, 2010). The predicted value is typically continuous.

***Expectation Maximization Algorithm***

Expectation Maximization (EM) algorithm was developed by Enders (2003) to mitigate information loss due to missing data. Essentially, it is an iterative method that performs imputation of the missing values. In the Expectation step, a series of regression equations are constructed to establish relationships between the variable with missing data and other variables. This process helps in developing estimated values for the empty cells. In the Maximization step, the covariance matrix is computed to minimize the residual variances after imputation. Iterations are repeated using the Maximum Likelihood method to minimize these residual variances each time. It relies on the assumption of multivariate normality of variables (Allison, 2002; Little & Rubin, 2002). This assumption is robust against some violations of categorical data. It is recognized that predictions may exhibit bias when dealing with Missing Not at Random (MNAR) as a missing data mechanism. Schafer and Graham's (2002) and Little and Rubin's (2002) study can be reviewed for further details.

***Random Forest Imputation***

Random Forest (RF) imputation relies on multiple regression and aims to impute missing data by creating a decision tree mechanism (Shah et al., 2014). It aims to predict missing cells by sampling randomly from the dataset. The Random Forest imputation method can be used for all types of variables, and unlike the EM method, it does not assume multivariate normality. It is known to perform well even for non-normally distributed datasets, and for more detailed information, studies by Doove et al. (2014) and T. Hayes & McArdle (2017) can be consulted.

Various studies have examined the effects of different approaches to handling missing data on various aspects of item response theory, such as item parameters (Finch, 2008), item difficulty and discrimination (Béland et al., 2018), methods for determining the number of factors (Goretzko, 2021), factor number and factor loadings (McNeish, 2017), results of confirmatory factor analysis (Lei & Shiverdecker, 2020), and reliability coefficients (Enders, 2004). This study aims to investigate the influence of different approaches to handling missing data on the reliability coefficients.

**Reliability Coefficients**

Reliability is one of the fundamental psychometric properties that need to be reported (Nunnally & Bernstein, 1994). While there are many different approaches to the concept of reliability, this study focuses on reliability coefficients in terms of internal consistency. Given the different assumptions they rely on, determining which reliability coefficient to report is essential for the reliability of research results. Among reliability coefficients, Cronbach's Alpha coefficient (Cronbach, 1951) is the most commonly reported in many social science research studies (Dunn et al., 2014; McNeish, 2018). However, Cronbach's Alpha coefficient is often reported without fulfilling its assumptions (Dunn et al., 2014). Despite many criticisms of using the Alpha coefficient, it continues to dominate among reported reliability coefficients (Edwards et al., 2021).

In order to properly utilize Cronbach's Alpha coefficient, many assumptions must be satisfied (Cronbach, 1951; A. Hayes & Coutts, 2020; McNeish, 2018). These are i) the presence of

unidimensionality, meaning that the items being measured are all connected to the same underlying construct; ii) the equality of factor loadings, also known as tau-equivalence, which implies that all items have equal relationships with the underlying construct; iii) the usage of continuous variables that follow a normal distribution; and iv) the assumption that error terms are uncorrelated. In order to calculate Alpha, it is essential to provide evidence of construct validity for unidimensionality, considering that psychological attributes are often multidimensional (McNeish, 2018). Furthermore, the equivalence of taus, or the parallelism of items in the scale, is linked to the degree to which items equally account for the measured attribute. However, in practice, it is more typical to have congeneric measurements, which means that the factor loadings are not equal. Research has shown that Alpha tends to underestimate the reliability of congeneric measures, as demonstrated by Edwards et al. (2021) and McNeish (2018). Other coefficients based on different assumptions than Alpha coefficient have also been developed, such as McDonald's Omega (McDonald, 1999), Heise & Bohrnstedt's Omega (Heise & Bohrnstedt, 1970), Hancock & Mueller's H (Hancock & Mueller, 2001), Armor's Theta (Armor, 1974), Gölbaşı-Şimşek & Noyan's Theta G (Gölbaşı-Şimşek & Noyan, 2013), and Gilmer-Feldt reliability coefficient (Feldt & Charter, 2003). When calculating reliability coefficients, estimations such as covariance matrices, variances, or factor loadings can be used. We compared the performance of Heise & Bohrnstedt's Omega, Hancock & Mueller's H, Gilmer-Feldt, Armor's Theta, Gölbaşı-Şimşek & Noyan's Theta G, and Gilmer-Feld coefficients in the current study. Table 1 provides the equations used to calculate the reliability coefficients examined in this study.

In order to obtain reliability coefficients in the presence of missing data, it is necessary to handle missing values in the cells where they occur through deletion or imputation approaches. Each missing data handling method may affect the values and parameters used to calculate reliability coefficients. The examination of how missing data handling methods affect reliability coefficients in datasets with missing data is becoming more critical and significant for researchers.

**Table 1.**
*Reliability Coefficients*

| Coefficients | Formula | Variables |
|---|---|---|
| Cronbach's Alpha (Cronbach, 1951) | $\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} s_i^2}{s_X^2}\right)$ | k: number of variables,<br>$s_i^2$: variance of each variable<br>$s_x^2$ : sum of variances of variables |
| Heise & Bohrnstedt's Omega (Heise & Bohrnstedt, 1970) | $\Omega = 1 - \frac{\sum_{i=1}^{n} \sigma_i^2 - \sum_{i=1}^{n} \sigma_i^2 h_i^2}{\sum_{i=1}^{n} \sum_{j=1}^{n} Cov(x_i, x_j)}$ | $h_i^2$ : communalities<br>x : variable |
| Hancock & Mueller's H (Hancock & Mueller, 2001) | $H = \frac{\sum_{i=1}^{k} \frac{l_i^2}{1-l_i^2}}{1+\sum_{i=1}^{k} \frac{l_i^2}{1-l_i^2}}$ | k : number of variables,<br>$l_i$ : i$^{th}$ items' standardized factor loading |
| Gölbaşı-Şimşek & Noyan's Teta G (Gölbaşı-Şimşek & Noyan, 2013) | $\theta_G = \frac{k}{k-m}\left(\frac{\sum_{i=1}^{m} \gamma_i - m}{\sum_{i=1}^{m} \gamma_i}\right)$ | m : number of factors<br>k : number of variables<br>$\gamma_i$ : i$^{th}$ eigenvalue |
| Armor's Teta (Armor, 1974) | $\theta = \left[\frac{p}{p-1}\right][1 - \left(\frac{1}{\lambda_i}\right)]$ | $\lambda_i$ : the largest eigenvalue obtained by principal component analysis,<br>p : number of variables |
| Gilmer-Feldt's coefficient (Feldt & Charter, 2003) | $r = \left[\frac{Q}{Q-W}\right](T/S_{tot}^2)$ | $S_{tot}^2$ : variance of total scores<br>$T$ : represents the sum of the last columns of the covariance matrix.<br>D should be calculated for Q and W. For D, the largest of the row sums of the covariance matrix is determined. The elements in this row are subtracted from the current row (e.g., ∑1 - A) and the largest row sum (∑h-A). D values are obtained by performing the process for each row. Q is calculated as the square of the sum of D's. W is calculated as the sum of the squares of each D. |

**Effects of Missing Data on Reliability Coefficients**

Most of the previous research on reliability coefficients was conducted with complete data. Enders' (2003) work considered incomplete data in 10 and 20 variables with unidimensional model. Factor loadings were λ = 0.55 and λ = 0.75. Data sets were generated with factor loadings, and the true reliability was calculated. Variables follow a normal distribution and are also categorical (3, 5, and 7). Missing mechanisms were used in three types (MCAR, MAR, and MNAR). Missing data ratios were at two levels (%15 and %30). EM, LD, PD, $Mean_{Column}$, and $Mean_{Person}$ were used to handle missing data. Only Cronbach's Alpha was analyzed in this study. Findings show that EM outperformed the other imputation methods, and LD and $Mean_{Column}$ may cause an underestimate of Cronbach's Alpha. Also, EM, LD, and Pairwise Deletion (PD) produced relatively high coverage rates.

In another study, Zhang & Yuan's (2016) work, Cronbach's Alpha and McDonald's Omega, were compared in terms of missing and outlying data. Listwise deletion and ML methods are used to handle missing data when tau equivalence is violated/is not violated. Findings show that listwise deletion causes the underestimate of Cronbach's Alpha and McDonald's Omega.

Also, Enders' (2004) study shows that EM is the best method to handle missing data for both missing data mechanisms (MCAR and MAR). The $Mean_{Person}$ method was the most negatively biased. The missing data ratio was 20%, and only Cronbach's Alpha was investigated.

In Sijtsma & Van Der Ark's (2003) study, predictive mean matching (PMM), two-way imputation (TW - $Mean_{Colum}$ and $Mean_{Person}$), response function (RF), and model-based response function (MRF) were used. Overall, PMM, TW, RF, and MRF resulted in slight biases on Cronbach's Alpha. The impact of imputation methods was not substantial.

Also, in the literature, there are several studies about the effects of missing data on reliability coefficients in real datasets (Parent, 2013; Şahin Kürşad & Nartgün, 2015). In real datasets, true reliability cannot be known; it can just be estimated. Therefore, this study adopts a simulation approach. Furthermore, this study is deemed significant for the following reasons: i) examining the impact of standard missing data handling methods found in statistical software on reliability coefficients, ii) investigating the effects of more contemporary and robust methods such as random forest, EM, as indicated by previous studies (T.

Hayes & McArdle, 2017; McNeish, 2017), alongside more straightforward methods on reliability coefficients, and iii) considering that the field of education primarily deals with categorical data, exploring the use of categorical variables and providing recommendations to researchers on which reliability coefficient to prefer in the presence of missing data.

The problem statement within the scope of the research is as follows: "In the examined simulation conditions, to what extent are the relative bias values of reliability coefficients:

- differ in complete datasets?
- differ in datasets with missing data?"

## Method

This study, conducted to determine how reliability coefficients are affected by different missing data handling methods, is a Monte Carlo simulation. Monte Carlo simulation studies involve generating data according to a specific distribution, analyzing the generated data using different methods, and comparing the results (Sigal & Chalmers, 2016).

### Simulation Conditions and Data Generation

This study investigates how reliability coefficients change according to missing data handling methods based on the number of items per factor, sample size, reliability level, missing data ratio, and missing data mechanism. The simulation conditions are presented in Table 2.

**Table 2.**
*Examined Conditions*

| | Simulation Factors | Simulation Conditions | Number of Conditions |
|---|---|---|---|
| Datasets with missing data | Sample size | 200, 1000 | 2 |
| | Missing data ratio | 5%, 10%, 20% | 3 |
| | Missing data mechanism | MCAR, MAR | 2 |
| | Test length | 8, 16 | 2 |
| | True reliability | 0.70, 0.90 | 2 |
| | | | 2x3x2x2x2 = 48 |
| | | | 1000 replications |
| Datasets without missing data | Sample size | 200, 1000 | 2 |
| | Test length | 8, 16 | 2 |
| | True reliability | 0.70, 0.90 | 2 |
| | | | 2x2x2=8 |
| | | | Total 56 conditions |
| | | | 1000 replications |

In addition to Table 2, eight conditions have been included for datasets without missing data, resulting in a total of 56 conditions. For each condition, 1000 replications were performed. Thus, 56,000 datasets were generated, and estimates were obtained for six different reliability coefficients from each dataset. Consequently, we conducted 336,000 different analyses. The fixed conditions of the study include a unidimensional structure and variables with five categories. Continuous datasets generated to exhibit multivariate normal distribution were transformed into categorical form using cutoff points as utilized by Uysal & Kılıç (2022)

Sample sizes of 200 and 1000 were added to the simulation conditions to represent small and large samples, respectively. These sample sizes are commonly preferred in many Monte Carlo simulation studies (Beauducel & Herzberg, 2006; Enders, 2004), facilitating the comparability of analysis results.

In this study, missing data rates of 5%, 10%, and 20% were considered. These rates have been examined in various studies on missing data (Cheema, 2014; H. J. Lee & Huber, 2021). It is noted that rates of 10% and above may lead to bias in the analyses (Bennett, 2001). Accordingly, the 5% rate represents a small proportion of missing data, the 10% rate represents the threshold where analyses may exhibit bias, and the 20% rate represents the expected level of missing data that may introduce bias in the analysis results.

The missing data mechanisms considered in the study are Missing Completely at Random (MCAR) and Missing at Random (MAR). In MCAR, missing data for a variable are unrelated to other variables (Allison, 2002; Enders, 2010; Graham, 2012; Little & Rubin, 2019). In MAR, there exists a relationship between the variable with missing data and other variables (Allison, 2002; Little & Rubin, 2019; Schafer & Graham, 2002). To ensure comparability with other studies in the literature, the MAR mechanism is included in this study.

Another condition included in the study is the number of items per factor. Brown (2006) suggests that there should be a minimum of 3 items per factor. However, there are studies suggesting that there should be at least 5 (Gorsuch, 2015) or at least 10 (Nunnally & Bernstein, 1994) items per factor. Since unidimensional structures are considered, the number of items per factor is set to 8 and 16 in this study. Thus, most of the recommendations regarding the number of items per factor in the literature are met. Additionally, this enables the examination of the impact of missing data handling methods on reliability coefficients in relatively short tests.

The reliability level represents the reliability coefficient in terms of internal consistency. In the literature, reliability coefficients of 0.70 and above are considered acceptable (McAllister & Bigley, 2002; Nunnally & Bernstein, 1994). Therefore, in this study, the true reliability conditions include the acceptable lower limit of 0.70 and the condition of 0.90, which can be interpreted as high reliability. The specified reliability levels, as also included in the study by Edwards et al. (2021), were obtained using McDonald's Omega coefficient (McDonald, 1970, 1999) with Formula 1:

$$\omega = r_{xx'} = \frac{(\sum_{i=1}^{k}\lambda_i)^2}{(\sum_{i=1}^{k}\lambda_i)^2+\sum_{i=1}^{k}(1-\lambda_i^{\,2})} \qquad (1)$$

In Formula 1, λ represents the factor loading of each item.

**Evaluation Criteria**

The relative bias (RB) values, which are based on the difference between the calculated reliability coefficients for each condition and the true reliability level, were calculated using Formula 2:

$$RB = \frac{\bar{r}_{xx'}-r_{xx'}}{r_{xx'}} \qquad (2)$$

Here, $\bar{r}_{xx'}$ represents the average reliability coefficient obtained from 1000 replications, while $r_{xx'}$ represents the true reliability level. As recommended by Flora & Curran (2004), we employed a criterion of |RB| < 0.10 to determine acceptable bias.

**Generating Datasets and Imputing Process**

The process of creating datasets with missing data was carried out using the "mice" package (van Buuren & Groothuis-Oudshoorn, 2011) based on the specified missing data ratios (5%, 10%, 20%) and missing data mechanisms (MCAR and MAR). After generating datasets with missing data, we employed the following imputation methods using relevant R packages. We employed the imputeTS package (Moritz & Bartz-Beielstein, 2017) for listwise deletion and zero imputation and the missMethods package (Rockel, 2022) for median imputation, mean imputation, and EM imputation. We employed the mice package (van Buuren & Groothuis-Oudshoorn, 2011) for random forest (RF) imputation with the default iteration setting as m = 5 and regression imputation.

The analysis of the data involved calculating Cronbach's Alpha coefficient for the complete datasets and datasets obtained through various missing data imputation methods using the psych package (Revelle, 2024). Additionally, we utilized the reliacoef package (Cho, 2023) to calculate Heise & Bohrnstedt's Omega, Gilmer-Field's reliability coefficient, and Hancock & Mueller's H coefficient. We wrote the custom R script to calculate Gölbaşı-Şimşek & Noyan's Teta G and Armor's Teta coefficients.

We utilized Pearson correlation matrices to calculate reliability coefficients from factor analysis for datasets imputed using mean, median and regression imputation methods, which resulted in continuous data.

## Findings

The one-way ANOVA results indicate that each simulation factor has a statistically significant effect on the RB values. The findings are summarized in Table 3.

**Table 3.**
*One-way ANOVA Results*

| | Factor | df | *F* | $\eta^2$ |
|---|---|---|---|---|
| Datasets with missing data | True reliability | 1 | 97.39* | 0.05 |
| | Test length | 1 | 284.42** | 0.12 |
| | Missing data mechanism and ratio | 5 | 23.84* | 0.06 |
| | Deletion/imputation methods | 6 | 76.06** | 0.19 |
| | Sample size | 1 | 89.91* | 0.04 |
| | Reliability coefficient | 5 | 455.70*** | 0.53 |
| Datasets without missing data | True reliability | 1 | 0.519 | 0.01 |
| | Test length | 1 | 4.457** | 0.10 |
| | Sample size | 1 | 2.195 | 0.05 |
| | Reliability coefficient | 5 | 12.606*** | 0.62 |

*Note. *p<0.05, **p<0.01,***p<0.001*

Upon examining Table 3, it is evident that the reliability coefficients have the most significant impact on RB values for datasets containing missing data, while the sample size has the most minor effect. For datasets without missing data, the most significant impact on RB values is associated with reliability estimation methods, with no significant effect observed for reliability level and sample size.

Figures 1-4 present the RB values obtained from simulation conditions. The parallel lines on the x-axis represent the acceptable range of |RB| as ±0.10.

**Figure 1.**
*Relative Bias Values for Datasets Without Missing Data*

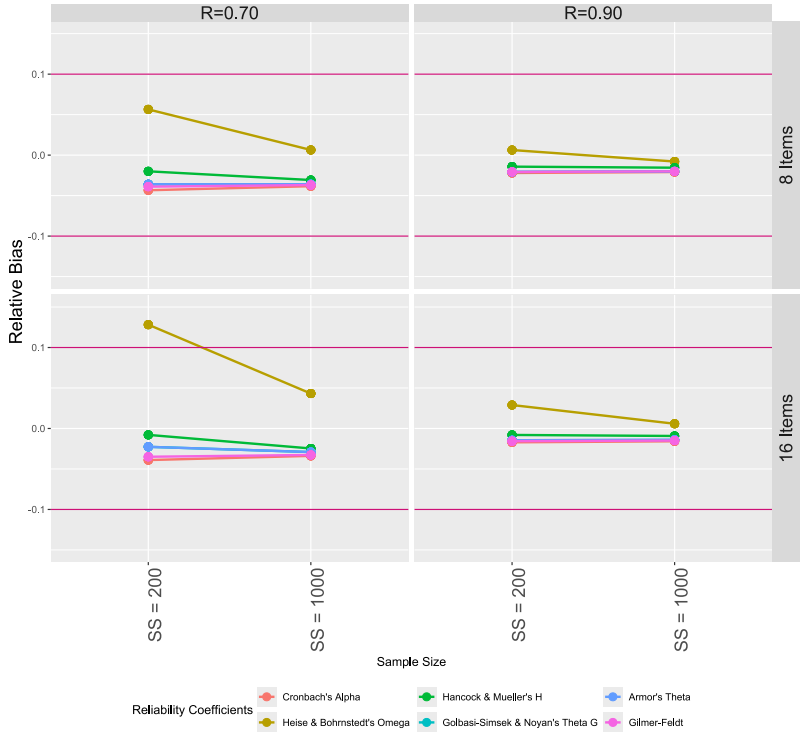



Figure 1 displays the relative bias (RB) values of reliability coefficients for datasets without missing data. It is observed that in the condition where the sample size is 200, the number of items is 16, and the reliability (R) is 0.70, Heise & Bohrnstedt's Omega coefficient overestimated reliability. However, all other reliability coefficients provided acceptable estimations across all conditions.

In Figure 2, the RB values of reliability coefficients are presented for datasets with a 5% missing data rate. Similar to datasets without missing data, in conditions where the sample size is 200, the number of items is 16, and reliability (R) is 0.70, Heise & Bohrnstedt's Omega coefficient overestimated reliability independent of the missing data mechanism and imputation method. However, all other reliability coefficients, except Heise & Bohrnstedt's Omega, provided reliability estimates within an acceptable RB range across all conditions, regardless of the number of items, sample size, imputation method, and missing data mechanism.

In Figure 3, the RB values of reliability coefficients are presented for datasets with a 10% missing data rate. Similar to datasets with a 5% missing data rate and datasets without missing data, Heise & Bohrnstedt's Omega coefficient overestimated reliability when the sample size was 200, the number of items was 16, and reliability (R) was 0.70. However, under the MAR missing data mechanism, when the true reliability was 0.70, and the number of items was 8, Cronbach's Alpha, Gilmer-Feldt's reliability coefficient, Armor's Theta, Gölbaşı-Şimşek & Noyan's Theta G, and Hancock & Mueller's H coefficient underestimated the true reliability when zero imputation was preferred. In all other conditions, all reliability coefficients (except Heise & Bohrnstedt's Omega when the number of items was 16, the sample size was 200, and R was 0.70) provided reliability estimates within an acceptable RB range.

In Figure 4, the RB values of reliability coefficients are presented for datasets with a 20% missing data rate. Under the MAR missing data mechanism, when the true reliability was 0.70, the number of items was 8, and the sample size was 200, Cronbach's Alpha, Gilmer-Feldt's reliability coefficient, Armor's Theta, and Gölbaşı-Şimşek & Noyan's Theta G underestimated the true reliability when zero imputation was preferred. Similarly, under the MAR missing data mechanism, when the true reliability was 0.70, the number of items was 16. The sample size was 1000, Cronbach's Alpha, Gilmer-Feldt's reliability coefficient, Armor's Theta, Gölbaşı-Şimşek & Noyan's Theta G, and Hancock & Mueller's H coefficient underestimated the true reliability when zero imputation was preferred. In all other conditions, all reliability coefficients (except Heise & Bohrnstedt's Omega when the number of items was 16, the sample size was 200, and R was 0.70) provided reliability estimates within an acceptable RB range.

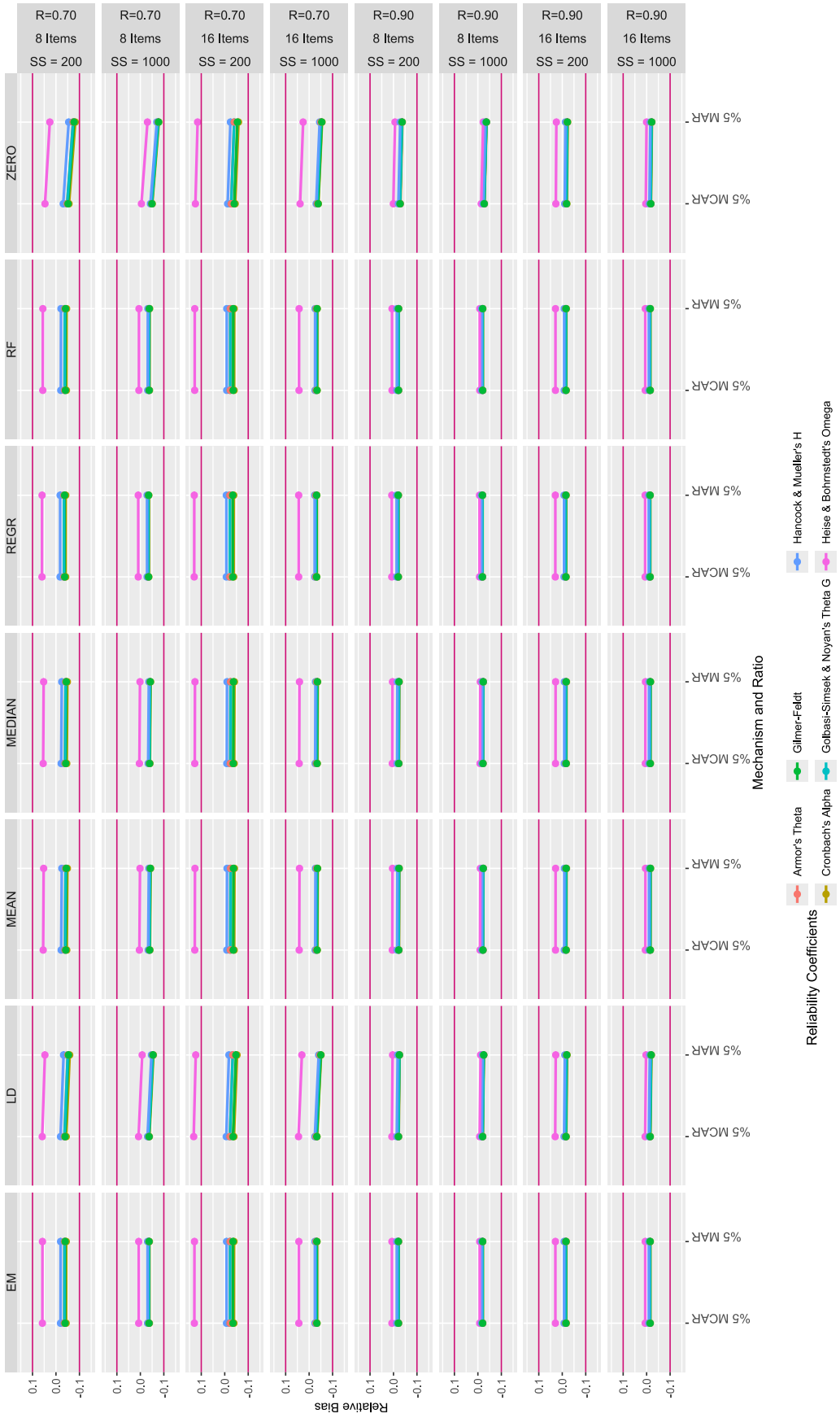



**Figure 2.**
*Relative Bias Values for Datasets with %5 Missing*

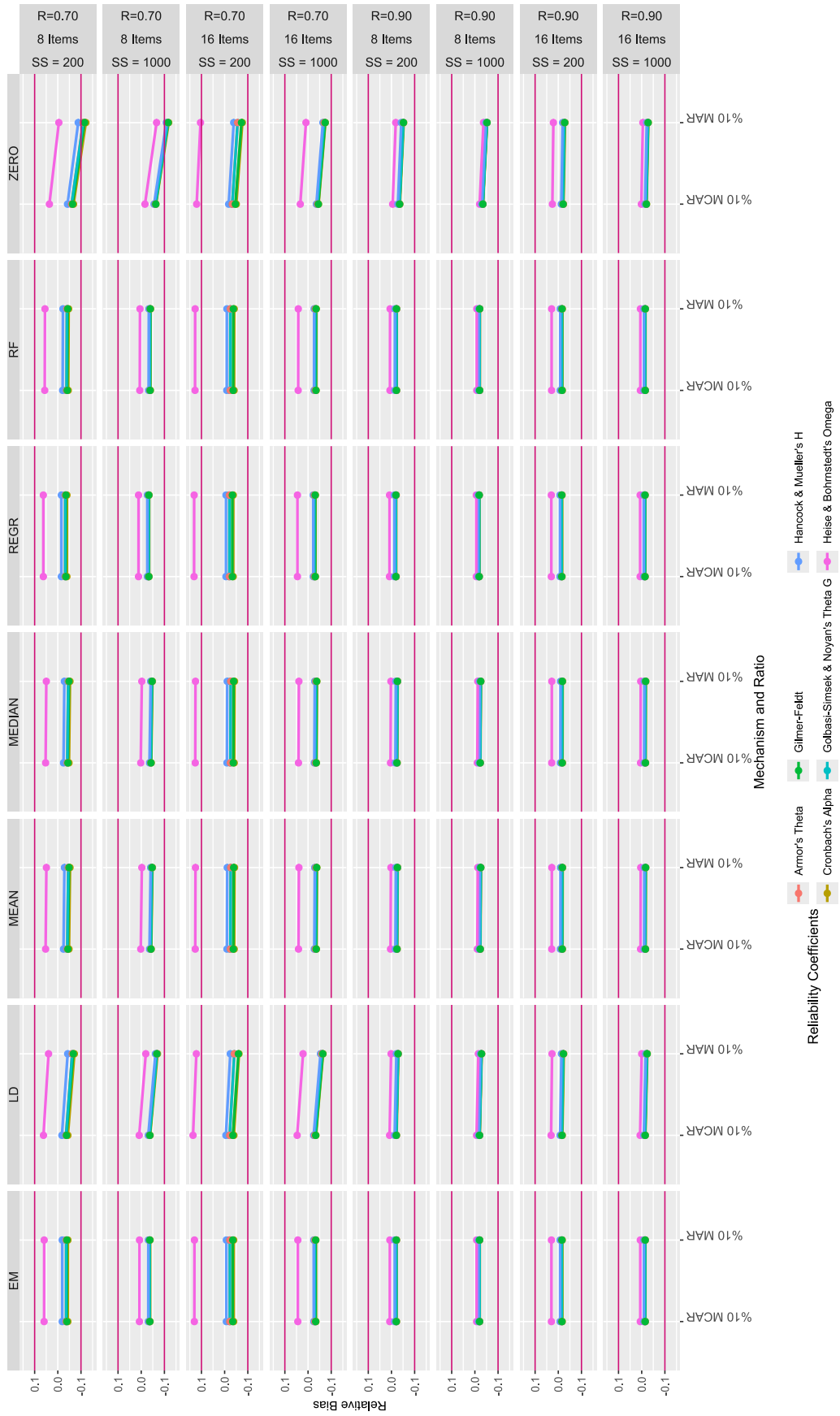



**Figure 3.**
*Relative Bias Values for Datasets with %10 Missing*

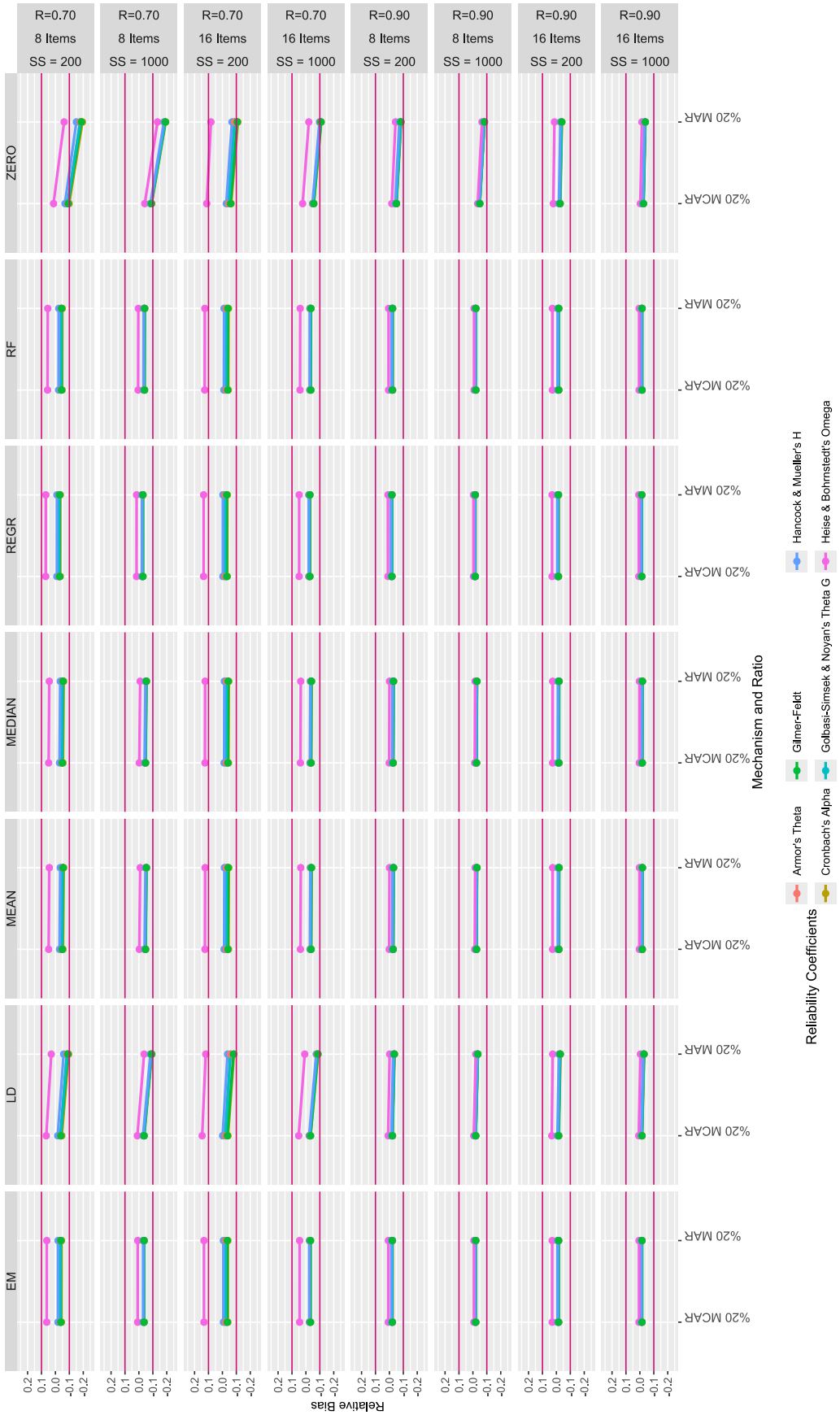



**Figure 4.**
*Relative Bias Values for Datasets with %20 Missing*

Upon examination of Figures 1, 2, 3, and 4, it was observed that Heise & Bohrnstedt's Omega coefficient tends to overestimate reliability, especially in conditions of low reliability (0.70), small sample size (200), and 16 items, regardless of missing data and imputation methods. Therefore, under these specific conditions, Heise & Bohrnstedt's Omega coefficient can be considered an upper limit for estimating reliability.

Using the zero-imputation method to handle missing data can yield misleading results, particularly when the missing data mechanism is MAR. When preferred alongside commonly used imputation methods such as mean imputation, median imputation, and regression imputation, as well as Random Forests imputation and expectation maximization methods, reliability coefficients generally provide estimations within acceptable limits ($|RB|<0.10$).

## Discussion

This study compares the relative bias values of reliability coefficients in unidimensional structures consisting of five-category ordinal indicators regarding the impact of missing data handling methods. Findings show that zero imputation method may cause to underestimation of reliability coefficients if missing data mechanism is MAR and when missing data ratio increases.

When the sample size gets lower for the complete data sets obtained with the listwise deletion method, it is seen that the reliability coefficients do not give biased results. This is in line with Enders' (2003) findings that when LD is preferred as a method of dealing with missing data in data sets with 15% missing data. Similar to Enders' (2003) study, the level of biased estimation for Cronbach's Alpha coefficient was within acceptable limits, and the bias decreased as the sample size increased. In the MAR and MCAR mechanism, the BM algorithm estimated Cronbach's Alpha coefficient with acceptable bias, similar to Enders' (2004) study. For the MCAR mechanism, LD, RF, ME, MD, RI, and EM algorithms gave unbiased results, which are theoretically and practically compatible. Analogous to these findings, in their study, Sheng and Shen (2012) observed that multiple imputation is efficient in enhancing dependability estimations, even when dealing with data distributions that are skewed. Franco-Martínez et al. (2022) highlighted the compensating nature of multiple imputation, indicating its ability to effectively handle intricate missing data patterns. Enders (2010) emphasized the effectiveness of regression imputation in producing unbiased estimates in several situations where data is missing.

When the effect size on the RB values obtained from the data sets with and without missing data is analyzed, it is seen that the largest effect size is in the reliability coefficients. Accordingly, the reliability coefficient is the most influential factor on the RB. In other words, the bias values obtained from the reliability coefficients differ from each other at a statistically significant level. In addition to this situation, reliability coefficients make unbiased predictions in general terms. The different values used in the calculation formulae of the reliability coefficients (Table 1) may have caused this. The "number of items," which is common in all formulae, has a moderate effect size in the data sets with and without missing data. These findings are also theoretically expected. Increasing the number of test items generally improves reliability estimates, reducing standard errors and bias (A. Hayes & Coutts, 2020; Sheng & Sheng, 2012; Trizano-Hermosilla & Alvarado, 2016; Turner et al., 2017).

The true reliability values (0.70 and 0.90) were obtained in line with McDonald's Omega coefficient formula. Since McDonald's Omega coefficient is calculated with reference to factor loadings, factor loadings vary at each true reliability level. The effect size on reliability levels estimated from sets with missing data is small ($\eta 2=0.05$). This may be related to the quality of the items (factor loadings).

## Results and Suggestions

This study demonstrates that due to their consistently unbiased outcomes across simulation conditions, the EM, RF, RI, $Mean_{Person}$, and $Median_{Item}$ imputation methods may be more beneficial than zero imputation and listwise deletion (LD) methods when estimating Cronbach's Alpha, Hancock & Mueller's H, Armor's Theta, Gölbaşı-Şimşek & Noyan's Theta G, and Gilmer-Feldt's reliability coefficient. It is noted that Heise & Bohrnstedt's Omega coefficient may overestimate reliability

independently of missing data issues. Therefore, although Heise & Bohrnstedt's Omega coefficient is based on factor loadings, indicating a congeneric measurement, it is recommended to assess its model-data fit at low-reliability levels due to low factor loadings before reporting. Considering that measurements in scale development and adaptation studies are often congeneric, Cronbach's Alpha coefficient has provided unbiased results even in datasets that do not meet this assumption. Despite this, researchers may still prefer Cronbach's Alpha coefficient in situations similar to simulation conditions when it yields similar estimates of true reliability. Still, this recommendation pertains solely to the impact of missing data imputation methods. It is crucial to note that there are simulation studies from various perspectives suggesting that Alpha tends to underestimate reliability (McNeish, 2018; Edwards et al., 2021). Furthermore, zero imputation directly affects the assumption of normal distribution, particularly for variables identified as normally distributed. Therefore, if zero imputation is preferred for such variables, a reevaluation of the normal distribution assumption is advisable. Additionally, the $Mean_{Item}$ and RI methods, despite dealing with categorical and discrete variables, perform continuous imputation, which alters the structure of the dataset. Given that educational sciences often work with categorical data, attention should be paid to this aspect in planned analyses following the reporting of reliability coefficients. It should be noted that this study is a Monte Carlo simulation, providing valid results for simulation conditions. Still, the validity of the results obtained decreases as the variables in real datasets deviate from these conditions. Therefore, the obtained results should be interpreted within these limitations. Future studies may manipulate the number and skewness of categories of variables to conduct simulation studies.

## Declarations

**Conflict of Interest:** No potential conflict of interest was reported by the author(s).

**Ethical Approval:** We declare that all ethical guidelines for authors have been followed by all authors. Ethical approval is not required as the datasets in the study were simulated.

## References

Allison, P. D. (2002). Missing data. Sage Publications.

Armor, D. J. (1974). Theta reliability and factor scaling. In H. Costner (Ed.), Sociological Methodology (pp. 17–50). Jossey-Bass.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. Journal of School Psychology, 48(1), https://doi.org/10.1016/j.jsp.2009.10.001

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. Structural Equation Modeling: A Multidisciplinary Journal, 13(2), 186–203. https://doi.org/10.1207/s15328007sem1302_2

Béland, S., Jolani, S., Pichette, F., & Renaud, J.-S. (2018). Impact of simple substitution methods for missing data on Classical test theory difficulty and discrimination. The Quantitative Methods for Psychology, 14(3), 180–192. https://doi.org/10.20982/tqmp.14.3.p180

Bennett, D. A. (2001). How can I deal with missing data in my study? Australian and New Zealand Journal of Public Health, 25(5), 464–469. https://doi.org/10.1111/j.1467-842X.2001.tb00294.x

Brown, T. A. (2006). Confirmatory factor analysis for applied research. Guilford Press.

Cheema, J. R. (2014). Some general guidelines for choosing missing data handling methods in educational research. Journal of Modern Applied Statistical Methods, 13(2), 53–75. https://doi.org/10.22237/jmasm/1414814520

Cho, E. (2023). reliacoef: Compute and compare unidimensional and multidimensional reliability coefficients (1.0.0) [R].

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), https://doi.org/10.1007/BF02310555

Dai, S. (2021). Handling missing responses in psychometrics: Methods and software. Psych, 3(4), 673–693. https://doi.org/10.3390/psych3040043

Dai, T., Du, Y., Cromley, J., Fechter, T., & Nelson, F. (2024). Analytic approaches to handle missing data in simple matrix sampling planned missing designs. The Journal of Experimental Education, 92(3), 531–558. https://doi.org/10.1080/00220973.2023.2196678

Doove, L. L., Van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. Computational Statistics & Data Analysis, 72, 92–104. https://doi.org/10.1016/j.csda.2013.10.025

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. British Journal of Psychology, 105 3, 399–412. https://doi.org/10.1111/bjop.12046

Edwards, A. A., Joyner, K. J., & Schatschneider, C. (2021). A simulation study on the performance of different reliability estimation methods. Educational and Psychological Measurement, 81(6), 1089–1117. https://doi.org/10.1177/0013164421994184

Enders, C. K. (2003). Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data. Psychological Methods, 8(3), 322–337. https://doi.org/10.1037/1082-989X.8.3.322

Enders, C. K. (2004). The impact of missing data on sample reliability estimates: Implications for reliability reporting practices. Educational and Psychological Measurement, 64, 419–436. https://doi.org/10.1177/0013164403261050

Enders, C. K. (2010). Applied missing data analysis. Guilford Press.

Fan, J., & Wu, W. (2022). A comparison of multiple imputation strategies to deal with missing nonnormal data in structural equation modeling. Behavior Research Methods. https://doi.org/10.3758/s13428-022-01936-y

Feldt, L. S., & Charter, R. A. (2003). Estimation of internal consistency reliability when test parts vary in effective length. Measurement and Evaluation in Counseling and Development, 36(1), 23–27. https://doi.org/10.1080/07481756.2003.12069077

Finch, W. H. (2016). Missing data and multiple imputation in the context of multivariate analysis of variance. The Journal of Experimental Education, 84(2), 356–372. https://doi.org/10.1080/00220973.2015.1011594

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. Psychological Methods, 9(4), 466–491. https://doi.org/10.1037/1082-989X.9.4.466

Gölbaşı-Şimşek, G., & Noyan, F. (2013). McDonald's ω t , Cronbach's α, and Generalized θ for composite reliability of common factors structures. Communications in Statistics - Simulation and Computation, 42(9), 2008–2025. https://doi.org/10.1080/03610918.2012.689062

Goretzko, D. (2021). Factor retention in exploratory factor analysis with missing data. Educational and Psychological Measurement, 82, 444–464. https://doi.org/10.1177/00131644211022031

Goretzko, D., Heumann, C., & Bühner, M. (2020). Investigating parallel analysis in the context of missing data: A simulation study comparing six missing data methods. Educational and Psychological Measurement, 80, 756–774. https://doi.org/10.1177/0013164419893413

Gorsuch, R. L. (2015). Factor analysis (Classic edition). Routledge, Taylor & Francis Group.

Graham, J. W. (2012). Missing data. Springer New York. https://doi.org/10.1007/978-1-4614-4018-5

Graham, J. W., Cumsille, P., & Shevock, A. E. (2013). Methods for handling missing data. In I. B. Weiner (Ed.), Handbook of Psychology, Second Edition (pp. 109–141). http://doi.org/https://doi.org/10.1002/9781118133880.hop202004

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), Structural equation modeling: Present and future—A festschrift in honor of Karl Jöreskog (pp. 195–216). Scientific Software International.

Hayes, A., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But... Communication Methods and Measures, 14, 1–24. https://doi.org/10.1080/19312458.2020.1718629

Hayes, T., & McArdle, J. J. (2017). Should we impute or should we weight? Examining the performance of two CART-based techniques for addressing missing data in small sample research with nonnormal variables. Computational Statistics & Data Analysis, 115, 35–52. https://doi.org/10.1016/j.csda.2017.05.006

Heise, D. R., & Bohrnstedt, G. W. (1970). Validity, invalidity, and reliability. Sociological Methodology, 2, 104. https://doi.org/10.2307/270785

Howell, D. C. (2007). The treatment of missing data. In W. Outhwaite & S. Turner, The SAGE Handbook of Social Science Methodology (pp. 212–226). SAGE Publications Ltd. https://doi.org/10.4135/9781848607958.n11

Lee, D. Y., Harring, J. R., & Stapleton, L. M. (2019). Comparing methods for addressing missingness in longitudinal modeling of panel data. The Journal of Experimental Education, 87(4), 596–615. https://doi.org/10.1080/00220973.2018.1520683

Lee, H. J., & Huber, J. C. Jr. (2021). Evaluation of multiple imputation with large proportions of missing data: How much is too much? Iranian Journal of Public Health. https://doi.org/10.18502/ijph.v50i7.6626

Lei, P.-W., & Shiverdecker, L. K. (2020). Performance of estimators for confirmatory factor analysis of ordinal variables with missing data. Structural Equation Modeling: A Multidisciplinary Journal, 27(4), 584–601. https://doi.org/10.1080/10705511.2019.1680292

Leite, W., & Beretvas, S. N. (2010). The performance of multiple imputation for likert-type items with missing data. Journal of Modern Applied Statistical Methods, 9(1), 64–74. https://doi.org/10.22237/jmasm/1272686820

Li, J., & Lomax, R. G. (2017). Effects of missing data methods in SEM under conditions of incomplete and nonnormal Data. The Journal of Experimental Education, 85(2), 231–258. https://doi.org/10.1080/00220973.2015.1134418

Little, R., & Rubin, D. (2002). Statistical analysis with missing data (1st ed.). Wiley. https://doi.org/10.1002/9781119013563

Little, R., & Rubin, D. (2019). Statistical analysis with missing data (3rd ed.). Wiley. https://doi.org/10.1002/9781119482260

McAllister, D. J., & Bigley, G. A. (2002). Work context and the definition of self: How organizational care influences organization-based self-esteem. Academy of Management Journal, 45(5), 894–904. https://doi.org/10.2307/3069320

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. British Journal of Mathematical and Statistical Psychology, 23(1), 1–21. https://doi.org/10.1111/j.2044-8317.1970.tb00432.x

McDonald, R. P. (1999). Test theory: A unified treatment. Lawrence Erlbaum Associates.

McNeish, D. M. (2017). Exploratory factor analysis with small samples and missing data. Journal of Personality Assessment, 99(6), 637–652. https://doi.org/10.1080/00223891.2016.1252382

McNeish, D. M. (2018). Thanks coefficient alpha, we'll take it from here. Psychological Methods, 23(3), 412–433. https://doi.org/10.1037/met0000144

Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: Time Series Missing Value Imputation in R. The R Journal, 9(1), 207. https://doi.org/10.32614/RJ-2017-009

Myers, T. A. (2011). Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. Communication Methods and Measures, 5(4), 297–310. https://doi.org/10.1080/19312458.2011.624490

Newman, D. A. (2014). Missing data. Organizational Research Methods. https://doi.org/10.1177/1094428114548590

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3nd ed.). McGraw-Hill.

Parent, M. C. (2013). Handling item-level missing data: Simpler is just as good. The Counseling Psychologist, 41(4), 568–600. https://doi.org/10.1177/0011000012445176

Revelle, W. (2024). psych: Procedures for psychological, psychometric, and personality research (R package version 2.4.1) [Computer software]. https://CRAN.R-project.org/package=psych

Rockel, T. (2022). missMethods: Methods for missing data (0.4.0) [R]. https://github.com/torockel/missMethods

Roth, P. L., Switzer, F. S., & Switzer, D. M. (1999). Missing data in multiple item scales: A monte carlo analysis of missing data techniques. Organizational Research Methods, 2(3), 211–232. https://doi.org/10.1177/109442819923001

Şahin Kürşad, M., & Nartgün, Z. (2015). Kayıp veri sorununun çözümünde kullanılan farklı yöntemlerin ölçeklerin geçerlik ve güvenirliği bağlamında karşılaştırılması. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 6(2), 254–267. https://doi.org/10.21031/epod.95917

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. Psychological Methods, 7(2), https://doi.org/10.1037/1082-989X.7.2.147

Scheffer, J. (2002). Dealing with missing data (1st ed.). Massey University. https://mro.massey.ac.nz/handle/10179/4355

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. American Journal of Epidemiology, 179(6), 764–774. https://doi.org/10.1093/aje/kwt312

Sheng, Y., & Sheng, Z. (2012). Is coefficient alpha robust to non-normal data? Frontiers in Psychology, 3(34). https://doi.org/10.3389/fpsyg.2012.00034

Sigal, M. J., & Chalmers, R. P. (2016). Play it again: Teaching statistics with monte carlo simulation. Journal of Statistics Education, 24(3), 136–156. https://doi.org/10.1080/10691898.2016.1246953

Sijtsma, K., & Van Der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. Multivariate Behavioral Research, 38(4), 505–528. https://doi.org/10.1207/s15327906mbr3804_4

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach's Alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. Frontiers in Psychology, 7. https://doi.org/10.3389/fpsyg.2016.00769

Turner, H. J., Natesan, P., & Henson, R. K. (2017). Performance Evaluation of confidence intervals for ordinal coefficient alpha. Journal of Modern Applied Statistical Methods, 16(2), 157–185. https://doi.org/10.22237/jmasm/1509494940

Uysal, İ., & Kılıç, A. (2022). Normal dağılım ikilemi. Anadolu Journal of Educational Sciences International, 12(1), 220–248. https://doi.org/10.18039/ajesi.962653

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3). https://doi.org/10.18637/jss.v045.i03

Van Ginkel, J. R., Van Der Ark, L. A., & Sijtsma, K. (2007). Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results. Multivariate Behavioral Research, 42(2), 387–414. https://doi.org/10.1080/00273170701360803

Wei, R., Wang, J., Su, M., Jia, E., Chen, S., Chen, T., & Ni, Y. (2018). Missing value imputation approach for mass spectrometry-based metabolomics data. Scientific Reports, 8(1), 663. https://doi.org/10.1038/s41598-017-19120-0

Zhang, Z. (2016). Missing data imputation: Focusing on single imputation. Annals of Translational Medicine, 4(1), 1–8. https://doi.org/10.3978/j.issn.2305-5839.2015.12.38

Zhang, Z., & Yuan, K.-H. (2016). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: Methods and software. Educational and Psychological Measurement, 76(3), 387–411. https://doi.org/10.1177/0013164415594658